Capability Provenance in Language Models: A Case Study in Social Reasoning

Glenn Matlin * **Chandreyi Chakraborty**
Saehee Eom † Mika Okamoto † Rayan Castilla †
Louis Jaburi Alvin Deng Taywon Min Lucia Quirke
Stella Biderman Mark Riedl

Georgia Institute of Technology, College of Computing MATS Program
EleutherAI KAIST AI Georgia Tech AI Safety Initiative
†Equal contribution

Website Code HuggingFace Discord

Abstract

We use training-data attribution as an interpretable tool for capability discovery, mapping which regions of the pretraining corpus support social reasoning versus STEM reasoning in OLMo3-7B. Training-data attribution measures how strongly each training document influences a model's predictions on a benchmark, but document-level scores are too noisy to identify which corpus regions support which capabilities. We compute gradient-based attribution (TrackStar via Bergson) over a working set drawn from the de-duplicated Dolma3 mix, aggregate influence across WebOrganizer's 24-format × 24-topic taxonomy (576 bins), and contrast benchmark pairs in a 2×2 design that varies domain (social vs. STEM) and capability type (reasoning vs. knowledge): SocialIQA and MMLU Social Sciences against ARC-Challenge and MMLU STEM. Social and STEM reasoning draw on qualitatively distinct corpus regions, and the contrast is sharper at the reasoning level than at the knowledge level. Targeted machine unlearning provides partial causal validation: forgetting high-attribution topics (e.g., Literature for SocialIQA) degrades the aligned benchmark more than within-topic random baselines. We open-source all code, data artifacts, influence scores, and checkpoints at https://github.com/HCAI-Lab-GT/capabilibara and https://huggingface.co/HCAI-Lab-GT.

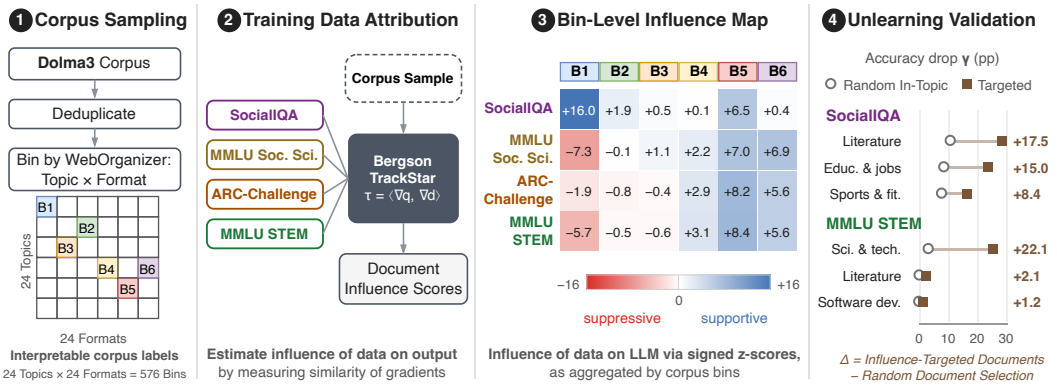


Figure 1: Capability provenance pipeline. **1)** Dolma3 stratified into 576 WebOrganizer bins. **2)** Bergson/TrackStar attributes probes to bins. **3)** Signed z-scores for curated bins B1 Lit.×Cust. Support, B2 Social×Q&A, B3 Travel×Q&A, B4 Sci.&Tech.×Academic, B5 Industrial×Docs, B6 Politics×Docs. **4)** Targeted unlearning beats random controls.

*Corresponding author: glenn@gatech.edu

1 Introduction

Training-data attribution (TDA) methods estimate how much each training example influences a model's predictions—quantifying which documents in the pretraining corpus matter most for a given behavior. Recent work using influence estimates has shown that knowledge-oriented tasks tend to rely on fewer documents, while reasoning tasks draw support from broader regions of the training corpus (Ruis et al., 2025). However, these analyses operate at the document level: they show how influence is distributed across training documents, but not *where in the corpus* distinct capabilities, such as social versus STEM reasoning, find their support.

Identifiable corpus regions can be inspected or reweighted, so a map from capabilities onto those regions is usable for auditing and for shaping behavior. We treat aggregate gradient-based attribution as a coarse mechanistic decomposition, complementary to internal-state interpretability because it analyzes inputs rather than activations. Document-level estimates support corpus-level claims only after they are aggregated in a structured taxonomy.

We study capability provenance in the open OLMo ecosystem (Soldaini et al., 2024; Groeneveld et al., 2024) using TrackStar (Chang et al., 2025) gradient-based attribution, implemented in the Bergson library (Quirke et al., 2026), over a stratified working set of 5,678,621 documents (10.5B tokens; 10,000/bin) from the de-duplicated Dolma3 mix of 1,263,587,168 ($\sim$1.26B) unique documents. WebOrganizer (Wettig et al., 2025) assigns each document a topic (e.g., Social Life, Literature, Software Development) and a format (e.g., Q&A Forum, Customer Support, Documentation). Our attribution unit is the resulting topic–format bin ($24 \times 24 = 576$ bins), such as Social Life $\times$ Q&A Forum or Literature $\times$ Customer Support (full taxonomy in Appendix A). We choose the four primary benchmarks as a controlled OLMES-compatible 2×2: domain (social vs. STEM) crossed with capability type (reasoning vs. knowledge) (Gu et al., 2025). The reasoning cells are **SocialIQA** (Sap et al., 2019) and **ARC-Challenge** (Clark et al., 2018), which ask likelihood-scored multiple-choice questions over social commonsense narratives and multi-step scientific problems; the knowledge cells are **MMLU Social Sciences** and **MMLU STEM** (Hendrycks et al., 2021b). Because all four targets are fixed-choice, above-chance OLMo3 evaluations under the same protocol, this pairing lets us ask whether corpus regions separate by domain, by capability type, or by their interaction (Figure 1).

The resulting profiles separate SocialIQA from the three comparison tasks at the topic and format levels. We then use targeted unlearning to test whether high-attribution regions are functionally necessary, and extend attribution to a held-out suite of social-reasoning probes spanning theory of mind, moral judgment, bias, and pragmatics.

We frame the study around three research questions, progressing from description to comparison to unlearning validation.

- RQ1 Where does each benchmark's attribution concentrate in the topic$\times$format taxonomy, do social and STEM benchmarks show distinct profiles, and what kinds of text characterize the high-influence bins?
- RQ2 Do social- and STEM-reasoning benchmarks draw attribution from different corpus regions, and does the same hold for the knowledge pair?
- RQ3 Are high-attribution corpus regions validated by targeted unlearning? Does forgetting them degrade the aligned benchmark more than matched random controls?

Our findings show that SocialIQA has a broader and sharply distinct training-data profile from social-domain knowledge and STEM reasoning, extending evidence that different capabilities can depend on different pretraining evidence (Ruis et al., 2025; Antoniades et al., 2024; Razeghi et al., 2024). Lexically, its high-influence bins carry a bimodal signature in which short interpersonal text and long-form documentation both contribute. Influence-targeted unlearning validates this association most clearly for SocialIQA, while weaker, null, and reversed effects on the comparison tasks delimit where attribution identifies functionally necessary documents (Zhao et al., 2024; Jin et al., 2025). The held-out probes share a provenance signature that SocialIQA captures only weakly, so the causally validated result samples one region of a broader capability space.

2 Related Work

Training-data attribution identifies which corpus parts shape model behavior. Classical influence functions estimate up-weighting effects without retraining (Koh & Liang, 2017); later work introduced scalable gradient-based approximations such as TracIn (Pruthi et al., 2020) and TRAK (Park et al., 2023), with improved inverse-Hessian methods for large models (Schioppa et al., 2022; Grosse et al., 2023). Subsequent research traces model outputs or factual knowledge back to training data (Akyurek et al., 2022; Chang et al., 2025) and builds retrieval or indexing systems for corpus-scale attribution (Lin et al., 2024; Liu et al., 2024; 2025). Concurrently, prior work identifies limitations of example-level estimates: brittleness to approximation choices (Basu et al., 2021; Grosse et al., 2023), failure to capture interactions between training examples (Guu et al., 2023), and inconsistent cross-task behavior (Jiao et al., 2025).

Prior corpus-scale capability-provenance work either characterizes top-k influential documents *post hoc* via language models or human annotation (Ruis et al., 2025; Razeghi et al., 2024), or unlearns putative dataset components on closed-data models via proxy datasets (Zhao et al., 2024). We instead apply structured taxonomic decomposition (WebOrganizer; Wettig et al., 2025) *before* inspection, on the released pretraining corpus; to our knowledge this is the first application of gradient-based corpus-scale TDA to *capability provenance* with taxonomic aggregation on a fully open data ecosystem.

Our approach relies on an open-model ecosystem whose pretraining corpus, training pipeline, and checkpoints are all released at the document level, such as Dolma (Soldaini et al., 2024) and OLMo3 (Ettinger et al., 2025). *Open-weight* and *open-data* are distinct: models that publish weights without their pretraining corpora (Longpre et al., 2024) cannot support document-level attribution. Among fully open families (OLMo, Pythia (Biderman et al., 2023), the Common Pile-trained models (Kandpal et al., 2025)), OLMo3 currently has the most mature evaluation methodology (OLMES; Gu et al., 2025).

Attribution scores are associational, and machine unlearning offers an intervention to complement them. Recent work explores unlearning for analyzing training data effects: NGDiff (Jin et al., 2025) forgets targeted data via gradient-normalized objectives without catastrophic loss of general ability, and Zhao et al. (2024) show that unlearning specific pretraining slices reveals which data components drive downstream performance. We adopt bin-level unlearning to validate our strongest attribution findings.

We ask whether different capability families—such as social reasoning, social-domain knowledge, and quantitative reasoning—draw support from systematically different regions of pretraining data. Recent tracing work shows that capabilities depend on distinct types of pretraining evidence (Ruis et al., 2025), memorization and generalization can be analyzed via attribution (Antoniades et al., 2024), and mathematical reasoning can be traced to specific data sources (Razeghi et al., 2024). We extend this line of work by asking whether such distinctions emerge at the aggregate corpus level. Complementary data-centric literature selects and audits pretraining corpora at scale (Albalak et al., 2024; Elazar et al., 2024). Capability-level provenance maps supply the evidence such curation efforts currently lack.

3 Method

We study capability provenance through data attribution, using targeted benchmarks as queries against an open-data model, complemented by validation via machine unlearning. Our setting is the OLMo3-7B/Dolma3 ecosystem, where we run the pipeline end to end, from corpus profiling to causal validation.

3.1 Corpus Construction

Training data attribution requires a model family whose pretraining corpus and checkpoints are both released. We therefore center the study on the OLMo3-7B open-data ecosystem (Ettinger et al., 2025): OLMo3-7B Base is the object of pretraining-data attribution over Dolma3, while OLMo3-7B Instruct supplies the benchmark completions and query gradients used

Table 1: Contrastive benchmark design: domain (social vs. STEM) crossed with capability type (reasoning vs. knowledge). Cell tint follows the manuscript's semantic benchmark palette; citations for the four benchmarks appear at first mention in the introduction.

Domain	**Reasoning**	**Knowledge**
Social	**SocialIQA** social reasoning	**MMLU Social Sciences** social knowledge
STEM	**ARC-Challenge** STEM reasoning	**MMLU STEM** STEM knowledge

for scoring. This split follows the base-document/instruct-query design of Ruis et al. (2025): document gradients and corpus-side curvature are computed on Base so the index remains tied to pretraining data, whereas query gradients are computed on Instruct because it follows the OLMES multiple-choice and chat-style prompts more reliably.

The 6T token mix was generated with quality-aware upsampling, in which higher-quality data is replicated more frequently in the training mixture to increase its effective weight (Soldaini et al., 2024). We de-duplicated the dataset to avoid duplicate attributions from upsampled documents. We then classify the Common Crawl and olmOCR portions of the de-duplicated corpus (1,098,646,162 documents, ∼87% of the population) with topic and format classifiers from WebOrganizer (Wettig et al., 2025), yielding taxonomies of 24 topics $\times$ 24 formats, resulting in 576 bins. The remaining code, mathematics, and reference families (Stack-Edu, FineMath, arXiv, Wikipedia) are retained in the population count but not taxonomized, since the web-oriented WebOrganizer taxonomy does not apply to them. These bins are the units of analysis for every corpus-level comparison that follows.

We estimate influence on a stratified sample of the de-duplicated dataset in which every taxonomy bin is equally represented. The Dolma3 corpus (Soldaini et al., 2024) exhibits substantial skew, so naive sampling would under-sample low-volume bins and yield unstable gradient estimates (see Appendices B.2 and B.3). We therefore construct the working set by sampling uniformly from each bin, which puts all bins on a common footing for comparison.

3.2 Training Data Attribution

With the stratified corpus established, we define the attribution queries. Each benchmark sample is a query to the attribution pipeline. The four core benchmarks below are the primary queries; a held-out suite of social-reasoning probes reuses the same query pipeline for the generalization analysis (construction and results in Appendix G).

We adopt a 2×2 contrastive design crossing domain (social vs. STEM) with capability type (reasoning vs. knowledge), shown in Table 1: SocialIQA for social commonsense reasoning (Sap et al., 2019), ARC-Challenge for science reasoning (Clark et al., 2018), and MMLU subdomains for social-science and STEM knowledge (Hendrycks et al., 2021b). The design distinguishes domain-specific effects, capability-specific effects, and domain-general patterns shared across all benchmarks. All four benchmarks are part of the OLMo3 base evaluation suite, evaluated with OLMES (Gu et al., 2025) following the exact protocol used during model development (Appendix C).

Unit of analysis

A WebOrganizer **bin** is one topic$\times$format cell. For each benchmark, the attribution method assigns that bin a signed, within-benchmark z-score: **positive** values indicate supportive influence on the benchmark query gradients, while **negative** values indicate suppressive or contrasting influence.

With the benchmarks established, we compute document-level attribution scores using TrackStar (Chang et al., 2025) as implemented in Bergson (Quirke et al., 2026). Gradient-

based training-data attribution estimates how strongly each training document affects a model's behavior on a downstream query by comparing the gradient it would induce during training with the gradient the model produces on the query. A document whose gradients consistently align with a benchmark's query gradients is judged to have positive influence on that benchmark, and one whose gradients consistently oppose them is judged to have negative influence. TrackStar is a scalable variant that projects these gradients into a low-rank subspace so the comparison runs at corpus scale. We build the corpus gradient index from OLMo3-7B Base and reduce benchmark query gradients from OLMo3-7B Instruct. The influence score compares pretraining-document gradients against the instruction-tuned model's task gradients, under the approximation used by Ruis et al. (2025) that the supervised-instruction-tuning second-order term is the identity. The output is a signed influence score per document–benchmark pair. Implementation details are in Appendix C.

3.3 Influence Aggregation

We aggregate attribution over taxonomy bins because individual document scores can be brittle under influence-function approximations and sensitive to example interactions (Basu et al., 2021; Grosse et al., 2023; Guu et al., 2023; Jiao et al., 2025). The WebOrganizer bins provide fixed, semantically named aggregation units (Wettig et al., 2025), and the resulting patterns are stable enough to interpret at corpus-region level. We aggregate with the mean rather than the median because the within-bin influence distribution is long-tailed and the tail carries the signal that drives the targeted unlearning forget sets; the paired unlearning statistics reported later use medians, where the goal is instead to resist outlier runs.

Formally, for each benchmark β and bin k, we compute the mean influence $\bar{s}$:

$$\bar{s}_{\beta,k} = \frac{1}{|D_k|} \sum_{j \in D_k} \frac{1}{|Q_\beta|} \sum_{q \in Q_\beta} s(j,q),$$

where $s(j,q)$ is the attribution score between training document j and benchmark query q, Q_β is the set of queries for benchmark β, and D_k is the set of documents in bin k.

This produces a 576×4 matrix of bin-level influence values, used in all subsequent analyses.

For each pair of benchmarks (β, β'), we compute the per-bin difference $\Delta z_k = z_\beta(k) - z_{\beta'}(k)$, where $z_\beta(k)$ is the standardized bin-level influence for benchmark β at bin k (computed via the aggregation rule above), identifying bins that are discriminative between benchmarks versus shared. We report bins where $|\Delta z|$ exceeds the bootstrap-derived noise floor as discriminative; confidence intervals are derived from 1,000-iteration resampling over documents within bin and shown in the corresponding figures.

For each benchmark β, we partition the eval queries Q_β into a *correct cohort* Q_β^+ (queries the model answers correctly under standard OLMES scoring) and an *incorrect cohort* Q_β^-. We then compute per-bin influence aggregates separately for each cohort using the same bin-level aggregation rule, yielding $z_\beta^+(k)$ and $z_\beta^-(k)$. The correctness differential $\Delta z_\beta^c(k) = z_\beta^+(k) - z_\beta^-(k)$ identifies bins whose content disproportionately drives the model's *correct* predictions on β, separating provenance-of-capability (positive Δz^c) from provenance-of-error (negative Δz^c), with bootstrap 95% confidence intervals as above.

To characterize the text in high-influence bins, we compute an open, LIWC-adjacent profile over the working set: in-repo pronoun, mental-state, function-word, and cognition counters, social/affective/topical categories from the open-source Empath analyzer (Fast et al., 2016), and direction-preserving proxies for the four LIWC summary variables (Analytic, Clout, Authentic, Tone). These are not LIWC.app outputs; the crosswalk, lexicons, and z-scoring baselines, with the expanded characterization, are detailed in Appendix H.

3.4 Validation via Targeted Machine Unlearning

While attribution identifies associational patterns, it does not show that the identified data is necessary for a given capability. We complement our influence analysis with targeted machine unlearning, treating topic-level forgetting as a diagnostic intervention, following prior uses of unlearning to analyze pretraining-data effects (Zhao et al., 2024). Because the causal claim concerns pretraining data, unlearning is applied to OLMo3-7B Base rather than the instruction-tuned checkpoint; this avoids treating post-training behavior as a direct forget-set target. Thus, the capability profiles reported for attribution use the deployment-realistic Instruct checkpoint, whereas the unlearning degradation γ is a base-model measurement. Following the NGDiff framework (Jin et al., 2025), we extend unlearning to the granular WebOrganizer taxonomy.

We structure these interventions around two axes: selection strategy (random vs. influence-guided) and data scope (single-topic vs. global). All unlearning runs train a rank-8 LoRA adapter on OLMo3-7B Base with the NGDiff gradient-normalized objective (Jin et al., 2025), then merge the adapter back into the base checkpoint for evaluation. We report raw accuracy damage, where positive values mean the unlearned checkpoint is less accurate than the Base checkpoint:

$$\gamma = A_{\text{baseline}} - A_{\text{unlearned}}$$

Detailed optimization parameters, resampling strategies, and the randomized-text stopping criterion are provided in Appendix I.

3.4.1 Experimental Conditions

The exploratory sweep runs, for each of the 24 topics, a **random baseline** (k=2,000 documents sampled uniformly in-topic) against **influence-guided** top-k selection for the target benchmark, plus a **global random control** with no topic alignment; this single-topic sweep and its net-effect correction are analyzed in Appendix I.

The paper's primary causal contrast is a paired, seed-replicated refinement of the same two axes. For each topic we forget the top-200 documents by per-document influence for the target benchmark, paired against 1,000 random documents from the same topic, with a flat 9,000-document retain set and three seeds per condition. We test the influence-versus-random difference with paired statistics at the topic level, which avoids inflating statistical power from repeated draws within the same topic (§4.3; full protocol in Appendix I).

4 Results

Unless otherwise noted, all influence scores reported in this section are z-score standardized within each benchmark over WebOrganizer bins. For each benchmark, we standardize the per-bin signed mean influence scores to have a mean of 0 and a standard deviation of 1 across bins; "absolute influence" refers to the absolute value of these standardized scores.

The working set contains 5,678,621 documents ($\sim$10.5B tokens) stratified across 576 WebOrganizer (Wettig et al., 2025) bins (10,000 per bin; see Appendix B.4 for bin fill rates).

4.1 Attribution Profiles Across Topics and Formats

Figure 2 summarizes the qualitative pattern that distinguishes social-reasoning attribution from the other three benchmarks. The social-knowledge and STEM benchmarks concentrate their strongest **positive influence** in documentation-like formats, while **SocialIQA**'s strongest format marginal is *Customer Support*, an interpersonal, dialogue-heavy format. At the topic level, *Literature* is **positive** for SocialIQA but **negative** for all three comparison benchmarks, and SocialIQA leads every comparison benchmark on *Social Life*: SocialIQA's support comes from informal, interpersonal text, and the comparison benchmarks' support comes from formal, expository text. Individual topic$\times$format cells hold too few documents

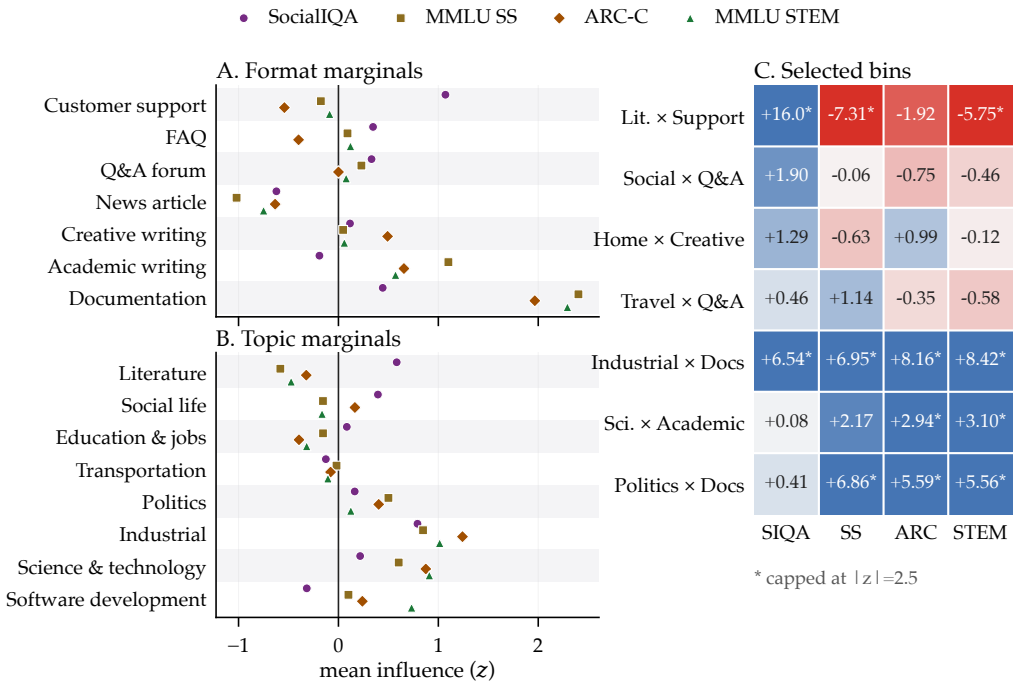


Figure 2: **SocialIQA** is the provenance-structure outlier, with a bin-level profile that correlates with each comparison benchmark at only $r \leq 0.21$ versus $r = 0.76$–0.86 among the three (Appendix D.1). **A–B** show marginal mean signed influence (z) over the 576-bin grid by format and topic on a shared z axis, rows sorted by the SocialIQA-minus-comparison contrast; *Documentation* drives the comparison benchmarks, whereas *Customer Support* and *Literature* are positive only for SocialIQA, which also leads *Social Life* by a wide margin. **C** gives seven topic–format cells ordered by the same contrast, spanning the range from SocialIQA-exclusive to comparison-exclusive support. *Literature* × *Customer Support* is extreme for SocialIQA (+16.0) yet negative for all three comparison benchmarks (−7.31, −1.92, −5.75), whereas *Politics* × *Documentation* inverts that pattern; colors are capped at $|z| = 2.5$, and asterisks mark cells at or beyond the cap. Full heatmaps in Figure 12.

to carry claims on their own, so we read cell-level views (Panel C) as illustrations; the full signed heatmaps are in Figure 12 (Appendix D).

SocialIQA's positive influence is distributed far more evenly across topics and formats than the other three benchmarks, whose strong positive mass concentrates in the Documentation column while News Article is a consistent negative (Appendix D). The single strongest SocialIQA cell, Literature × Customer Support, sits where the topic and format marginals predict it: an interpersonal format inside a narrative topic.

The per-topic and per-format marginals, reported in Appendix D, show that the observed patterns are not isolated to a single extreme bin. The near-neutral *Creative Writing* and *Transportation* rows are useful negative checks: the social-reasoning signal is not simply any narrative-looking format or broad human-centered topic. A whole-profile comparison confirms the effect is not driven by one bin, with SocialIQA's 576-bin influence profile correlating with the three comparison benchmarks at $r \leq 0.21$ versus $r = 0.76$–0.86 among them, making it the only benchmark with a significant profile-divergence statistic (Appendix D.1).

At the topic level, the clearest social-reasoning divergence is the Literature sign flip: it has positive mean influence for SocialIQA while remaining among the most negatively influential topics for the other three benchmarks, and it is the strongest SocialIQA–ARC-Challenge topic contrast in Appendix E. This places more of SocialIQA's positive attribution in Literature-labeled and narrative/interpersonal corpus regions. In contrast, ARC-Challenge, MMLU Social Sciences, and MMLU STEM display more similar topic profiles,

with the Industrial and Science & Technology topics contributing strongly positive influence across those benchmarks; Science & Technology contributes substantially less for SocialIQA.

At the format level, marginal influence concentrates in fewer categories at larger magnitudes than at the topic level, while preserving the same contrast between SocialIQA and the other benchmarks. Documentation is the highest-influence format for ARC-Challenge, MMLU Social Sciences, and MMLU STEM, with the next-strongest formats (Academic Writing for ARC-Challenge and MMLU Social Sciences; Legal Notices for MMLU STEM) at clearly smaller magnitudes. SocialIQA again has the most distinct profile: Customer Support leads its format marginals by a wide margin, with Documentation, FAQ, and Q&A Forum close behind (Appendix D).

RQ1 attribution profile takeaway

SocialIQA's positive influence is diffuse and narrative/interpersonal; **MMLU Social Sciences**, **ARC-Challenge**, and **MMLU STEM** concentrate in documentation-like and technical regions.

4.2 Cross-Domain Contrastive Analysis

The paired-topic heatmaps in Appendix E show a sharper social–STEM separation for reasoning (SocialIQA–ARC-Challenge) than for knowledge (MMLU Social Sciences–MMLU STEM), while preserving the same direction: Software Development tilts STEM and Education & Jobs tilts social. The asymmetry is quantitative as well as visual: the reasoning pair's top-3 topic contrasts span $|\Delta z| \leq 0.91$ versus ≤ 0.63 for the knowledge pair, roughly $1.4\times$ the dynamic range, so in this controlled 2×2 design capability type differentiates corpus provenance more than domain does.

RQ2 contrast takeaway

The social–STEM split is sharper for reasoning than knowledge: **SocialIQA**−**ARC-Challenge** separates corpus regions more strongly than **MMLU Social Sciences**−**MMLU STEM**, though the direction is preserved.

Three additional analyses, detailed in the appendix, refine this picture. A correctness partition reveals a reasoning-vs-knowledge divergence: the text most associated with correct answers on both reasoning benchmarks (**SocialIQA**, **ARC-Challenge**) is dialogue-rich, interpersonal writing, and the same text is associated with errors on both knowledge benchmarks (**MMLU Social Sciences**, **MMLU STEM**); the sharpest single example is the *Literature* $\times$ *Customer Support* cell (Appendix F). A wider held-out suite of social, Theory-of-Mind, moral, and pragmatic probes runs through the same pipeline, analyzed in Appendix G. Lexical profiling shows **SocialIQA**'s high-influence bins are bimodal: dialogue-rich interpersonal text alongside long-form documentation (Appendix H).

To test how far the SocialIQA lens generalizes, we extend attribution to the OLMES-gated held-out suite of Appendix G: nine probes spanning theory of mind, moral and normative judgment, social bias, and pragmatics, scored against the same gradient index and aggregated over the same 576-bin taxonomy. Seven of the probes share a provenance block (mean within-block $r = 0.70$) whose supportive bins carry a documentary, formal-register signature, while SocialIQA sits outside it ($r = 0.22$ to the block, 0.20 across all nine held-out probes; Figure 34, Appendix G.9). A held-out unlearning test contributes one causal data point beyond the four-benchmark design (Appendix G.10). We read this as scoping the SocialIQA result: social reasoning is not monolithic, and the causally validated finding anchors one region of a broader space of socially relevant capabilities, each with its own corpus support.

4.3 Empirical Validation via Machine Unlearning

We validate the attribution results with machine unlearning, comparing influence-targeted document removal against in-topic random removal across the four benchmarks.

In the faithful replication, which unlearns the top-200 documents by per-document influence within each of the 24 topics, influence-targeted unlearning produces larger accuracy drops than random in-topic unlearning most clearly for **SocialIQA**: across the available matched topic–seed cells, the paired difference $\gamma_{\text{influence}} - \gamma_{\text{random}}$ is significantly positive (Wilcoxon BH-adjusted $p \approx 10^{-5}$; full per-benchmark statistics in Appendix J, Table 49 and Figure 43). **MMLU STEM** is weaker: it reaches significance on the pooled Wilcoxon test, but not on the paired t-test over topic means, and its bootstrap median interval crosses zero. **MMLU Social Sciences** does not reach significance, while **ARC-Challenge** has a negative median paired difference, meaning influence-targeted forgetting is slightly less damaging than random in-topic forgetting under this construction. High-attribution documents thus identify capability-relevant corpus regions, not merely broad topic membership, with SocialIQA as the clearest case.

Appendix J reports the full unlearning suite, including convergence-speed and influence–drop diagnostics, which we treat as supporting evidence rather than separate main claims.

RQ3 validation takeaway

Influence-targeted unlearning reduces **SocialIQA** accuracy more than **same-topic random controls**. On the comparison tasks the paired effect weakens, disappears, or reverses, so attribution does not uniformly isolate functionally necessary documents across benchmarks.

5 Discussion

What generalizes across our settings is a relationship between influence targeting and selective, benchmark-specific damage. Within OLMo3-7B, the knowledge-oriented benchmarks draw concentrated support from documentation-led corpus regions and correlate with one another, while the social probes decompose by probe (Appendix G). *Which* corpus support is functionally necessary, down to which benchmark shows it, is probe-sensitive; the causal tests support the selectivity itself. We accordingly read attribution maps per probe and benchmark, and reserve broader claims for the contrasts that replicate.

6 Conclusions and Future Work

Within our OLMo3-7B/Dolma3 setting, SocialIQA draws positive attribution from a broader, more evenly distributed set of topic$\times$format regions than the comparison tasks, with strong influence from both interactional and expository writing. Influence-targeted unlearning validates this result most clearly for SocialIQA.

A fixed taxonomy turns noisy document-level scores into corpus units a reader can compare or perturb. Influence estimates stay approximate; the taxonomy is what makes corpus-level interpretation more stable than a standalone document ranking.

Direct extensions include repeating the pipeline on a larger model from the same family with an identical unlearning protocol, testing whether causal selectivity and category associations survive scale, and applying it to other fully open pretraining ecosystems to test how capability attribution distributions vary with corpus composition.

Limitations

Training-data attribution estimates influence; it does not prove causal necessity for individual documents, and large web corpora—heterogeneous, source-correlated, and shaped by filtering and upsampling decisions (Soldaini et al., 2024)—strain the assumptions behind influence approximations (Basu et al., 2021; Grosse et al., 2023). Bin-level aggregation mitigates the resulting instability without eliminating it, and the bins are classifier-defined compressions rather than exact semantic partitions (Wettig et al., 2025); we read them as structured summaries and test the strongest findings with targeted unlearning (§3.4).

Our provenance claims are anchored to the de-duplicated Dolma3 population (Soldaini et al., 2024), but the empirical analysis operates on the stratified working set drawn from its WebOrganizer-binned portion (§3). Results characterize benchmark-relevant structure within that working set; they should not be read as exhaustively characterizing every document in the full corpus. Corpus-frequency weighting changes the question from per-bin capability signal to population contribution. A post-hoc token-mass weighting check preserves the sign of each bin's influence but changes which bins dominate population-level totals, because high-volume bins can outweigh lower-volume high-z bins. For example, SocialIQA's strongest balanced bin remains Literature $\times$ Customer Support, while the largest token-weighted positive contributions come from higher-mass technical/tutorial bins. We therefore report the main results as balanced taxonomy-level contrasts and treat corpus-weighted summaries as a separate population-scope diagnostic.

The core causal validation rests on the four primary benchmarks. The held-out probe suite broadens the picture, but it is attribution-first evidence: apart from the single held-out Theory-of-Mind unlearning test (Appendix G.10), the probes carry no matching unlearning validation, and we interpret them only under the tiering policy of Appendix G. Other social-reasoning benchmarks (e.g., ToMi and BigToM (Gandhi et al., 2023)) may reveal different attribution patterns.

The unlearning experiments test whether removing high-attribution topics harms benchmark performance, but they do not identify the mechanisms by which training documents shape model behavior: topic-level forgetting establishes that a corpus region is functionally necessary without decomposing that dependence into features, patterns, or individual examples.

Our deepest corpus-profile analyses use a single open-data ecosystem (OLMo3-7B/Dolma3), for the reason given in §2: document-level TDA requires a released pretraining corpus, and OLMo3 is the best-instrumented family that qualifies. Whether the high-level attribution patterns hold across other open-data models trained on web-scale corpora remains open; at 37,000 H200-equivalent GPU-hours for the present analysis, testing this is a separate follow-up study rather than an incremental experiment.

Author Contributions

Glenn Matlin led the project as first author; Chandreyi Chakraborty was its second-largest contributor. Saehee Eom, Mika Okamoto, and Rayan Castilla contributed equally as joint third authors. Roles follow the CRediT taxonomy.

Glenn Matlin (corresponding): Conceptualization; Data curation; Formal analysis; Funding acquisition; Investigation; Methodology; Project administration; Resources; Software; Supervision; Validation; Visualization; Writing – original draft; Writing – review & editing.

Chandreyi Chakraborty: Conceptualization; Formal analysis; Investigation; Methodology; Project administration; Supervision; Visualization; Writing – original draft; Writing – review & editing. **Saehee Eom**: Data curation; Investigation; Software; Validation; Writing – review & editing. **Mika Okamoto**: Investigation; Methodology; Visualization; Writing – original draft; Writing – review & editing. **Rayan Castilla**: Investigation; Visualization; Writing – review & editing. **Louis Jaburi**: Methodology; Writing – review & editing. **Alvin Deng**: Data curation; Investigation; Methodology; Visualization; Writing – review & editing. **Taywon Min**: Writing – review & editing. **Lucia Quirke**: Resources; Software. **Stella Biderman**: Supervision. **Mark Riedl**: Funding acquisition; Supervision; Writing – review & editing.

Acknowledgments

This research was supported in part through research cyberinfrastructure resources and services provided by the Partnership for an Advanced Computing Environment (PACE) at the Georgia Institute of Technology, Atlanta, Georgia, USA. RRID:SCR_027619. We additionally thank the PACE team for the operational support that enabled the attribution and unlearning experiments in this work.

We are grateful to the ML Alignment & Theory Scholars (MATS) Program for funding and program support throughout this research, and in particular to Ryan Kidd and Lauren Vaughn for their mentorship, logistical coordination, and ongoing encouragement.

Glenn Matlin thanks the Berkeley Existential Risk Initiative (BERI) for funding support of this research, and the Georgia Tech AI Safety Initiative for its support and research community.

We thank Modal Labs and CEO Charles Frye for additional compute support and for providing infrastructure that accelerated the unlearning validation experiments.

Reproducibility Statement

All experiments use the OLMo3-7B/Dolma3 open-data ecosystem (Ettinger et al., 2025): OLMo3-7B Base anchors corpus indexing and unlearning against the de-duplicated 6T Dolma3 training mix (Soldaini et al., 2024) ($\sim$1.26B unique documents), while OLMo3-7B Instruct supplies the benchmark completions and query gradients for attribution scoring. The reproducibility target is the full analysis workflow over the sampled working set (5,678,621 documents, 10.5B tokens across 576 bins), not attribution over the entire upstream population. Attribution uses TrackStar (Chang et al., 2025) via Bergson v0.9.0; unlearning uses NGDiff (Jin et al., 2025) with LoRA (rank 8). The end-to-end pipeline—taxonomy enrichment of $\sim$1.26B documents, gradient index construction, attribution scoring across six benchmarks, and 24-topic unlearning—consumed approximately 37,000 H200 144 GB-equivalent GPU-hours (Appendix L). All pipeline parameters, compute accounting, and storage requirements are fully specified in Appendices C, I, and L.

Following dataset and model documentation norms (Bender & Friedman, 2018; Gebru et al., 2021; Mitchell et al., 2019), we openly release all artifacts, including attribution and aggregation code, sampling scripts, bin-level manifests, the aggregate 576×4 influence matrix, unlearning checkpoints, and analysis notebooks. Because the upstream Dolma3 corpus and OLMo3 weights are publicly available, the full pipeline is reproducible end-to-end. Detailed resource accounting is in Appendix L.

Ethics Statement

This study analyzes the relationship between pretraining data and model capabilities using publicly available resources: the open-weight OLMo3 model (Ettinger et al., 2025) and the openly released Dolma3 corpus (Soldaini et al., 2024). No new data was collected, and no human subjects were involved.

Our methodology operates at the level of aggregate corpus regions (WebOrganizer taxonomy bins) rather than individual documents; like datasheets and model cards (Bender & Friedman, 2018; Gebru et al., 2021; Mitchell et al., 2019), the release discloses structure and withholds raw examples. This reduces the risk of surfacing or amplifying specific training examples. The unlearning experiments modify model behavior through parameter updates and do not alter the underlying training data.

Training-data attribution tools could in principle be used to identify memorized or sensitive content in pretraining corpora (Carlini et al., 2021). We mitigate this by releasing only aggregate bin-level statistics and influence matrices, and by paraphrasing any qualitative examples drawn from high-influence bins. All released artifacts operate at corpus level and do not support individual-document retrieval.

The benchmarks used in this study (SocialIQA, MMLU Social Sciences, MMLU STEM, ARC-Challenge) are established public evaluation resources (Sap et al., 2019; Hendrycks et al., 2021b; Clark et al., 2018), as are the probes in the held-out suite (Appendix G). Our contrastive design probes the structure of capability provenance; it does not evaluate or make claims about the social reasoning abilities of any population or demographic group. In particular, the attribution profiles we report for the bias-related probes (BBQ (Parrish et al., 2022)) are descriptive provenance measurements of where in the corpus those probes draw influence, not a fairness audit of the model; StereoSet (Nadeem et al., 2021) appears only as native-likelihood bias evidence, never as a direct-answer attribution surface. We make no claim about the model's bias levels, and none of our results certify its behavior toward any group.

References

Ekin Akyurek, Tolga Bolukbasi, Frederick Liu, Binbin Xiong, Ian Tenney, Jacob Andreas, and Kelvin Guu. Towards Tracing Knowledge in Language Models Back to the Training Data. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 2429–2446. Association for Computational Linguistics, December 2022. doi: 10.18653/v1/2022.findings-emnlp.180.

Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, Colin Raffel, Shiyu Chang, Tatsunori Hashimoto, and William Yang Wang. A survey on data selection for language models, 2024. URL https://arxiv.org/abs/2402.16827.

Antonis Antoniades, Xinyi Wang, Yanai Elazar, Alfonso Amayuelas, Alon Albalak, Kexun Zhang, and William Yang Wang. Generalization vs. Memorization: Tracing Language Models' Capabilities Back to Pretraining Data. In *ICML 2024 Workshop on Foundation Models in the Wild*, 2024.

Samyadeep Basu, Phil Pope, and Soheil Feizi. Influence Functions in Deep Learning Are Fragile. In *International Conference on Learning Representations*, 2021.

Emily M Bender and Batya Friedman. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018. ISSN 2307-387X. doi: 10.1162/tacl_a_00041.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling, May 2023.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting Training Data from Large Language Models, June 2021.

Chunkit Chan, Cheng Jiayang, Yauwai Yim, Zheye Deng, Wei Fan, Haoran Li, Xin Liu, Hongming Zhang, Weiqi Wang, and Yangqiu Song. NegotiationToM: A benchmark for stress-testing machine theory of mind on negotiation surrounding. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 4211–4241. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.244.

Tyler A. Chang, Dheeraj Rajagopal, Tolga Bolukbasi, Lucas Dixon, and Ian Tenney. Scalable Influence and Fact Tracing for Large Language Model Pretraining. In *The Thirteenth International Conference on Learning Representations*, 2025.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. ToMBench: Benchmarking Theory of Mind in Large Language Models, 2024.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge, March 2018.

Yanai Elazar, Akshita Bhagia, Ian Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, Hanna Hajishirzi, Noah A. Smith, and Jesse Dodge. What's in my big data?, 2024. URL https://arxiv.org/abs/2310.20707.

Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, Pradeep Dasigi, Robert Berry, Saumya Malik, Saurabh Shah, Scott Geng, Shane Arora, Shashank Gupta, Taira Anderson, Teng Xiao, Tyler Murray, Tyler Romero, Victoria Graf, Akari Asai, Akshita Bhagia, Alexander Wettig, Alisa Liu, Aman Rangapur, Chloe Anastasiades, Costa Huang, Dustin Schwenk, Harsh Trivedi, Ian Magnusson, Jaron Lochner, Jiacheng Liu, Lester James V. Miranda, Maarten Sap, Malia Morgan, Michael Schmitz, Michal Guerquin, Michael Wilson, Regan Huff, Ronan Le Bras, Rui Xin, Rulin Shao, Sam Skjonsberg, Shannon Zejiang Shen, Shuyue Stella Li, Tucker Wilde, Valentina Pyatkin, Will Merrill, Yapei Chang, Yuling Gu, Zhiyuan Zeng, Ashish Sabharwal, Luke Zettlemoyer, Pang Wei Koh, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. OLMo 3, 2025.

Chongyu Fan, Jiancheng Liu, Licong Lin, Jinghan Jia, Ruiqi Zhang, Song Mei, and Sijia Liu. Simplicity prevails: Rethinking negative preference optimization for LLM unlearning. In *Advances in Neural Information Processing Systems 38*, 2025.

Ethan Fast, Binbin Chen, and Michael S. Bernstein. Empath: Understanding Topic Signals in Large-Scale Text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 4647–4657, San Jose California USA, May 2016. ACM. ISBN 978-1-4503-3362-7. doi: 10.1145/2858036.2858535.

Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah D. Goodman. Understanding Social Reasoning in Language Models with Language Models, 2023.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for Datasets. *Communications of the ACM*, 64(12):86–92, 2021. ISSN 0001-0782. doi: 10.1145/3458723.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, William Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah Smith, and Hannaneh Hajishirzi. OLMo: Accelerating the Science of Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15789–15809. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.841.

Roger Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, Evan Hubinger, Kamilė Lukošiūtė, Karina Nguyen, Nicholas Joseph, Sam McCandlish, Jared Kaplan, and Samuel R. Bowman. Studying Large Language Model Generalization with Influence Functions, August 2023.

Yuling Gu, Oyvind Tafjord, Hyunwoo Kim, Jared Moore, Ronan Le Bras, Peter Clark, and Yejin Choi. SimpleToM: Exposing the Gap between Explicit ToM Inference and Implicit ToM Application in LLMs, 2024.

Yuling Gu, Oyvind Tafjord, Bailey Kuehl, Dany Haddad, Jesse Dodge, and Hannaneh Hajishirzi. OLMES: A Standard for Language Model Evaluations. In *Findings of the*

Association for Computational Linguistics: NAACL 2025, pp. 5020–5048. Association for Computational Linguistics, April 2025. doi: 10.18653/v1/2025.findings-naacl.282.

Kelvin Guu, Albert Webson, Ellie Pavlick, Lucas Dixon, Ian Tenney, and Tolga Bolukbasi. Simfluence: Modeling the Influence of Individual Training Examples by Simulating Training Runs, 2023.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning ai with shared human values. In *International Conference on Learning Representations*, 2021a. URL https://arxiv.org/abs/2008.02275.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations*, 2021b.

Cathy Jiao, Yijun Pan, Emily Xiao, Daisy Sheng, Niket Jain, Hanzhang Zhao, Ishita Dasgupta, Jiaqi W Ma, and Chenyan Xiong. DATE-LM: Benchmarking Data Attribution Evaluation for Large Language Models, 2025.

Xiaomeng Jin, Zhiqi Bu, Bhanukiran Vinzamuri, Anil Ramakrishna, Kai-Wei Chang, Volkan Cevher, and Mingyi Hong. Unlearning as multi-task optimization: A normalized gradient difference approach with an adaptive learning rate. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 11278–11294. Association for Computational Linguistics, April 2025. doi: 10.18653/v1/2025.naacl-long.563.

Zhijing Jin, Sydney Levine, Fernando Gonzalez Adauto, Ojasv Kamal, Maarten Sap, Mrinmaya Sachan, Rada Mihalcea, Josh Tenenbaum, and Bernhard Schölkopf. When to make exceptions: Exploring language models as accounts of human moral judgment. In *Advances in Neural Information Processing Systems 35*, pp. 28458–28473, 2022. doi: 10.52202/068431-2063.

Ewa Kacewicz, James W. Pennebaker, Matthew Davis, Moongee Jeon, and Arthur C. Graesser. Pronoun use reflects standings in social hierarchies. *Journal of Language and Social Psychology*, 33(2):125–143, September 2013. ISSN 1552-6526. doi: 10.1177/0261927x13502654.

Nikhil Kandpal, Brian Lester, Colin Raffel, Sebastian Majstorovic, Stella Biderman, Baber Abbasi, Luca Soldaini, Enrico Shippole, A. Feder Cooper, Aviya Skowron, John Kirchenbauer, Shayne Longpre, Lintang Sutawika, Alon Albalak, Zhenlin Xu, Guilherme Penedo, Loubna Ben Allal, Elie Bakouch, John David Pressman, Honglu Fan, Dashiell Stander, Guangyu Song, Aaron Gokaslan, Tom Goldstein, Brian R. Bartoldson, Bhavya Kailkhura, and Tyler Murray. The Common Pile v0.1: An 8TB Dataset of Public Domain and Openly Licensed Text, June 2025.

Pang Wei Koh and Percy Liang. Understanding Black-box Predictions via Influence Functions. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 1885–1894. PMLR, July 2017.

Huawei Lin, Jikai Long, Zhaozhuo Xu, and Weijie Zhao. Token-wise Influential Training Data Retrieval for Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 841–860. Association for Computational Linguistics, August 2024. doi: 10.18653/v1/2024.acl-long.48.

Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. Infini-gram: Scaling Unbounded n-gram Language Models to a Trillion Tokens. In *First Conference on Language Modeling*, 2024.

Jiacheng Liu, Taylor Blanton, Yanai Elazar, Sewon Min, Yen-Sung Chen, Arnavi Chheda-Kothary, Huy Tran, Byron Bischoff, Eric Marsh, Michael Schmitz, Cassidy Trier, Aaron Sarnat, Jenna James, Jon Borchardt, Bailey Kuehl, Evie Yu-Yen Cheng, Karen Farley, Taira Anderson, David Albright, Carissa Schoenick, Luca Soldaini, Dirk Groeneveld,

Rock Yuren Pang, Pang Wei Koh, Noah A. Smith, Sophie Lebrecht, Yejin Choi, Hannaneh Hajishirzi, Ali Farhadi, and Jesse Dodge. OLMoTrace: Tracing Language Model Outputs Back to Trillions of Training Tokens. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 178–188. Association for Computational Linguistics, July 2025. doi: 10.18653/v1/2025.acl-demo.18.

Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff, Nathan Khazam, Jad Kabbara, Kartik Perisetla, Xinyi Wu, Enrico Shippole, Kurt Bollacker, Tongshuang Wu, Luis Villa, Sandy Pentland, and Sara Hooker. A large-scale audit of dataset licensing and attribution in AI. *Nature Machine Intelligence*, 6(8):975–987, August 2024. ISSN 2522-5839. doi: 10.1038/s42256-024-00878-8.

Matteo Marcuzzo, Alessandro Zangari, Andrea Albarelli, Jose Camacho-Collados, and Mohammad Taher Pilehvar. Morables: A Benchmark for Assessing Abstract Moral Reasoning in LLMs with Fables. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 27715–27739, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.1411.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 220–229. ACM, 2019. doi: 10.1145/3287560.3287596.

Moin Nadeem, Anna Bethke, and Siva Reddy. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 5356–5371, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.416.

Matthew L. Newman, James W. Pennebaker, Diane S. Berry, and Jane M. Richards. Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin*, 29(5):665–675, May 2003. ISSN 1552-7433. doi: 10.1177/0146167203029005010.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. TRAK: Attributing Model Behavior at Scale. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 27074–27113. PMLR, 2023.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2086–2105. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-acl.165.

James W. Pennebaker, Cindy K. Chung, Joey Frazee, Gary M. Lavergne, and David I. Beaver. When small words foretell academic success: The case of college admissions essays. *PLoS ONE*, 9(12):e115844, December 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0115844.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating Training Data Influence by Tracing Gradient Descent. In *Advances in Neural Information Processing Systems*, volume 33, pp. 19920–19930, 2020.

Lucia Quirke, Nora Belrose, Louis Jaburi, William Li, davidoj, Michael Mulet, darkness8i8, Guillaume Martres, Goncalo Paulo, Dipika Khullar, and neverix. EleutherAI/bergson: V0.9.0. Zenodo, March 2026.

Yasaman Razeghi, Hamish Ivison, Sameer Singh, and Yanai Elazar. Backtracking Mathematical Reasoning of Language Models to the Pretraining Data. In *The Second Tiny Papers Track at ICLR 2024*, 2024.

Laura Ruis, Maximilian Mozes, Juhan Bae, Siddhartha Rao Kamalakara, Dwarak Talupuru, Acyr Locatelli, Robert Kirk, Tim Rocktäschel, Edward Grefenstette, and Max Bartolo. Procedural Knowledge in Pretraining Drives Reasoning in Large Language Models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social IQa: Commonsense Reasoning about Social Interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4463–4473. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1454.

Maarten Sap, Ronan LeBras, Daniel Fried, and Yejin Choi. Neural Theory-of-Mind? On the Limits of Social Intelligence in Large LMs, 2022.

Andrea Schioppa, Polina Zablotskaia, David Vilar, and Artem Sokolov. Scaling Up Influence Functions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 8179–8186, June 2022. doi: 10.1609/aaai.v36i8.20791.

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. Clever Hans or Neural Theory of Mind? Stress Testing Social Reasoning in Large Language Models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2257–2273, St. Julian's, Malta, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.eacl-long.138.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Evan Walsh, Luke Zettlemoyer, Noah Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. Dolma: An Open Corpus of Three Trillion Tokens for Language Model Pretraining Research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15725–15788. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.840.

Settaluri Sravanthi, Meet Doshi, Pavan Tankala, Rudra Murthy, Raj Dabre, and Pushpak Bhattacharyya. PUB: A Pragmatics Understanding Benchmark for Assessing LLMs' Pragmatics Capabilities. In *Findings of the Association for Computational Linguistics ACL 2024*, pp. 12075–12097, Bangkok, Thailand and virtual meeting, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.719.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13003–13051. Association for Computational Linguistics, July 2023. doi: 10.18653/v1/2023.findings-acl.824.

Alexander Wettig, Kyle Lo, Sewon Min, Hannaneh Hajishirzi, Danqi Chen, and Luca Soldaini. Organize the Web: Constructing Domains Enhances Pre-Training Data Curation. In *Proceedings of the 42nd International Conference on Machine Learning*. arXiv, 2025. doi: 10.48550/arXiv.2502.10341.

Yang Zhao, Li Du, Xiao Ding, Kai Xiong, Zhouhao Sun, Shi Jun, Ting Liu, and Bing Qin. Deciphering the Impact of Pretraining Data on Large Language Models through Machine Unlearning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 9386–9406. Association for Computational Linguistics, August 2024. doi: 10.18653/v1/2024.findings-acl.559.

Appendix Contents

A WebOrganizer Taxonomy **18**

B Corpus Characterization **20**
B.1 De-duplication and Provenance . . . 20
B.2 Corpus Distribution . . . 20
B.3 Sampling Strategy . . . 25
B.4 Realized Sample Fill Rates . . . 28
B.5 Document Length Statistics . . . 29
B.6 Influence Score Distributions . . . 29

C Attribution Pipeline Details **32**
C.1 Background: TrackStar and Bergson . . . 32
C.2 Query Construction . . . 32
C.3 Corpus Indexing . . . 32
C.4 Gradient Computation . . . 32
C.5 Preconditioner . . . 33
C.6 Pipeline Architecture . . . 33

D Bin-Level Influence Heatmaps and Format Marginals **33**
D.1 Do the benchmark profiles differ from each other? . . . 34

E Paired Topic Comparisons **38**

F Correctness Differential **40**

G Held-out Probe Analysis for Social Reasoning **42**
G.1 Setup and Query Construction . . . 42
G.2 Direct-Answer Gate and Tiering . . . 43
G.3 Split-Sensitive Masks . . . 43
G.4 Pragmatics and BBH Family . . . 44
G.5 Theory of Mind and Negotiation . . . 47
G.6 Moral, Ethics, and Bias Probes . . . 47
G.7 Attribution Diagnostics and Claim Boundary . . . 56
G.8 Register Robustness of the Block Signature . . . 60
G.9 Probe Geometry . . . 61
G.10 Causal Validation: Held-out ToM Unlearning . . . 61
G.11 Data-Side Levers for Safety-Relevant Capabilities . . . 62

H Bin Characterization **63**

H.1 Lexical Profiling Protocol . . . 63
H.2 Top-20 Lexical Profiles and Format Clusters . . . 64
H.3 OpenLIWC-Style Result Surfaces . . . 65
H.4 Per-Benchmark Top-20 Bin Tables . . . 73

I Unlearning Experimental Setup **84**
I.1 Objective and Optimization . . . 84
I.2 Data Configuration and Sampling . . . 84
I.3 Stopping Criterion: Randomized-Text Perplexity . . . 85
I.4 Experimental Conditions (Validation) . . . 85
I.5 Evaluation Protocol . . . 85

J Unlearning Results **86**
J.1 Paired Significance of the Influence-vs-Random Contrast . . . 86
J.2 Single-Bin Topic Sweep . . . 87
J.3 Global Random Control . . . 89
J.4 Topic-Level Efficiency Analysis . . . 89
J.5 Influence–Drop Diagnostic . . . 89

K Selective Unlearning: Does Taxonomy Decomposition Help? **94**
K.1 Cross-Method Check: Bin-Level vs. Per-Document Forget-Sets . . . 95
K.2 ARC Unlearning Across Task Definitions (Coverage Robustness) . . . 98

L Compute and Resource Accounting **100**
L.1 Taxonomy Enrichment . . . 100
L.2 Attribution Pipeline . . . 100
L.3 Unlearning Experiments . . . 101
L.4 Aggregate Compute Summary . . . 101

M LLM Disclosure **101**

A WebOrganizer Taxonomy

We use the WebOrganizer taxonomy (Wettig et al., 2025), comprising 24 topic categories and 24 format categories. Wettig et al. constructed the taxonomy by reviewing existing web taxonomies (Curlie.org, Google AdSense, Wikipedia ontology) and iteratively refining category definitions: Llama-3.1-405B-Instruct classified CommonCrawl samples, and the authors manually reviewed annotations to refine boundaries. Classification of our deduplicated Dolma3 corpus uses the official trained WebOrganizer classifiers (gte-base-en-v1.5 backbone; see Appendix L.1 for throughput, hardware, and reported accuracy). Topic and format definitions, adapted from the original paper and reproduced here for reference, appear in Tables 2 and 3. Figure 3 summarizes the topic–format cross-product used throughout the bin-level analyses.

Taxonomy unit

Throughout the paper, a WebOrganizer **bin** means one topic×format cell. Bin-level attribution aggregates noisy document scores into these repeated units so the same cells can be compared across **social reasoning**, **social knowledge**, **STEM reasoning**, and **STEM knowledge**.

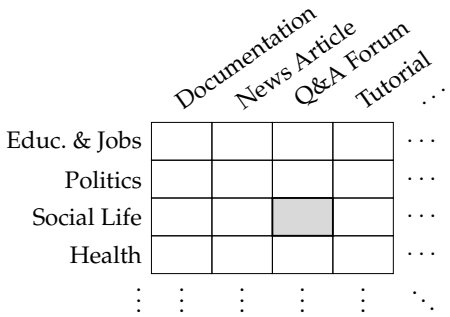


Figure 3: Schematic of the WebOrganizer cross-product taxonomy. Each of 24 topic categories pairs with each of 24 format categories to yield 576 bins; the shaded cell illustrates one example bin (Social Life × Q&A Forum). Empty cells indicate that this figure conveys taxonomy *structure*; bin populations are visualized separately in the joint count panels in Figure 6.

Table 2: WebOrganizer topic categories and descriptions (adapted from Wettig et al. 2025).

Topic	Description
Adult	Adult content
Art & Design	Includes architecture
Crime & Law	Includes law enforcement; financial crime falls under Finance & Business
Education & Jobs	Pedagogy, training, certification, academia
Entertainment	Music, movies, TV, celebrities, humor, nightlife
Fashion & Beauty	Clothing, accessories, cosmetics
Finance & Business	Taxes, investments, insurance, corporate communication, marketing, HR
Food & Dining	Recipes, groceries, beverages, restaurants
Games	Video games, board games, gambling
Hardware	Computer hardware, phones, televisions, consumer electronics
Health	Medicine, wellness, mental health, veterinary science, nutritional science
History	Includes geography, archaeology
Home & Hobbies	Real estate, furniture, home improvement, DIY, gardening, pets
Industrial	Mining, agriculture, manufacturing, utilities, construction
Literature	Literary criticism, linguistics, philosophy, humanities
Politics	Social issues, campaigns, legislative process, geopolitics, activism
Religion	Includes spirituality
Science & Technology	Physics, chemistry, biology, mathematics, engineering
Social Life	Family, friends, relationships, community
Software	Use of software and the internet
Software Development	Algorithms, coding, web development
Sports & Fitness	Martial arts, motor sports, outdoor activities
Transportation	Cars, taxis, public transit, aviation, shipping, logistics
Travel	Hospitality, hotels, sight-seeing, cruises

Table 3: WebOrganizer format categories and descriptions (adapted from Wettig et al. 2025).

Format	Description
About (Organization)	Organizational self-description or introduction
About (Personal)	Personal website self-description or profile
Academic Writing	Research papers, abstracts, theses, literature reviews
Audio Transcript	Written record of spoken language (interviews, lectures, podcasts)
Comment Section	Discussion forum with multiple posts or comments
Content Listing	Navigation page: sitemap, catalog, search results
Creative Writing	Short stories, novel chapters, poetry, song lyrics
Customer Support	Troubleshooting guides for a general audience
Documentation	Technical writing, API docs, README files, source code
FAQ	Frequently Asked Questions format
Knowledge Article	Objective, neutral style; moderated platform or reputable source
Legal Notices	Terms of service, disclaimers, privacy policies, licenses
Listicle	List-format articles ("Top 10" lists, Buzzfeed-style)
News Article	Journalism on current events by news organizations
News (Organization)	Press releases, organizational blog posts
Nonfiction Writing	Long reads, profiles, editorials, essays, memoirs
Personal Blog	Individual writing on personal experiences and opinions
Product Page	Descriptions and promotions for products or services
Q&A Forum	Explicit question-and-answer format (Quora, Stack Exchange)
Spam / Ads	SEO keyword stuffing, short ads, or no apparent purpose
Structured Data	Multiple data entries with common structure (tables, JSON, CSV)
Truncated	Incomplete content: paywalled, login-required, or multimedia descriptions
Tutorial	Recipes, DIY instructions, WikiHow, how-to guides
User Review	User-posted reviews (Yelp, TripAdvisor)

B Corpus Characterization

B.1 De-duplication and Provenance

OLMo3-7B (Ettinger et al., 2025) was trained on the 6T Dolma3 mix (Soldaini et al., 2024) (`allenai/dolma3_mix-6T-1025-7B`), distributed across 65,718 shards. The mix uses quality-aware upsampling, so the same document can appear in multiple shards (rationale and impact on attribution discussed in §3). We de-duplicated at the document level using a Bloom filter over document identifiers (0.1% false-positive rate, 10 hash functions); Table 4 reports raw and de-duplicated shard counts.

Table 4: De-duplication summary for the Dolma3 6T training mix.

Metric	Value
Total shards	65,718
Total document slots (with upsampling)	∼3.87B
Unique documents after de-duplication	1,263,587,168
Duplication rate	67.3%
Bloom filter size	2.27 GB

Source composition. The de-duplicated corpus draws from six source categories (percentages reported by Soldaini et al., 2024): Common Crawl (76%), olmOCR PDFs (14%), Stack-Edu (7%), FineMath (2.6%), arXiv (0.9%), and Wikipedia (0.04%).

B.2 Corpus Distribution

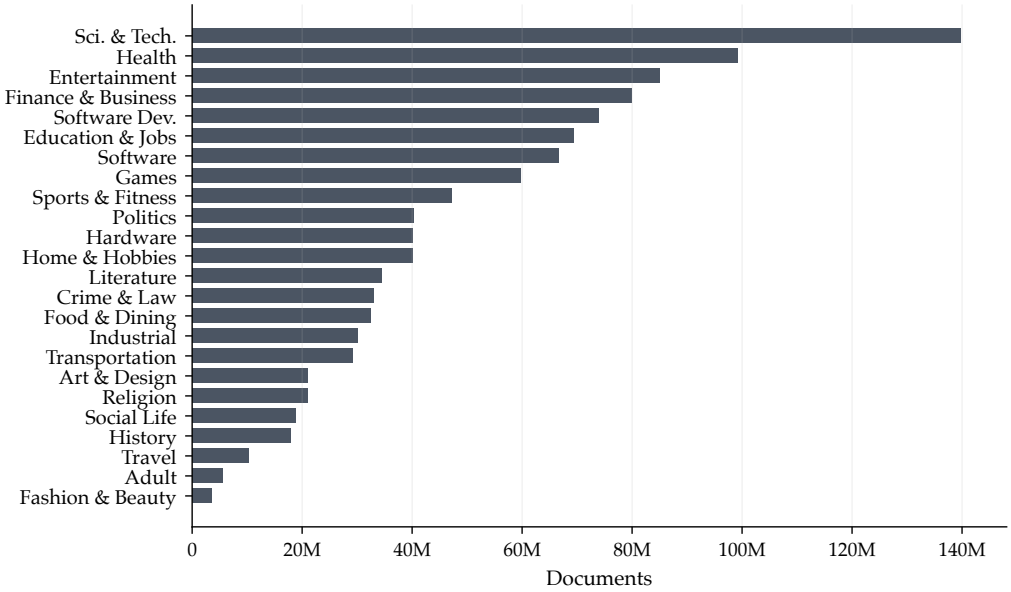
Before sampling, we analyzed all 576 WebOrganizer (Wettig et al., 2025) bins in the binned (Common Crawl + olmOCR) portion of the de-duplicated corpus (1,098,646,162 documents). All bins are populated (zero empty bins), but the distribution is highly skewed: the largest

bin contains over 100,000× more documents than the smallest, motivating our stratified sampling design (§B.3). Table 5 quantifies the distributional skew: the Gini coefficient is 0.68 for documents and 0.73 for tokens, and only ∼236 of 576 bins carry appreciable document mass (effective bins, computed as $\exp(H)$ where H is Shannon entropy).

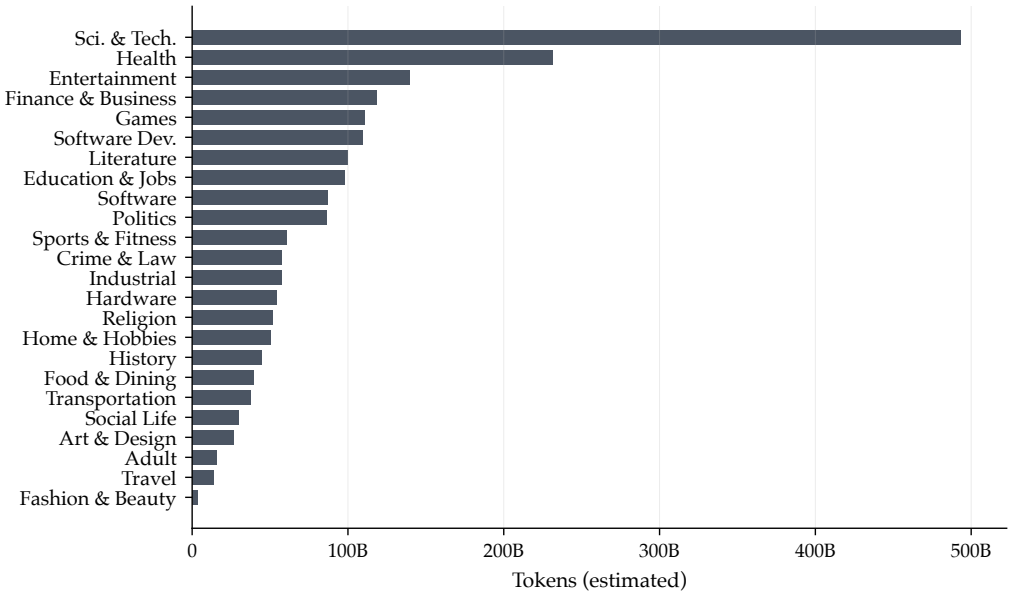
Table 5: Corpus concentration metrics for the de-duplicated Dolma3 training mix across 576 WebOrganizer bins. The Gini coefficient and effective-bin counts quantify the distributional skew that motivates stratified sampling.

Metric	**Value**
Total bins (24 × 24)	576
Empty bins (0 docs)	0
Near-empty bins (<100 docs)	0
Near-empty bins (<1,000 docs)	3
Gini coefficient (documents)	0.6837
Gini coefficient (tokens)	0.7337
Effective bins, documents (exp H)	236.0
Effective bins, tokens (exp H)	158.6
Top-10 bin token share	27.8%
Top-50 bin token share	53.1%
Top-100 bin token share	71.9%

Figures 4, 5, and 6 visualize these properties across the WebOrganizer taxonomy. The marginal distributions reveal substantial concentration: a handful of topics (Science & Technology, Health, Entertainment, Finance & Business) and formats (Product Page, News Article, Nonfiction Writing, Personal Blog) dominate counts, while combinations such as Social Life × Documentation or Adult × Legal Notices contribute orders of magnitude fewer documents. The joint heatmaps expose the full 24×24 structure, showing that skew compounds when topic and format interact. Table 6 contrasts the twenty bins that carry the most token mass with the twenty that carry the least, making the heavy-head/empty-tail asymmetry concrete: the top 20 account for roughly 36% of all training tokens, while the bottom 20 contribute essentially zero.

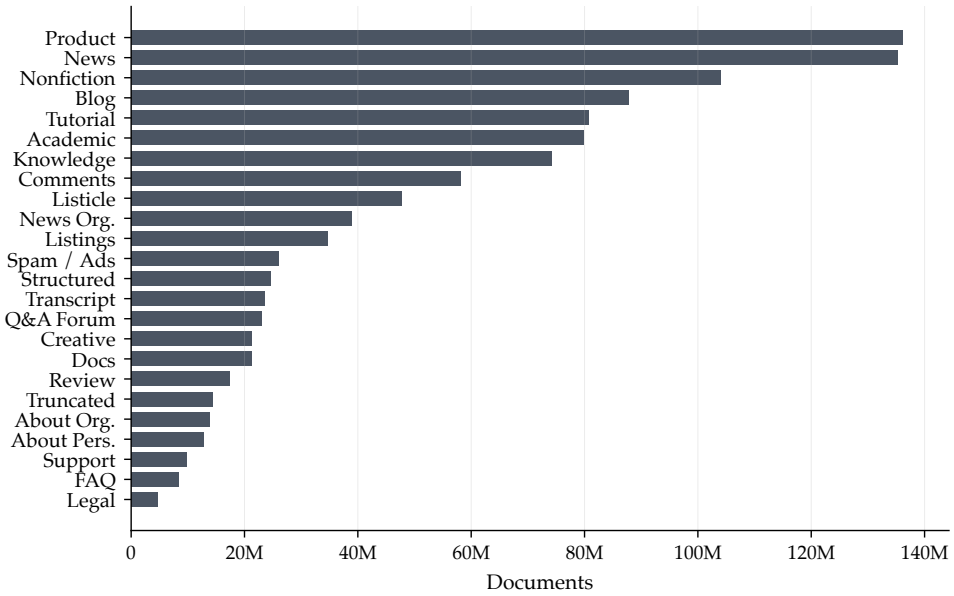


(a) Document count.

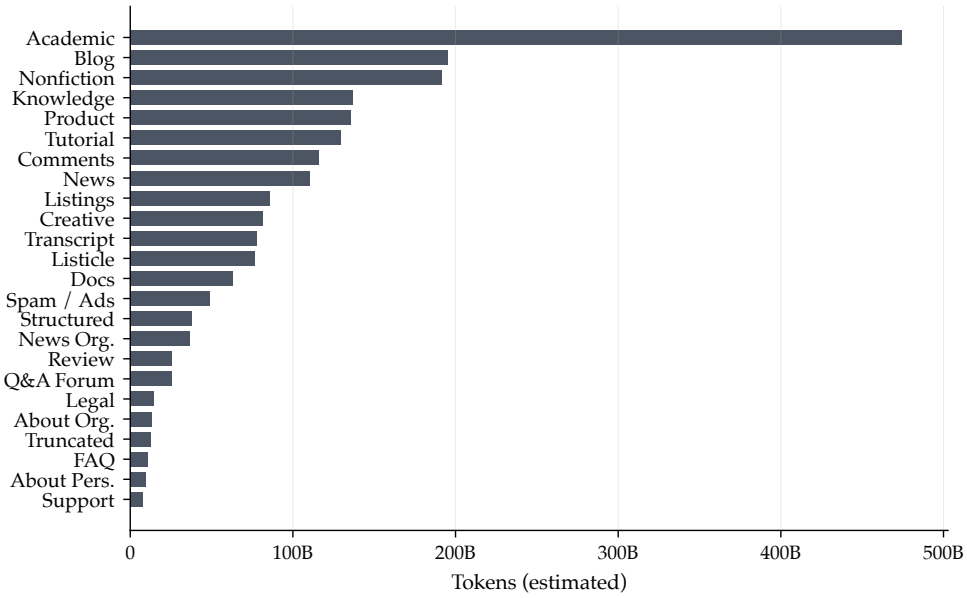


(b) Token count.

Figure 4: Marginal distribution by WebOrganizer topic in the WebOrganizer-binned portion of the de-duplicated Dolma3 corpus (1,098,646,162 documents). Both panels show all 24 topics, sorted by count within each panel. The distributions are highly skewed: *Science & Tech.*, *Health*, *Entertainment*, and *Finance & Business* dominate document counts, while token counts also reflect topic-specific document length.



(a) Document count.



(b) Token count.

Figure 5: Marginal distribution by WebOrganizer format in the WebOrganizer-binned portion of the de-duplicated Dolma3 corpus (1,098,646,162 documents). Both panels show all 24 formats, sorted by count within each panel. *Product Page*, *News Article*, *Nonfiction Writing*, and *Personal Blog* dominate document counts; token counts additionally reflect format-specific length variation.

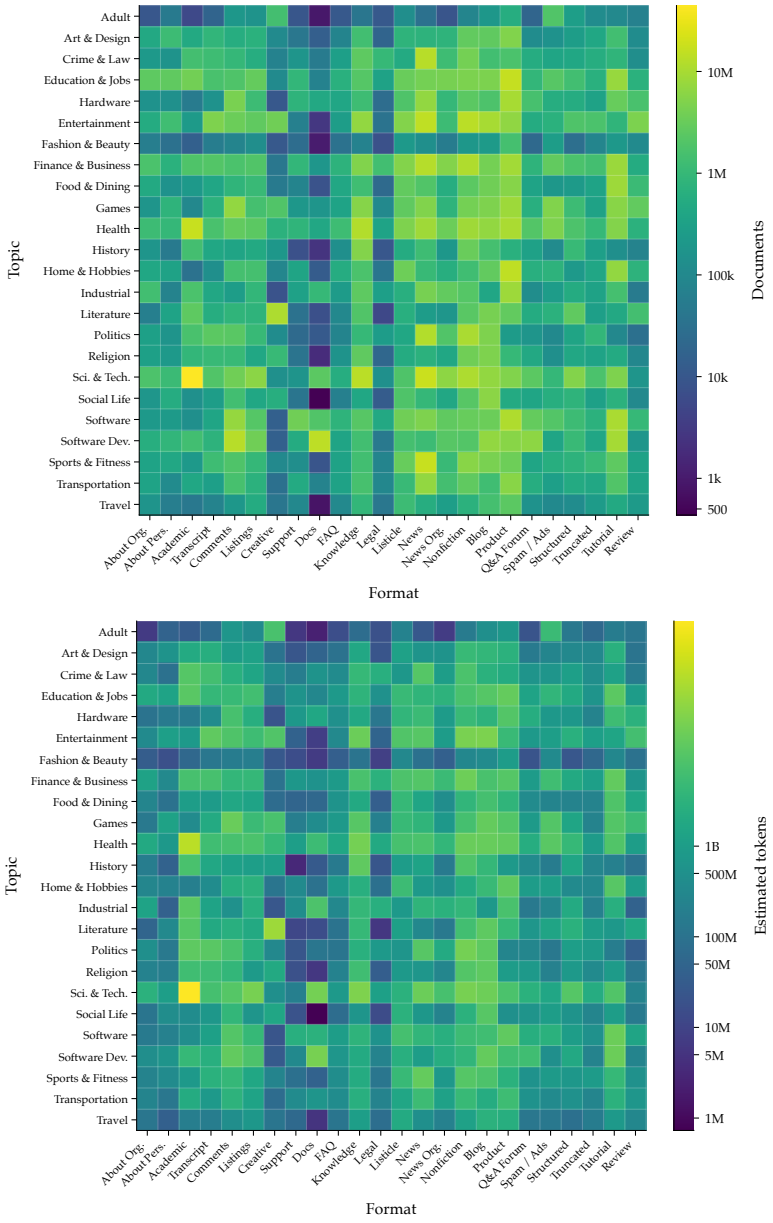


Figure 6: Joint topic–format distribution across the 576 WebOrganizer bins. **(a)** Document counts ($\log_{10}$ scale). **(b)** Token counts ($\log_{10}$ scale). Token counts amplify the skew because longer-form formats (Academic, Docs) carry disproportionate per-document token mass.

Table 6: Top-20 and bottom-20 bins by token count in the de-duplicated Dolma3 corpus. The top 20 bins (of 576) carry ∼36% of all training tokens while the bottom 20 collectively carry essentially zero (<0.01%), exposing the corpus's heavy-head/empty-tail asymmetry. The high-mass bins concentrate in technical and reasoning-aligned topic–format combinations; the empty-tail bins concentrate in documentation and administrative formats (Documentation, Customer Support, Legal Notices) across lower-volume topics. Cf. Table 5 for aggregate concentration metrics.

(a) Top 20 by token mass

Rank	Topic	Format	Docs	Tokens (B)	Share (%)
1	Science & Technology	Academic Writing	44.44M	304.60	14.39
2	Health	Academic Writing	16.57M	81.51	3.85
3	Literature	Creative Writing	10.71M	46.64	2.20
4	Science & Technology	Knowledge Article	13.30M	24.85	1.17
5	Entertainment	Personal Blog	9.63M	23.88	1.13
6	Entertainment	Nonfiction Writing	13.58M	21.99	1.04
7	Software Development	Documentation	14.27M	21.98	1.04
8	Science & Technology	Nonfiction Writing	10.85M	21.62	1.02
9	Science & Technology	Content Listing	5.90M	21.52	1.02
10	Politics	Nonfiction Writing	10.23M	20.52	0.97
11	Health	Knowledge Article	11.82M	20.40	0.96
12	Science & Technology	Documentation	2.47M	19.88	0.94
13	Finance & Business	Nonfiction Writing	11.11M	17.89	0.84
14	Science & Technology	Personal Blog	6.75M	17.20	0.81
15	Entertainment	Knowledge Article	6.94M	16.68	0.79
16	Science & Technology	News Article	17.94M	16.67	0.79
17	Software	Tutorial	10.40M	16.25	0.77
18	Software Development	Tutorial	9.41M	16.25	0.77
19	Games	Comment Section	6.87M	15.98	0.76
20	Health	Nonfiction Writing	8.59M	15.90	0.75

(b) Bottom 20 by token mass

Rank	Topic	Format	Docs	Tokens (B)	Share (%)
557	Religion	Customer Support	37K	0.02	0.00
558	Adult	Legal Notices	11K	0.02	0.00
559	Fashion & Beauty	About (Personal)	32K	0.02	0.00
560	Adult	FAQ	11K	0.02	0.00
561	Literature	Documentation	7.2K	0.02	0.00
562	Fashion & Beauty	Customer Support	27K	0.02	0.00
563	Social Life	Legal Notices	13K	0.01	0.00
564	Literature	Customer Support	34K	0.01	0.00
565	Fashion & Beauty	Legal Notices	7.7K	0.01	0.00
566	Entertainment	Documentation	2.9K	0.01	0.00
567	Adult	News (Organization)	9.0K	0.01	0.00
568	Fashion & Beauty	Documentation	1.1K	0.01	0.00
569	Adult	About (Organization)	10K	0.01	0.00
570	Adult	Customer Support	9.5K	0.01	0.00
571	Literature	Legal Notices	5.4K	0.01	0.00
572	Religion	Documentation	1.9K	0.01	0.00
573	Travel	Documentation	816	0.01	0.00
574	History	Customer Support	7.7K	0.00	0.00
575	Adult	Documentation	935	0.00	0.00
576	Social Life	Documentation	436	0.00	0.00

B.3 Sampling Strategy

To mitigate the distributional skew documented above, we employ stratified sampling with a fixed per-bin budget. This design is not intended to preserve the natural Dolma3 topic–format proportions. Instead, it makes a conditional capability-provenance estimate possible: given a corpus region, which capabilities does that region support or interfere

with? A proportional sample at the same working-set budget would be dominated by high-mass corpus regions and would leave many low-volume cells too sparse for stable influence-function estimates. In the manifest-backed representative draw used for the diagnostics below, all 576 bins are represented, but 426 fall below the 10,000-document-per-bin working-set target. Populating the smallest cells proportionally would approach sampling the whole corpus, which is computationally infeasible for this attribution pipeline.

Figures 7 and 8 visualize this trade-off. The stratified allocation puts taxonomy cells on comparable footing so that small but semantically meaningful regions can be measured; the representative allocation answers a different question, namely which corpus regions dominate total attribution mass under the natural corpus distribution.

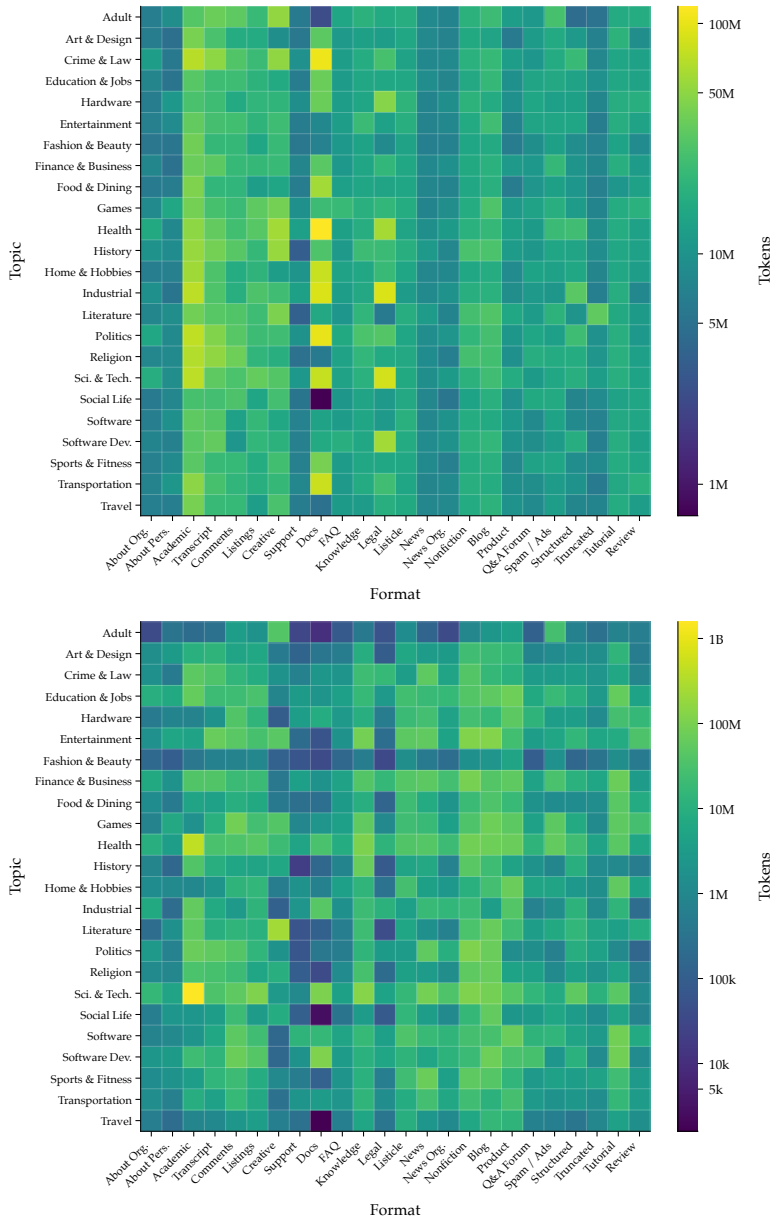


Figure 7: Stratified versus representative sampling across WebOrganizer bins (token counts, $\log_{10}$ scale). **(a)** The stratified working set gives each populated taxonomy cell comparable measurement budget. **(b)** A manifest-backed representative draw with the same 5,678,621-document budget follows the natural corpus skew: all 576 bins are covered, but 426 fall below the 10,000-document-per-bin working-set target.

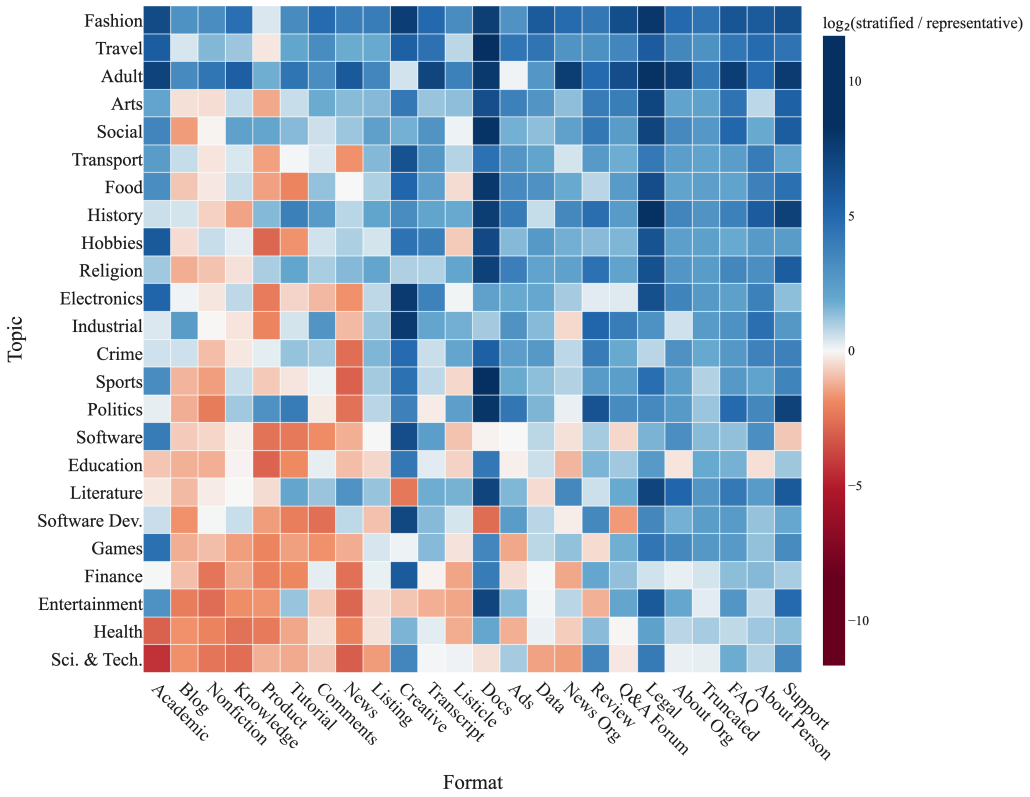


Figure 8: Difference between stratified and representative sampling allocations ($\log_2$ share ratio). Positive values indicate bins up-sampled by stratification; negative values indicate bins down-sampled relative to their corpus proportion. The representative side is a manifest-backed draw with the same document budget as the stratified working set; cells with zero share in both samples render at the midpoint of the diverging scale.

B.4 Realized Sample Fill Rates

The stratified sampling design targets 10,000 documents per bin. Table 7 summarizes realized fill rates: 17 of 576 bins are underfilled, concentrated in the Documentation and Legal Notices format columns where the upstream Dolma3 population simply contains fewer than 10,000 documents for those topic combinations. All underfilled bins still contribute to analysis; their influence estimates are noisier due to smaller sample sizes but are not excluded.

Table 7: Realized bin fill rates for the 10,000-document-per-bin stratified sample across 576 WebOrganizer bins.

Category	**Bins**
Full (10,000 docs realized)	559
Near-full (5,000–9,999 docs)	10
Partial (1,000–4,999 docs)	4
Sparse (<1,000 docs)	3
Total bins	**576**
Underfilled (below target)	17
Smallest realized bin	Social Life / Documentation (436 docs)

B.5 Document Length Statistics

Document length varies substantially across the WebOrganizer taxonomy and may confound attribution patterns if longer documents accumulate more gradient signal. Figure 9 shows the ECDF of document word counts across the de-duplicated corpus; the joint topic–format heatmap in Figure 10 reveals that mean document length varies by more than an order of magnitude across taxonomy cells, motivating per-document normalization in the attribution pipeline.

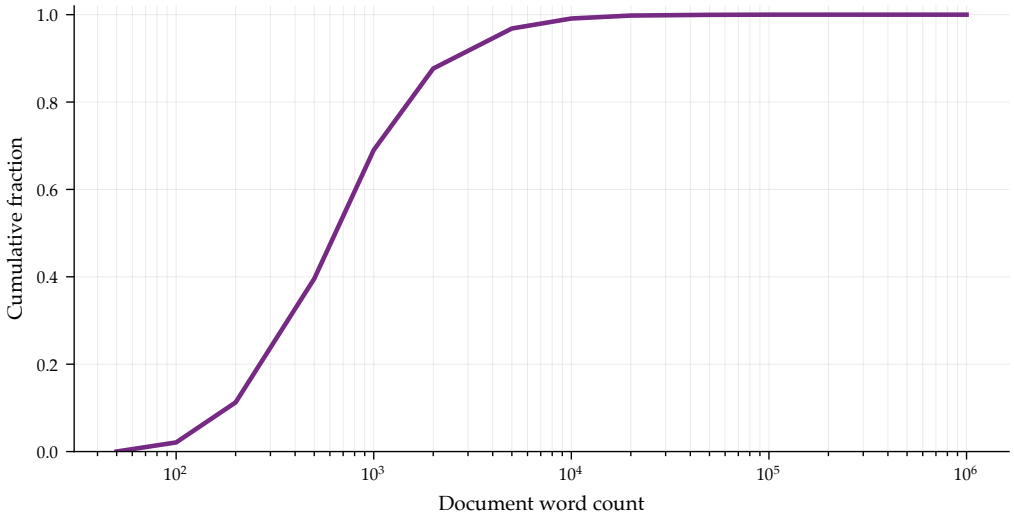


Figure 9: Empirical cumulative distribution (ECDF) of document word counts over an 86.7M-document EDA sample of the de-duplicated Dolma3 corpus, plotted on a log-scaled x-axis. The curve's steep climb between 100 and 2,000 words confirms that most documents are short, while the remaining upper tail stretches to 10^5 words, motivating per-document normalization in the attribution pipeline (§3).

B.6 Influence Score Distributions

Figure 11 characterizes the influence score distribution across the four primary benchmarks before any aggregation, confirming approximate symmetry around zero with benchmark-specific tail behavior. Topic-level cross-benchmark summaries appear in Appendix D.

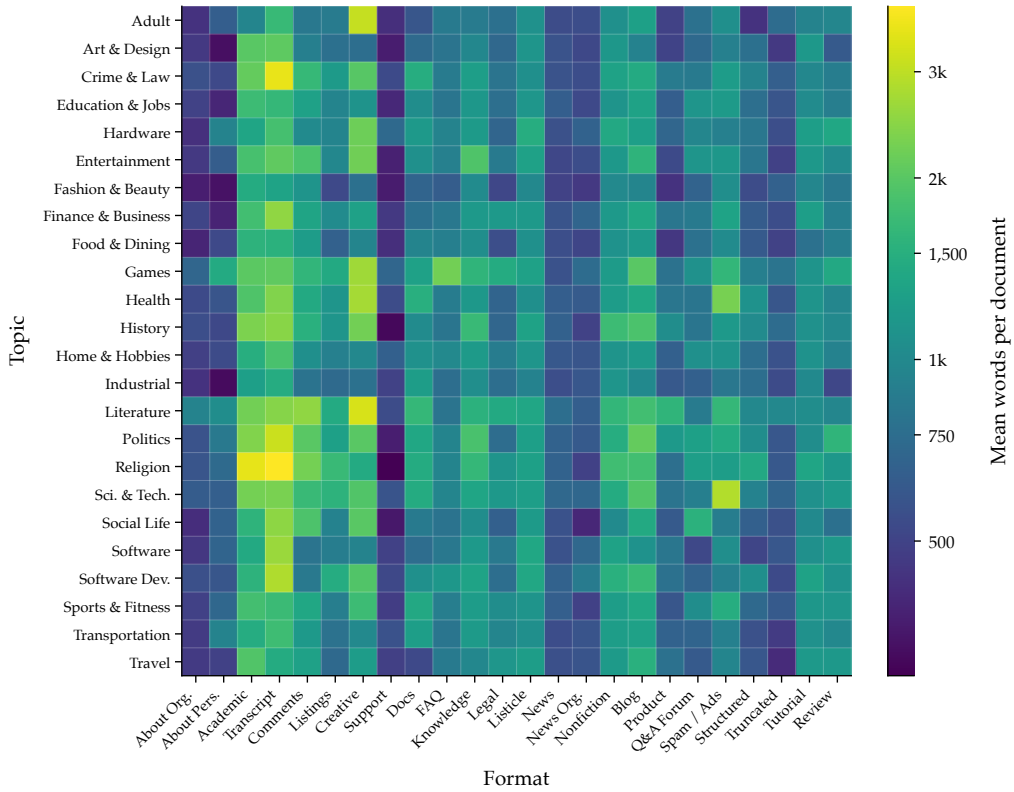


Figure 10: Joint topic–format heatmap of mean document word count on a log color scale, computed over the same 86.7M-document EDA sample as Figure 9. Colorbar ticks report mean words per document. Length variation across bins exceeds an order of magnitude, motivating per-document normalization in the attribution pipeline.

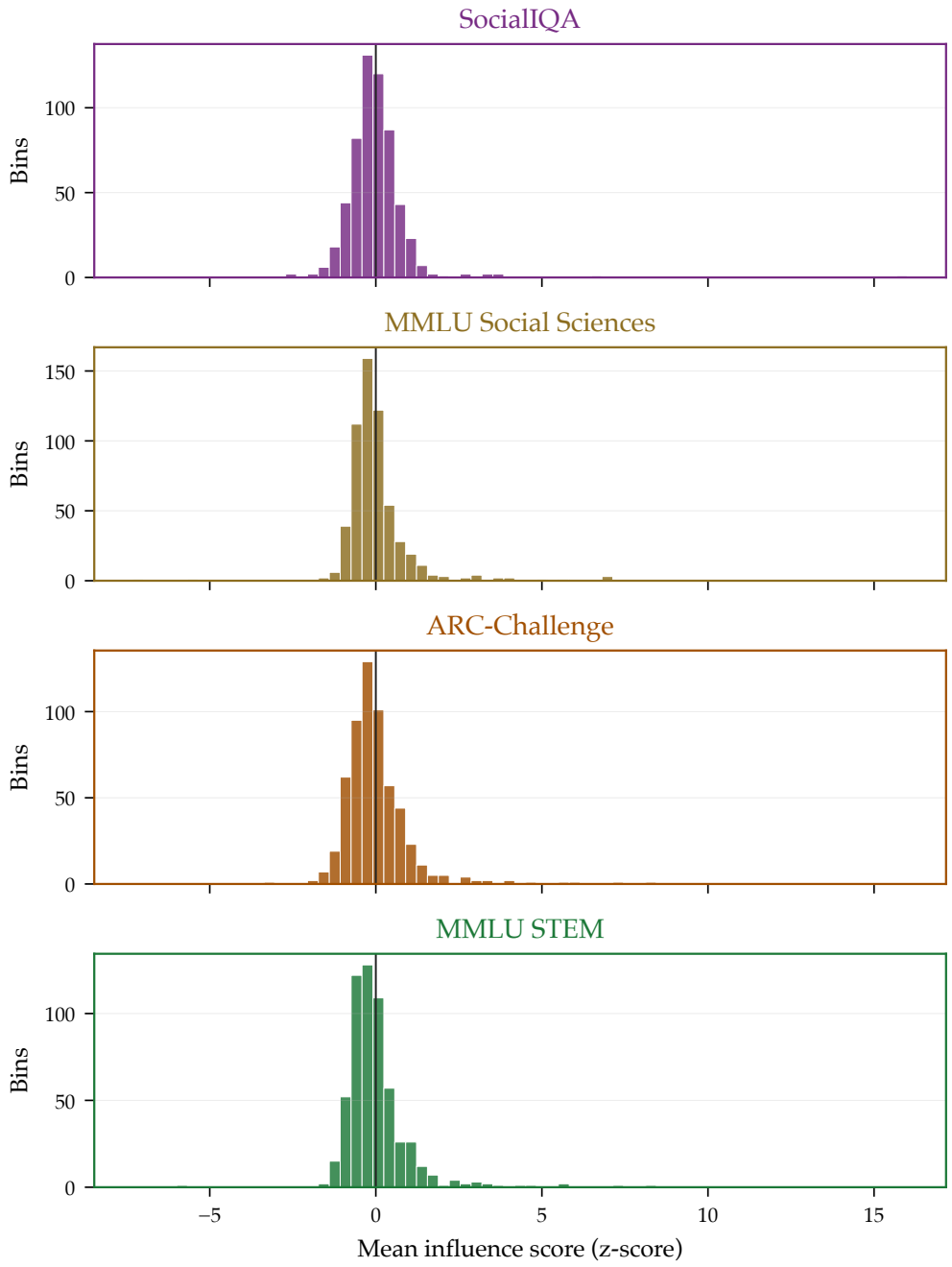


Figure 11: Per-benchmark influence score distributions across all 576 WebOrganizer bins. Distributions are centered near zero with comparable spread, while benchmark-specific behavior appears in the tails.

C Attribution Pipeline Details

C.1 Background: TrackStar and Bergson

We compute gradient-based training-data attribution following TrackStar (Chang et al., 2025) as implemented in Bergson. TrackStar projects per-example gradients into a low-dimensional space using Rademacher random projections and computes influence as

$$s(j,i) = \sum_{k \in \mathcal{K}} \hat{g}_k(x_j)^\top \hat{g}_k(q_i),$$

where $\hat{g}_k(\cdot)$ is the unit-normalized projected gradient at module k. §C.5 describes the mixed preconditioner used in practice; the unpreconditioned form above is retained only for exploratory runs.

C.2 Query Construction

Table 8: Core benchmarks used as attribution queries in the main design.

Benchmark	Queries	Evaluation format	Query type
SocialIQA (Sap et al., 2019)	10,000	3-shot MC, Accuracy	**Social reasoning**
MMLU Social Sciences (Hendrycks et al., 2021b)	3,077	4-shot MC, Accuracy	**Social knowledge**
ARC-Challenge (Clark et al., 2018)	1,172	4-shot MC, Accuracy	**STEM reasoning**
MMLU STEM (Hendrycks et al., 2021b)	3,018	4-shot MC, Accuracy	**STEM knowledge**

Each benchmark sample is an attribution query: it asks which training documents most influence the model's behavior on that instance. We evaluate with OLMES (Gu et al., 2025) following the OLMo3 base evaluation suite: SocialIQA, MMLU Social Sciences, ARC-Challenge, and MMLU STEM are scored as multiple-choice accuracy, and MMLU subscores are macro-averaged. OLMES outputs are converted to Bergson-compatible JSONL records (query and correct completion), so the recorded gradients reflect the true cross-entropy loss. The 10,000 SocialIQA queries come from the OLMES Hugging Face `train` split used for evaluation packaging, not from the standard SocialIQA validation partition.

C.3 Corpus Indexing

Corpus characterization (de-duplication, distribution, fill rates) is in Appendix B. For attribution, each document is stored as a JSONL record with two fields—`id` (stable corpus identifier) and `text` (document text)—sharded round-robin across 65,718 files for corpus storage. The stratified attribution working set is re-sharded into ∼316 shard files of 17,996 documents each for gradient indexing (Appendix L).

C.4 Gradient Computation

For each document, we compute per-sample gradients of cross-entropy loss with respect to all linear-layer weights (Q/K/V/O and MLP projections); Table 9 lists the projection dimension, precision, and token-batching parameters used at index build time.

Table 9: Gradient computation parameters.

Parameter	Value
Projection	Rademacher random
Dimension	d=16 (256 per module)
Precision	fp32
Token batching	4,096 tokens
Truncation	Enabled
Normalization	Unit norm

C.5 Preconditioner

We use a two-model attribution setup following Ruis et al. (2025): document gradients and base-side second-order information are computed with OLMo3-7B Base, while query gradients are computed with OLMo3-7B Instruct, the checkpoint used to score benchmark completions. Following that setup, we treat the supervised-instruction-tuning second-order contribution as identity and attribute only to pretraining data. This approximation keeps the metric tied to pretraining curvature rather than adding an unobserved post-training curvature term. The TrackStar preconditioner therefore supplies the base-curvature metric in which base-document gradients and instruct-query gradients are compared:

$$s(j,i) = \sum_k \tilde{g}_k^{\text{base}}(x_j)^\top \tilde{g}_k^{\text{inst}}(q_i),$$

where $\tilde{g}$ denotes projected, unit-normalized, preconditioned gradients for module k.

The mixed preconditioner is constructed in three steps:

1. Build a value preconditioner from 100k random training documents.
2. Build a query preconditioner from evaluation queries.
3. Combine with 1,000-component downweighting.

Exploratory runs skip preconditioning; all reported results include it. We also computed Base-query variants as calibration checks; the reported figures and tables use the Base-document/Instruct-query setup because the Instruct checkpoint follows the OLMES prompts more reliably.

C.6 Pipeline Architecture

The pipeline consists of three stages:

1. **Reduce**: compute the mean query gradient per benchmark.
2. **Score**: compute document influence scores against the prebuilt gradient index.
3. **Aggregate**: collect and rank the top-k documents.

GPU work is concentrated in index construction: one build job per working-set shard ($\sim$316 shards of 17,996 documents each), plus per-benchmark query-gradient builds. Once the index is built, scoring a query set runs as a parallel CPU pass. Compute accounting and hardware are in Appendix L.

D Bin-Level Influence Heatmaps and Format Marginals

Figure 12 shows the full signed bin-level influence heatmaps for all four benchmarks on a shared color scale; the main text (Figure 2) summarizes their most diagnostic format and topic contrasts. Figures 13 and 14 show the topic and format marginals that summarize these heatmaps. Table 10 names the top-3 positive and top-3 negative formats per benchmark.

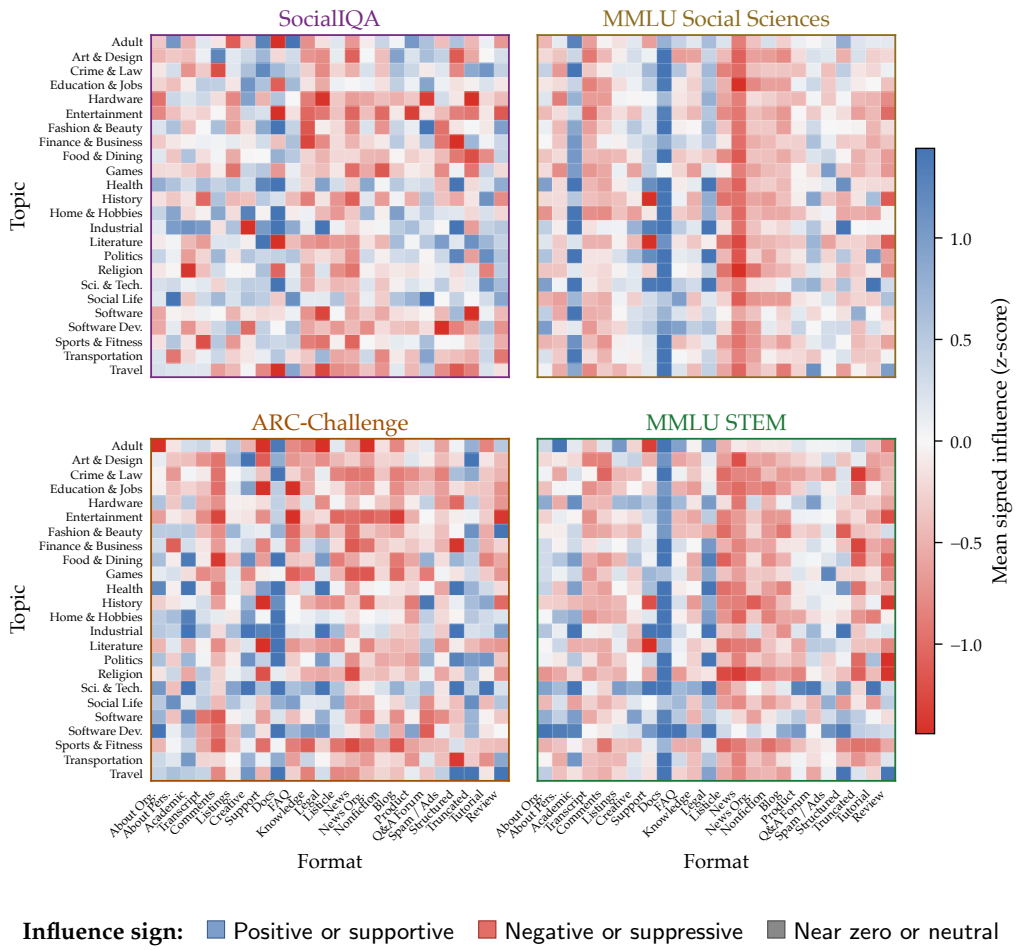


Figure 12: Signed mean influence (z-scores) by topic (rows) and format (columns). Top row: social benchmarks; bottom row: STEM benchmarks. The shared 95th-percentile color scale uses **blue for positive influence**, white near zero, and **red for negative influence**; more extreme cells saturate. **SocialIQA**'s positive mass is dispersed across many topic–format cells, whereas the other three benchmarks—MMLU Social Sciences most visibly—concentrate positive influence in documentation-like formats, giving them a stronger column (format) structure.

Documentation is the strongest positive format for MMLU Social Sciences (z=+2.40), MMLU STEM (+2.29), and ARC-Challenge (+1.97), and remains positive for SocialIQA (+0.44). This cross-benchmark consistency parallels the topic-level positivity of Industrial in Figure 13. SocialIQA diverges at the peak: *Customer Support* (+1.07) is its top format and does not appear in any other benchmark's top-3, mirroring the dialogue-rich character of SocialIQA's top-influence bins (Appendix H). *News Article* registers as a consistent negative across all four benchmarks (−0.62 to −1.02), distinct from the topic-level Literature flip. Substantively, the documentation-like formats the knowledge and STEM benchmarks rely on carry more direct factual content, whereas SocialIQA's interpersonal formats (Customer Support, FAQ, Q&A Forum) carry dialogue-rich social interaction.

D.1 Do the benchmark profiles differ from each other?

The per-bin tests above ask whether a bin's influence differs from zero within a benchmark. A complementary question is whether the four profiles differ *from each other*. Because influence is z-standardized within each benchmark, the marginal score distributions are

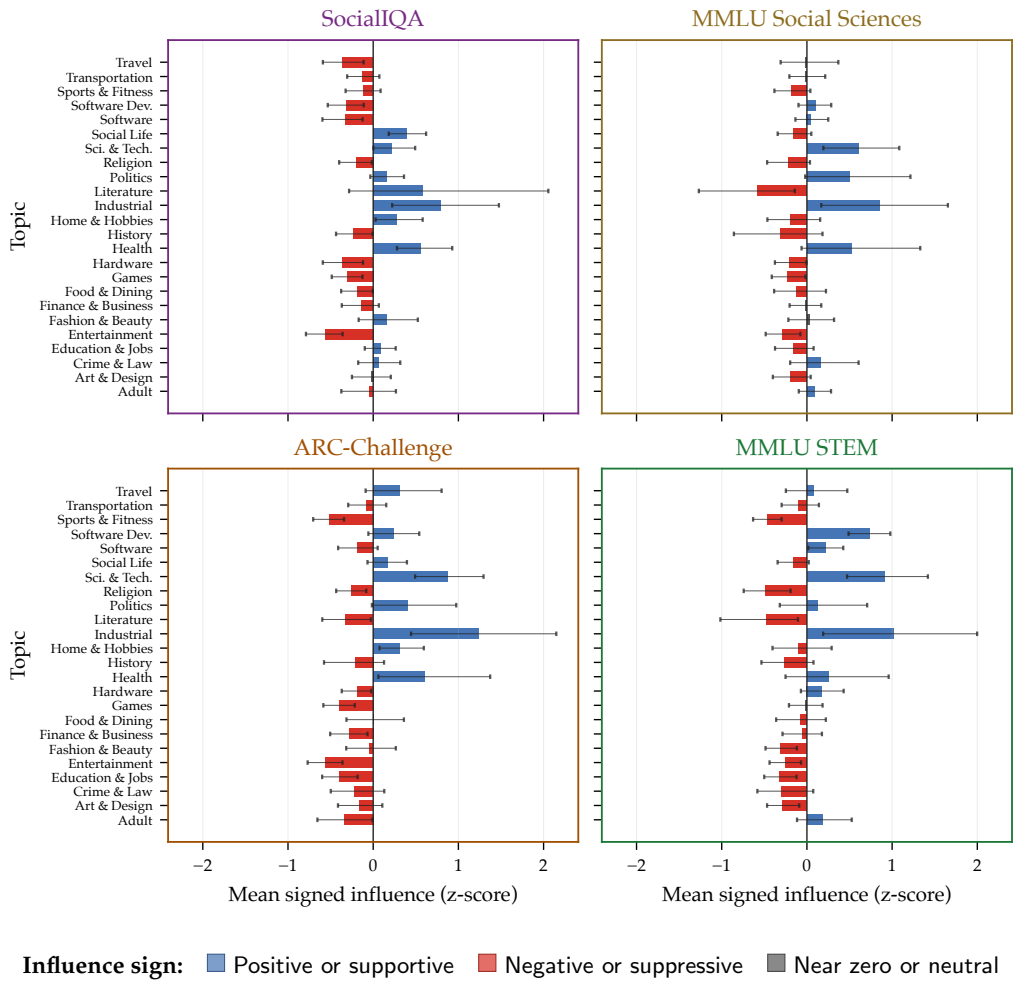


Figure 13: Mean signed influence by WebOrganizer topic, aggregated across formats. Values are z-scores; bars are colored by influence sign, with positive values indicating supportive influence and negative values indicating suppressive influence.

identical by construction, so we compare *positions* rather than magnitudes, treating each benchmark's profile as a vector over the 576 bins and measuring how much two benchmarks concentrate influence on the same bins (Pearson correlation; Figure 15).

SocialIQA's full-grid profile correlates with the three comparison benchmarks at $r = 0.06$, 0.21, and 0.09, whereas those three correlate at $r = 0.76$–0.86 among themselves. An outlier statistic Δ_b, defined as the mean correlation among the other three benchmarks minus benchmark b's mean correlation to them, is positive only for SocialIQA at $\Delta = +0.68$ (95% bin-bootstrap CI $[+0.34, +0.96]$), with a within-bin label-permutation null giving $p \approx 10^{-4}$. The same ranking holds independently on the format marginals ($\Delta = +0.70$, $[+0.35, +1.07]$) and the topic marginals ($+0.37$, $[+0.13, +0.88]$), so SocialIQA's status as the structural outlier does not rest on any single format. Among the cited marginals, the SocialIQA-minus-comparison-mean gap is CI-separated for Documentation (-1.78, $[-2.43, -1.15]$), Customer Support ($+1.34$, $[+0.18, +3.32]$), and Social Life ($+0.45$, $[+0.19, +0.71]$); the Literature gap is positive but its CI includes zero ($+1.04$, $[-0.13, +2.98]$), as do the Creative Writing and Transportation negative checks.

What is resampled, and the independence assumption. All intervals are nonparametric (10,000 resamples, fixed seed) and take the z-scored bin as the unit of analysis. The full-grid

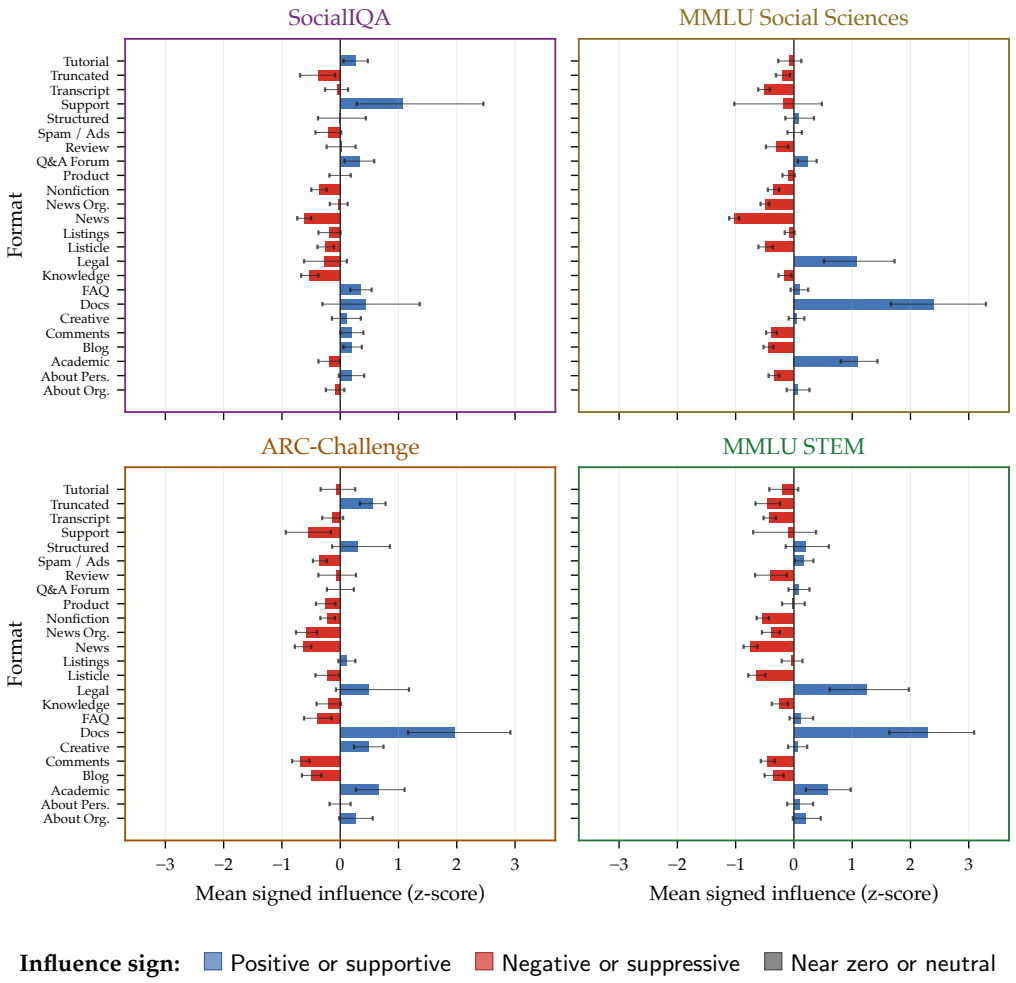


Figure 14: Mean signed influence by WebOrganizer format, aggregated across topics. Values are within-benchmark z-scores; bars are colored by influence sign. Documentation-like formats dominate the STEM and knowledge benchmarks, while SocialIQA peaks on more dialogue-rich formats.

Table 10: Top-3 positive and top-3 negative WebOrganizer formats per benchmark, ranked by mean signed influence z-score (averaged across topics within each format). Values are derived from the same z-scored bin scores that drive Figure 14. *Documentation* appears in every benchmark's top-3; SocialIQA's top format *Customer Support* does not appear in any other benchmark's top-3; *News Article* appears in all four bottom-3 lists.

Benchmark	Sign	Top-3 formats
SocialIQA	Positive	+1.07 Customer Support +0.44 Documentation +0.35 FAQ
	Negative	−0.38 Truncated −0.53 Knowledge Article −0.62 News Article
MMLU Social Sciences	Positive	+2.40 Documentation +1.10 Academic Writing +1.08 Legal Notices
	Negative	−0.49 News (Organization) −0.51 Audio Transcript −1.02 News Article
ARC-Challenge	Positive	+1.97 Documentation +0.66 Academic Writing +0.56 Truncated
	Negative	−0.58 News (Organization) −0.63 News Article −0.68 Comment Section
MMLU STEM	Positive	+2.29 Documentation +1.24 Legal Notices +0.57 Academic Writing
	Negative	−0.54 Nonfiction Writing −0.64 Listicle −0.75 News Article

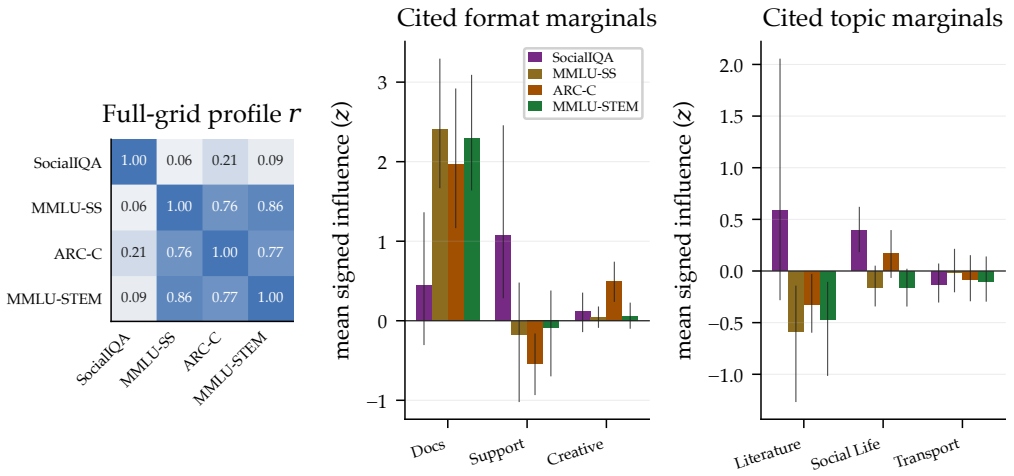


Figure 15: Between-benchmark profile divergence shows **SocialIQA** is the structural outlier independent of any single format. Each benchmark's bin-level influence is a vector over the 576 (topic×format) bins; we compare *positions* (which bins carry influence), not magnitudes, since scores are z-standardized within each benchmark. **Left:** pairwise Pearson correlation of the full-grid profiles, where SocialIQA correlates with each comparison benchmark at $r \leq 0.21$ while the three comparison benchmarks correlate at $r = 0.76$–0.86 among themselves. **Center, Right:** the format and topic marginals the main text cites, with 95% bin-bootstrap CIs. SocialIQA separates from the comparison set on Documentation, Customer Support, and Social Life; the Literature gap is positive but not CI-separated, and the Creative Writing and Transportation negative checks do not separate.

quantities (the profile correlations, c_b, and Δ_b) use a *bin* bootstrap: each replicate draws the N=576 topic×format bins with replacement and applies the *same* resampled set to all four benchmarks before recomputing the 4×4 correlation matrix, so the cross-benchmark

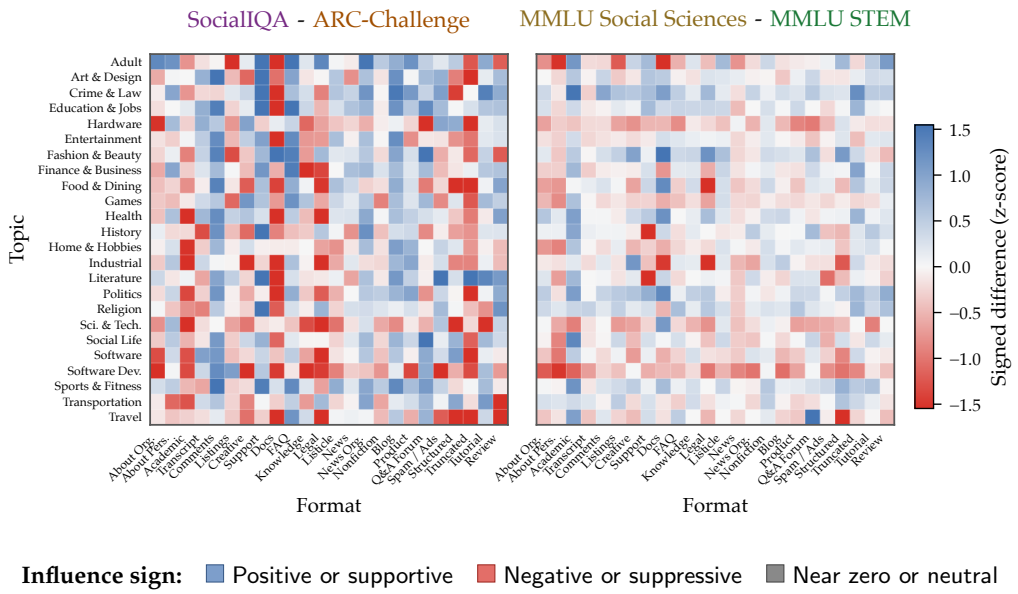


Figure 16: Bin-level social-vs-STEM separation. Signed differences compare **SocialIQA − ARC-Challenge** and **MMLU Social Sciences − MMLU STEM** over WebOrganizer topic–format bins. The shared 95th-percentile color scale uses **blue for positive differences favoring the social benchmark** and **red for negative differences favoring the STEM benchmark**; more extreme cells saturate.

pairing is preserved and the interval reflects sensitivity to bin composition. The marginal intervals instead resample the *collapsed* axis: the format-marginal CIs draw the 24 topics (the axis averaged over to form a format profile) and the topic-marginal CIs draw the 24 formats; the cited single-cell gap CIs (e.g. Documentation) likewise resample the 24 topics or formats composing that marginal. The omnibus p-value is a separate *benchmark-label* permutation that assumes only exchangeability, not independence: within each bin we independently permute the four benchmark labels (reassigning that bin's four standardized values across benchmarks), recompute $\Delta_{\text{SocialIQA}}$, and report $p = (1 + \#\{\Delta^{\text{perm}} \geq \Delta^{\text{obs}}\})/(1 + 10{,}000)$. The bin bootstrap treats bins as exchangeable resampling units and does *not* model the topic×format dependence among them, which could make the full-grid interval mildly anti-conservative; we therefore rely on the agreement of three schemes resting on different assumptions: the bin bootstrap, the topic- and format-resampled marginal bootstraps, and the label permutation. We use bin-level rather than document-level resampling deliberately: with $\sim 10^4$ documents per bin a document bootstrap would render almost any difference significant, so bin composition is the binding, conservative source of uncertainty, the effect sizes carry the claim, and the permutation p is confirmatory.

E Paired Topic Comparisons

Figure 16 provides the full bin-level social-vs-STEM difference heatmaps. Figures 17 and 18 provide direct side-by-side comparisons of per-topic influence between contrastive benchmark pairs. Table 11 names the three topics most strongly tilting toward each benchmark in each pair.

Software Development is consistently STEM-leaning in both contrasts ($\Delta = -0.56$ for SocialIQA vs. ARC-Challenge and $\Delta = -0.63$, the most STEM-leaning topic, for MMLU Social Sciences vs. MMLU STEM); in the reasoning contrast the most STEM-leaning topics are *Travel* (-0.67) and *Science & Technology* (-0.66). *Literature* is the strongest social-leaning topic in the reasoning contrast ($\Delta = +0.91$, sign-flipped between the two benchmarks); the knowledge contrast has no single dominant social-leaning topic of comparable magnitude—*Crime & Law*

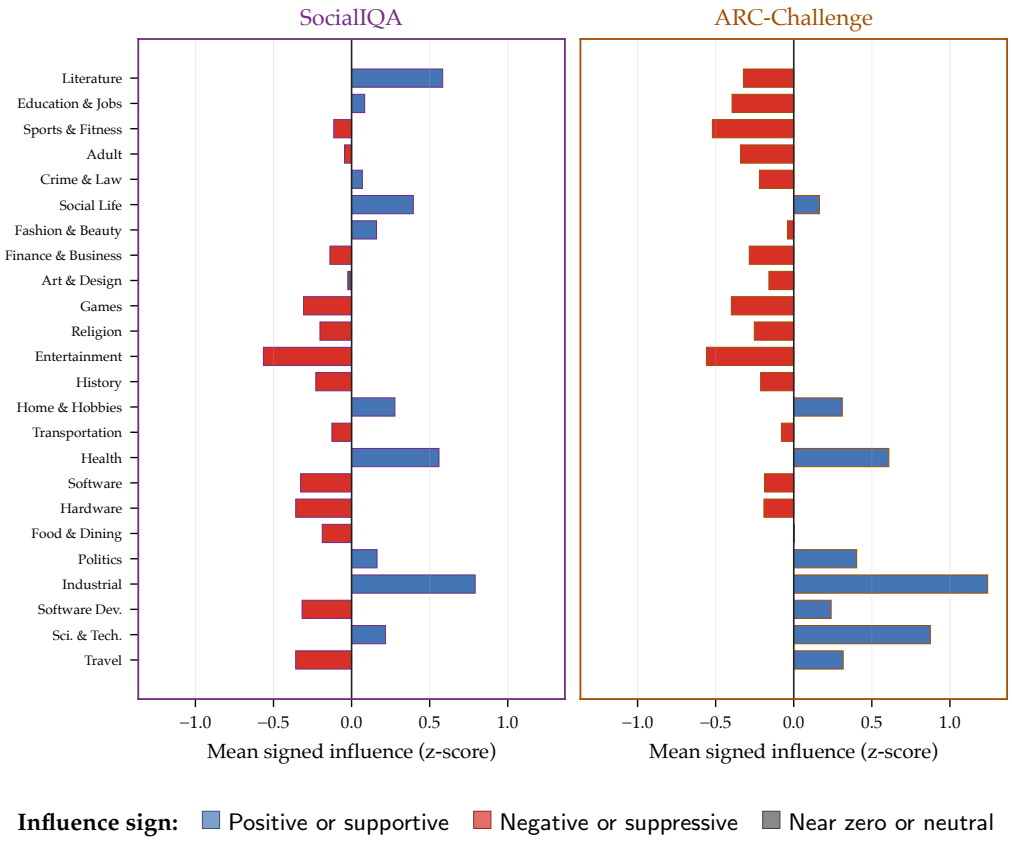


Figure 17: Paired topic-level influence comparison: SocialIQA vs. ARC-Challenge. Each topic shows signed mean influence for both benchmarks on matched topic axes. Topics where signs diverge (e.g., Literature, Software Development) represent the strongest contrastive signal between social and STEM reasoning.

Table 11: Top-3 topics tilting toward each side of the two contrastive benchmark pairs, ranked by the difference $\Delta = z_A - z_B$ of topic-level marginal influence z-scores. Reasoning contrasts (SocialIQA vs. ARC-Challenge) span $|\Delta|$ up to 0.91, while knowledge contrasts (MMLU Social Sciences vs. MMLU STEM) span $|\Delta|$ up to 0.63—quantifying the attenuation noted in the main text (§4). Positive and negative columns follow the signed-difference convention; *Software Development* is consistently STEM-leaning in both pairs.

Contrast pair	**Top-3 tilting A** positive Δ	**Top-3 tilting B** negative Δ
SocialIQA vs. **ARC-Challenge**	Literature (+0.91) Education & Jobs (+0.48) Sports & Fitness (+0.41)	Travel (−0.67) Science & Technology (−0.66) Software Development (−0.56)
MMLU Social Sciences vs. **MMLU STEM**	Crime & Law (+0.46) Politics (+0.38) Fashion & Beauty (+0.34)	Software Development (−0.63) Hardware (−0.38) Science & Technology (−0.31)

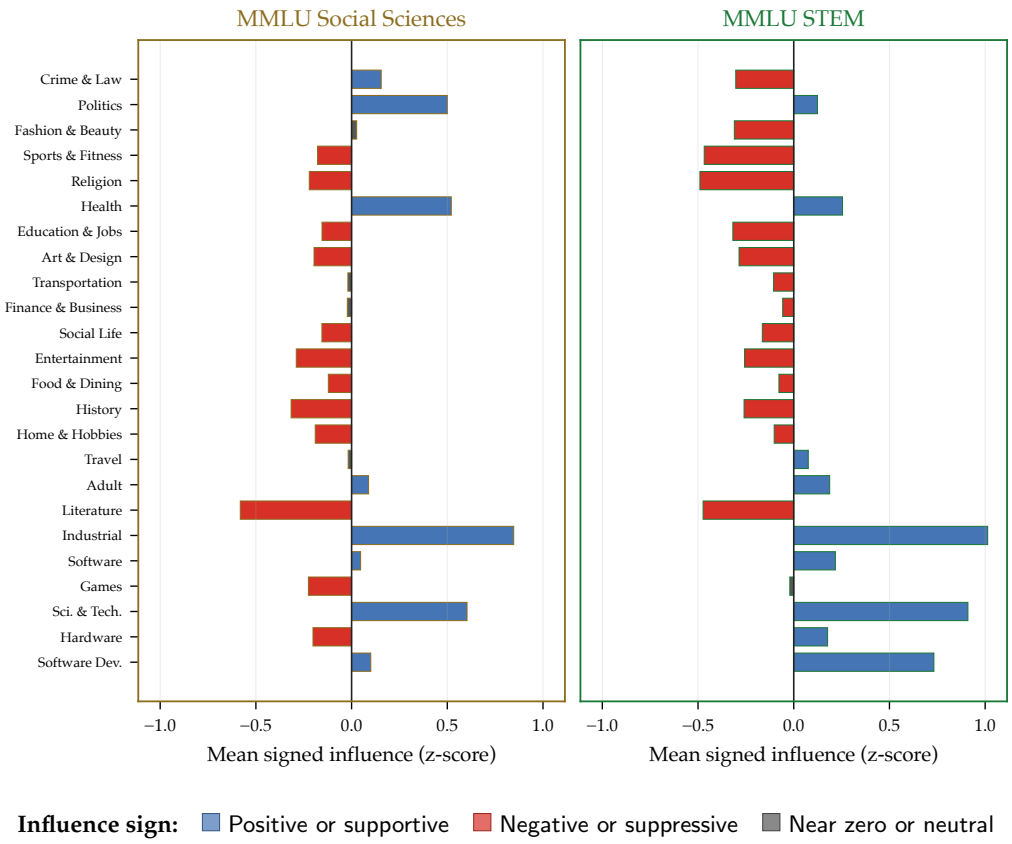


Figure 18: Paired topic-level influence comparison: MMLU Social Sciences vs. MMLU STEM. The knowledge-domain contrast is directionally consistent with the reasoning contrast (Figure 17) but weaker in magnitude, consistent with the main text findings.

(+0.46) and *Politics* (+0.38) lead but at roughly half that size. The maximum gap of 0.91 z-units in the reasoning pair versus 0.63 in the knowledge pair quantifies the attenuation reported in the main text: reasoning contrasts are sharper than knowledge contrasts.

F Correctness Differential

For each benchmark we separately compute influence over the correctly answered and incorrectly answered query subsets and report the difference $\Delta z = z_{\text{correct}} - z_{\text{incorrect}}$ per WebOrganizer (Wettig et al., 2025) bin. SocialIQA shows the strongest correctness differential, with maximum $|\Delta z|$=13.1, about 2.2× the next benchmark, ARC-Challenge (+6.0); the two MMLU benchmarks stay weak ($|\Delta z| \leq 2.2$) (Table 12). The composite heatmap (Figure 19) shows high-magnitude positive and negative cells distributed across many topic–format bins of SocialIQA's panel and consistent coloring within topic rows. ARC-Challenge's panel is dominated by a strongly positive *Literature* × *Customer Support* cell, with a few further cells (for example History & Geography × Customer Support and Industrial × Documentation) also reaching the edge of the shared color scale; the two MMLU panels remain uniformly muted. A provenance caveat applies to the ARC splits specifically: their per-query correct/incorrect alignment is reconstructed positionally from recovered evaluation artifacts rather than re-verified against original per-query identifiers, so the ARC correctness partitions carry weaker provenance than the other benchmarks' splits.

Reasoning-vs-knowledge sign flip. *Literature* × *Customer Support* is the dominant outlier and the top-positive correctness bin for *both* reasoning benchmarks: it is the strongest

Table 12: Per-benchmark correctness-differential extremes: bin with the strongest positive Δz (correct minus incorrect; appears more on correctly answered queries) and the strongest negative Δz (appears more in errors). *Literature* $\times$ *Customer Support* is the top-positive bin for both reasoning benchmarks, SocialIQA (+13.1) and ARC-Challenge (+6.0), and the strongest *negative* bin for both knowledge benchmarks, MMLU Social Sciences and MMLU STEM: a reasoning-vs-knowledge sign flip on a single bin. Brackets give bootstrap 95% CIs (2,000 resamples over queries); the four top-positive extremes exclude zero, while all four top-negative extremes include zero. SocialIQA's differential is the largest, about 2.2$\times$ ARC-Challenge's; the two MMLU benchmarks stay weak ($|\Delta z| \leq 2.2$). MMLU STEM's top-positive bin lands in a small, underfilled Documentation cell (Fashion & Beauty $\times$ Documentation; Table 6), while MMLU Social Sciences' top-positive Health $\times$ Documentation bin is fully populated.

Benchmark	**Top positive bin**		**Top negative bin**	
	Δz	Bin (Topic $\times$ Format)	Δz	Bin (Topic $\times$ Format)
SocialIQA	+13.11 [2.8, 24.3]	Literature $\times$ Customer Support	−4.60 [−7.7, 0.8]	Health $\times$ Documentation
MMLU Social Sciences	+0.86 [0.1, 1.8]	Health $\times$ Documentation	−2.19 [−5.4, 1.1]	Literature $\times$ Customer Support
ARC-Challenge	+6.00 [0.2, 10.8]	Literature $\times$ Customer Support	−1.31 [−2.6, 0.3]	Politics $\times$ Legal Notices
MMLU STEM	+1.83 [0.5, 3.2]	Fashion & Beauty $\times$ Documentation	−2.17 [−6.1, 1.6]	Literature $\times$ Customer Support

correctness-differential bin for SocialIQA (Δz=+13.1) and for ARC-Challenge (Δz=+6.0), while being the strongest bin associated with *incorrect* answers on the two knowledge benchmarks, MMLU Social Sciences (−2.2) and MMLU STEM (−2.2). Bootstrap 95% CIs over queries (2,000 resamples) qualify the flip: the positive arm is significant on both reasoning benchmarks (SocialIQA [2.8, 24.3], ARC-Challenge [0.2, 10.8], both excluding zero), whereas the negative arm on the knowledge benchmarks is not (MMLU Social Sciences [−5.4, 1.1], MMLU STEM [−6.1, 1.6], both spanning zero), so the suppressive side of the flip is directional rather than statistically resolved. The underlying mechanism differs by benchmark: for SocialIQA both cohorts are positive (z_{correct}=+17.6, $z_{\text{incorrect}}$=+4.5), so this content actively supports correct social reasoning; for ARC-Challenge both are negative (z_{correct}=−0.8, $z_{\text{incorrect}}$=−6.8), so it is net-harmful but far less so on correctly answered queries. Either way, the same corpus content—short, dialogue-rich book-club Q&A text whose lexical profile is detailed in Appendix H—separates correct reasoning from correct factual recall in this diagnostic. SocialIQA's strongest *negative* bin is the mirror image: Health $\times$ Documentation (Δz=−4.6), a documentation-format bin that is positive for the knowledge benchmarks. The two MMLU benchmarks have weak correctness differentials overall ($|\Delta z| \leq 2.2$). MMLU STEM's top-positive bin additionally lands in a small, underfilled Documentation cell (Fashion & Beauty $\times$ Documentation, one of the corpus's empty-tail bins; Table 6), so its signal is support-limited on top of being weak; MMLU Social Sciences' top-positive Health $\times$ Documentation bin is fully populated, so its weakness is not a support artifact.

The asymmetry is itself informative. The strong correctness differentials of the two reasoning benchmarks (SocialIQA and ARC-Challenge) suggest that specific corpus regions contribute differently to correct vs. incorrect reasoning, while the weaker signal for the two knowledge benchmarks (both MMLUs) is consistent with a more diffuse relationship between corpus content and item-level accuracy.

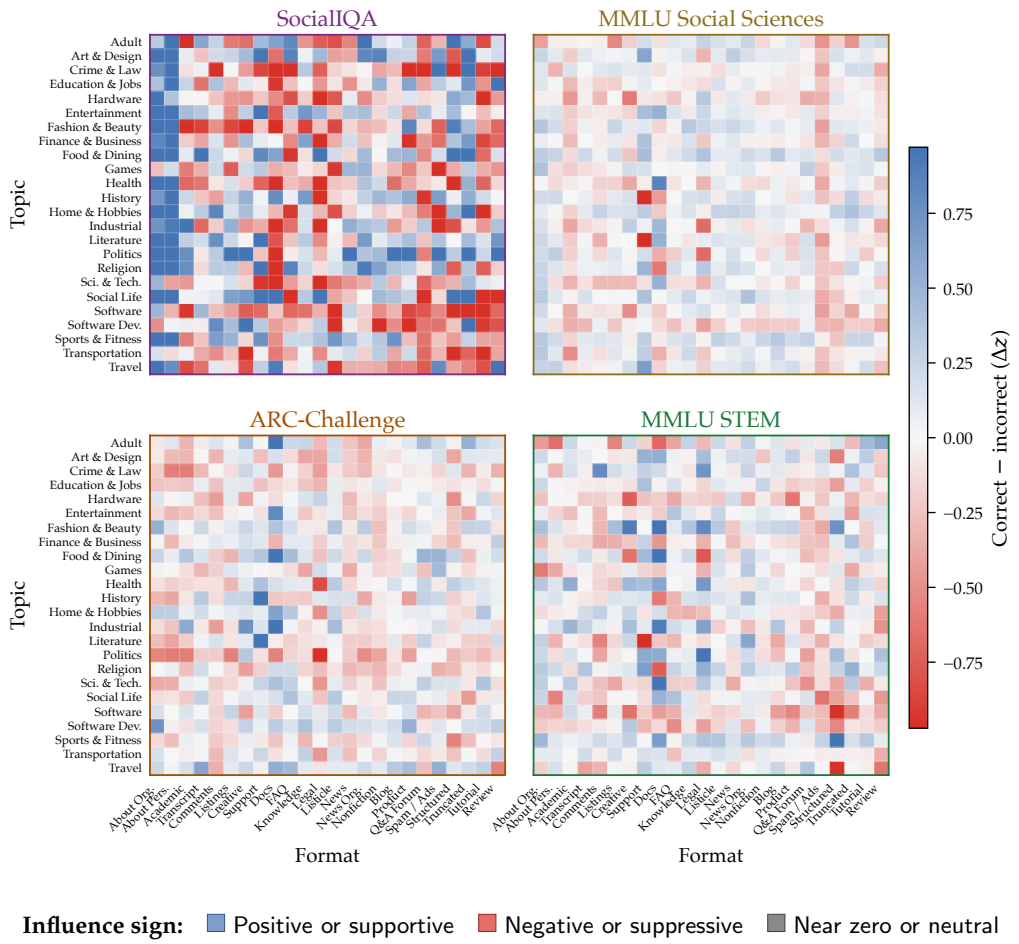


Figure 19: Correctness differential: mean influence on correctly answered queries minus mean influence on incorrectly answered queries, per WebOrganizer bin. **Positive** values support correct responses more than errors; **negative** values indicate the reverse; more extreme cells saturate on the shared 95th-percentile color scale.

G Held-out Probe Analysis for Social Reasoning

This appendix collects the paper's held-out probes: benchmarks that were not part of the four core attribution-query design in Table 8. The goal is breadth, not a separate unlearning claim. OLMES emits per-query predictions, correctness labels, subset metadata, chance baselines, and Wilson intervals. TrackStar then aggregates signed influence over the 24 topic by 24 format taxonomy. We first check whether the model solves each probe above chance, because attribution is only interpretable as capability evidence when the evaluated mask has usable prediction signal.

G.1 Setup and Query Construction

Table 13 records the held-out query surface. The table is separate from Table 8 because the latter defines the four core in-paper attribution queries. Here, the query set is an audit of additional social, pragmatic, moral, bias, and Theory-of-Mind probes. StereoSet (Nadeem et al., 2021) is included only to document the boundary: its official evaluation is native likelihood, so it is not used in the direct-answer TrackStar attribution headline.

Probe	Queries	Scoring format	Role	Caveat
BBQ	58,492	Direct MC plus bias	Main	Bias benchmark; interpret with category-level bias diagnostics.
BBH Disambiguation QA	250	Forced choice	Main	Small-N BBH row, but clear margin above chance.
BBH Causal Judgment	187	Forced choice	Caution	Wilson interval reaches chance.
BBH Snarks	178	3-shot CoT MC	Supporting	Pragmatic BBH check; attribution is noisier because the query set is small.
ToMBench	2,860	Direct MC	Main	Public merged source caveat retained from the OLMES card.
SimpleToM	3,441	Direct suite	Supporting/control	Mental-state QA is positive; behavior and judgment QA are controls.
NegotiationToM	28,560	Direct MC, desire/belief	Main	Intention and all-round splits are not in the current scored suite.
PUB	26,742	Direct MC	Main with retained mask	pub_2/pub_3 are excluded from the main PUB mask.
MMLU moral/humanities	1,652	OLMES MC subjects	Main	Moral and humanities subject grouping.
ETHICS	19,967	Direct MC	Secondary/control	Non-justice retained as secondary; justice is a failure control.
MORABLES	709	Direct MC	Secondary	Low absolute accuracy and CC-BY-NC-4.0 caveat.
MoralExceptQA/RBQA	148	Forced choice	Secondary	Small-N row with CC-BY-NC-SA-4.0 caveat.
StereoSet	4,229	Native likelihood	Excluded from attribution headline	Reported only as native-likelihood bias evidence, not direct-answer TrackStar attribution.

Table 13: Held-out query construction and inclusion policy. Query counts are the exported OLMES or diagnostic request rows used by the held-out artifact bundle.

G.2 Direct-Answer Gate and Tiering

Table 14 defines the tiering rule before attribution figures are interpreted. Main held-out probes must be clearly above chance, relevant to social, pragmatic, moral, negotiation, or bias attribution, and supported by the updated OLMES and TrackStar artifacts. Secondary rows remain part of the audit trail but carry less weight. Failure controls and excluded rows define the boundary of the direct-answer claims.

Tier	Benchmarks or masks	Role in the paper
Main	BBQ; BBH Disambiguation QA; ToMBench; MMLU moral/humanities; NegotiationToM desire/belief; PUB excluding pub_2/pub_3	Primary held-out attribution surfaces: social, pragmatic, moral, negotiation, and bias benchmarks that clear chance by a substantial margin.
Supporting	SimpleToM mental-state QA; BBH Snarks	Included because they probe relevant pragmatic or mental-state behavior, but not used as the strongest held-out evidence.
Secondary	ETHICS non-justice; MoralExceptQA/RBQA; MORABLES	Above chance but lower-weight because of smaller margins, small N, low absolute accuracy, or license caveats.
Caution, control, excluded	BBH Causal Judgment; SimpleToM behavior QA; SimpleToM judgment QA; ETHICS justice; pub_2/pub_3; StereoSet	Reported to define the boundary of the claims. StereoSet is excluded from direct-answer attribution claims because its official metric is native likelihood.

Table 14: Held-out benchmark tiers used in the paper. Tiers are assigned before interpreting TrackStar attribution so the attribution claims follow the scored capability evidence.

Table 15 reports the scored direct-answer results. The strongest rows are BBQ, BBH Disambiguation QA, ToMBench, MMLU moral/humanities, NegotiationToM, and PUB after removing failed PUB subsets. SimpleToM is subset-dependent: mental-state QA clears chance, while behavior QA and judgment QA are controls rather than positive SimpleToM-wide evidence.

G.3 Split-Sensitive Masks

Table 16 records the masks that require subset handling. PUB uses the retained aggregate after removing pub_2 and pub_3, both of which fall below their chance baselines. ETHICS is retained as a secondary non-justice aggregate because the justice subset fails as a capability

Benchmark or mask	N	Acc.	Chance	Δ	95% CI	Tier
BBQ	58,492	0.719	0.333	+0.385	[0.715, 0.722]	Main
BBH Disambiguation QA	250	0.636	0.332	+0.304	[0.575, 0.693]	Main
ToMBench	2,860	0.641	0.333	+0.308	[0.623, 0.658]	Main
MMLU moral/humanities	1,652	0.533	0.250	+0.283	[0.509, 0.557]	Main
NegotiationToM	28,560	0.430	0.200	+0.230	[0.425, 0.436]	Main
PUB excluding pub_2/pub_3	21,730	0.671	0.442	+0.229	[0.665, 0.677]	Main
SimpleToM mental-state QA	1,147	0.685	0.500	+0.185	[0.658, 0.711]	Supporting subset
MoralExceptQA/RBQA	148	0.635	0.500	+0.135	[0.555, 0.708]	Secondary
ETHICS non-justice	17,263	0.566	0.500	+0.066	[0.559, 0.574]	Secondary
MORABLES	709	0.260	0.200	+0.060	[0.229, 0.293]	Secondary
BBH Causal Judgment	187	0.567	0.500	+0.067	[0.495, 0.636]	Caution
SimpleToM behavior QA	1,147	0.493	0.500	-0.007	[0.464, 0.521]	Failure control
SimpleToM judgment QA	1,147	0.456	0.500	-0.044	[0.427, 0.485]	Failure control

Table 15: Direct-answer held-out OLMES results. Δ is accuracy minus the benchmark-specific chance baseline; Wilson intervals are computed on accuracy.

check. SimpleToM is reported by subset because the overall average mixes one above-chance mental-state row with behavior and judgment rows that do not support a positive SimpleToM-wide claim.

Subset or mask	N	Acc.	Chance	Δ	95% CI	Handling
SimpleToM mental-state QA	1,147	0.685	0.500	+0.185	[0.658, 0.711]	Supporting subset; the only SimpleToM subset clearly above chance.
SimpleToM behavior QA	1,147	0.493	0.500	-0.007	[0.464, 0.521]	Failure control; not used for the SimpleToM attribution claim.
SimpleToM judgment QA	1,147	0.456	0.500	-0.044	[0.427, 0.485]	Failure control; below chance.
PUB excluding pub_2/pub_3	21,730	0.671	0.442	+0.229	[0.665, 0.677]	Main PUB mask.
pub_2	2,506	0.152	0.200	-0.048	[0.139, 0.167]	Failure control; excluded from the PUB main mask.
pub_3	2,506	0.131	0.200	-0.069	[0.118, 0.145]	Failure control; excluded from the PUB main mask.
ETHICS non-justice	17,263	0.566	0.500	+0.066	[0.559, 0.574]	Secondary ETHICS mask.
ETHICS justice	2,704	0.426	0.500	-0.074	[0.407, 0.444]	Failure control; excluded from the ETHICS non-justice mask.

Table 16: Split-sensitive held-out results. PUB and ETHICS use retained-mask aggregates because specific subsets fail as capability checks; SimpleToM is reported by subset because its overall average hides opposite behavior across subsets.

Figure 20 shows the retained PUB (Sravanthi et al., 2024) mask rather than the failed PUB subsets.

G.4 Pragmatics and BBH Family

BBH Snarks remains the established pragmatic held-out check already used in the manuscript. Snarks presents pairs of statements and asks the model to identify which is sarcastic. We treat it as a pragmatic social-reasoning probe because sarcasm interpretation depends on speaker intent, social norms, and the gap between literal meaning and communicative purpose; prior social-reasoning stress tests show that such cues can be brittle for LLMs (Sap et al., 2022; Shapira et al., 2024).

We score 178 Snarks queries against the same 5.68M-document working set. Figure 21 shows the signed influence heatmap on the canonical axis ordering. Q&A Forum and Creative Writing show consistently positive influence in both Snarks and SocialIQA, and Social Life topic content is positive for both. With only 178 Snarks queries distributed across 576 bins, the per-bin signal is noisier than for the primary benchmarks. We therefore treat Snarks as supporting pragmatic evidence rather than one of the strongest held-out rows.

Table 17 keeps Snarks with the BBH-family probes (Suzgun et al., 2023). BBH Disambiguation QA is the stronger BBH-family direct-answer result in the current scored bundle, while BBH Causal Judgment is marginal because its confidence interval reaches chance.

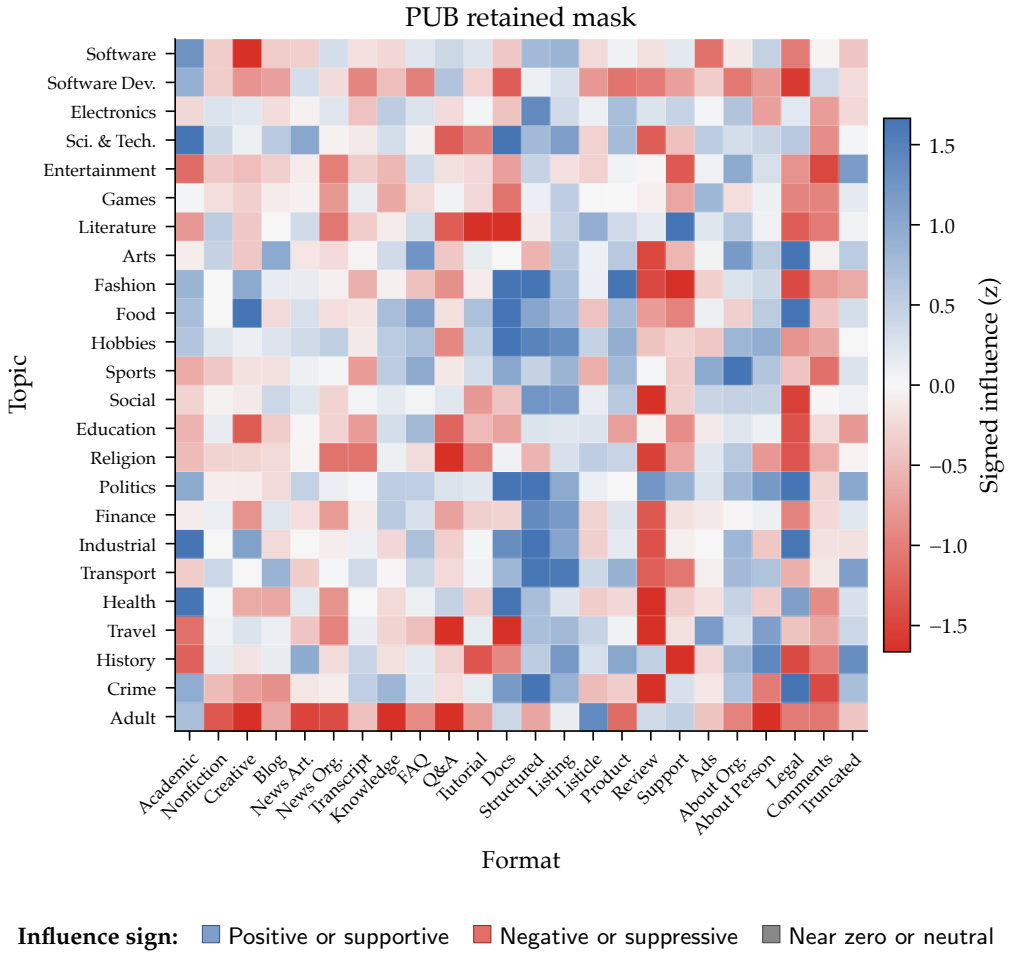
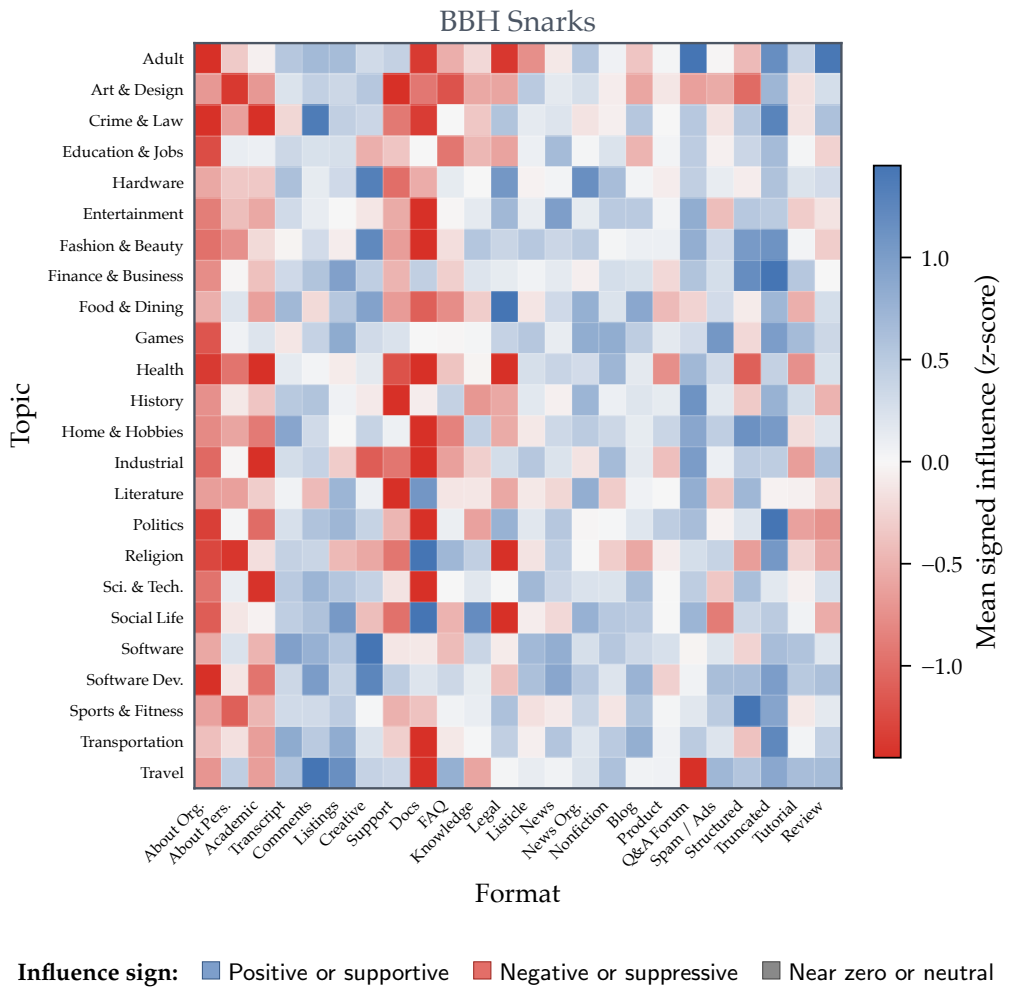


Figure 20: Signed TrackStar influence for PUB after excluding pub_2 and pub_3. The figure follows the split-sensitive mask used in Table 16.

Benchmark	N	Acc.	Chance	Δ	95% CI	Interpretation
BBH Snarks	178	0.579	0.500	+0.079	n/a	Existing pragmatic BBH held-out check; useful as supporting evidence.
BBH Disambiguation QA	250	0.636	0.332	+0.304	[0.575, 0.693]	Stronger current BBH-family held-out probe; used as main evidence.
BBH Causal Judgment	187	0.567	0.500	+0.067	[0.495, 0.636]	Caution row; point estimate is above chance but the interval reaches chance.

Table 17: BBH-family held-outs. BBH Snarks uses the existing manuscript evaluation, while BBH Disambiguation QA and BBH Causal Judgment come from the current scored OLMES bundle.



Influence sign: Positive or supportive Negative or suppressive Near zero or neutral

Figure 21: Signed mean influence for BBH Snarks (held-out benchmark) on the canonical topic–format axis ordering. The shared sign convention uses **blue for positive influence** (supporting sarcasm detection), white near zero, and **red for negative influence**; more extreme cells saturate. Dialogue-rich formats such as Q&A Forum and Creative Writing carry positive mass, consistent with the SocialIQA pattern from the primary analysis; topic-level marginals are noisier given the smaller query set (n=178).

Figure 22 shows the stronger BBH-family held-out heatmap with the full topic and format axes visible.

G.5 Theory of Mind and Negotiation

The ToM-facing held-out probes separate stable direct-answer evidence from subset-limited evidence. ToMBench (Chen et al., 2024) clears its chance baseline by a large margin and is treated as main held-out evidence. NegotiationToM (Chan et al., 2024) also clears chance for the scored desire and belief splits, but intention and all-round variants are not in the current suite. SimpleToM (Gu et al., 2024) is not treated as a single positive row: the mental-state subset is supporting evidence, while behavior QA and judgment QA are controls.

Figures 23–25 show the corresponding full held-out attribution panels. These panels use the same sign convention as the primary benchmark heatmaps and keep the split policy visible in the figure captions.

G.6 Moral, Ethics, and Bias Probes

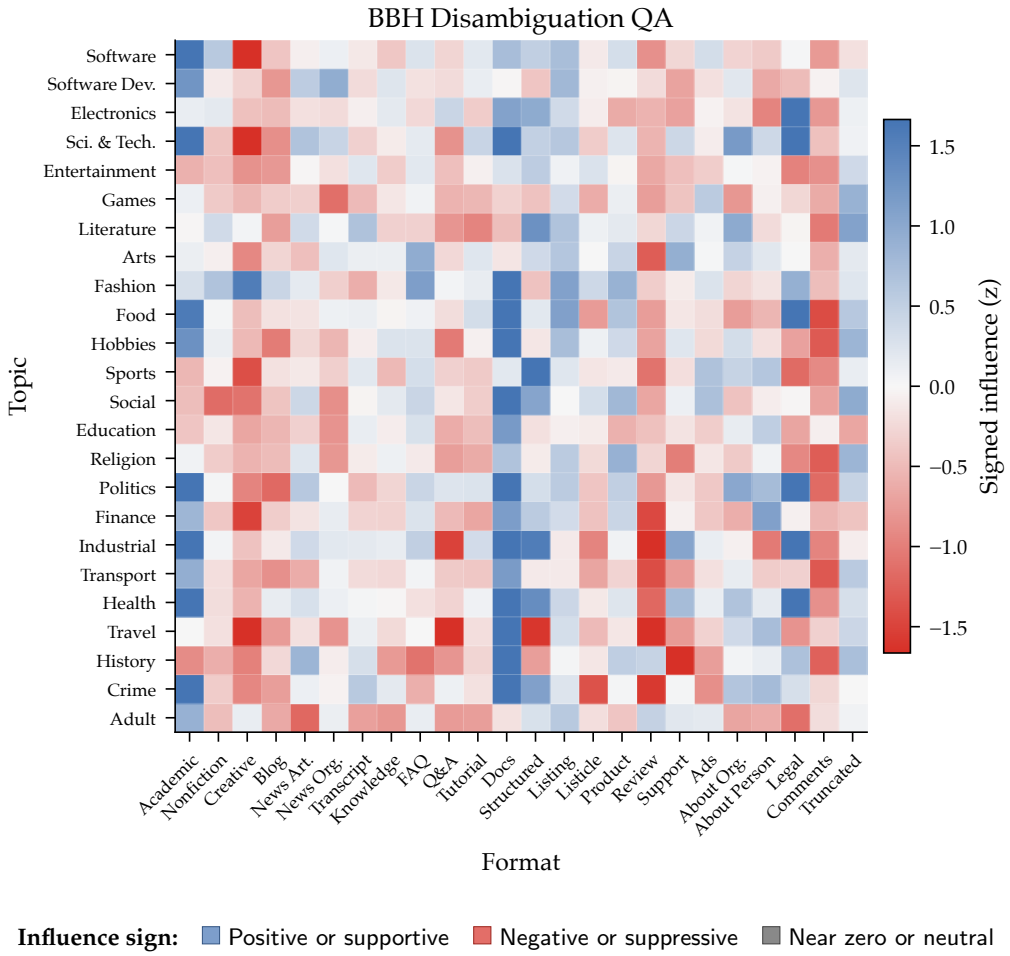
The moral and ethics probes give breadth without carrying equal evidential weight. MMLU moral/humanities is a main held-out probe because its four-choice accuracy is well above chance. ETHICS (Hendrycks et al., 2021a) non-justice is retained as secondary evidence because the justice subset fails. MORABLES (Marcuzzo et al., 2025) and MoralExceptQA/RBQA (Jin et al., 2022) are secondary because of lower absolute accuracy, small N, and license caveats.

BBQ (Parrish et al., 2022) needs benchmark-specific treatment in addition to aggregate accuracy. Its direct-answer score makes it a strong capability check, but BBQ is also a bias benchmark. Table 18 reports ambiguous and disambiguated BBQ bias scores by category so the paper does not collapse bias-oriented behavior into ordinary multiple-choice accuracy alone.

Category	Context	N	Acc.	Non-unknown N	Bias
ALL	ambig	29,246	0.5408	13,431	-0.0301
ALL	disambig	29,246	0.8964	28,129	-0.0189
Age	ambig	1,840	0.3620	1,174	-0.0804
Age	disambig	1,840	0.8141	1,792	-0.0446
Disability_status	ambig	778	0.4023	465	-0.1093
Disability_status	disambig	778	0.8355	738	-0.0515
Gender_identity	ambig	2,836	0.5049	1,404	0.0254
Gender_identity	disambig	2,836	0.8942	2,635	0.0269
Nationality	ambig	1,540	0.5305	723	-0.0734
Nationality	disambig	1,540	0.9071	1,494	-0.0094
Physical_appearance	ambig	788	0.4251	453	-0.0647
Physical_appearance	disambig	788	0.7843	743	-0.0417
Race_ethnicity	ambig	3,440	0.5302	1,616	0.0640
Race_ethnicity	disambig	3,440	0.9125	3,342	0.0078
Race_x_SES	ambig	5,580	0.5842	2,320	-0.0362
Race_x_SES	disambig	5,580	0.9285	5,421	-0.0279
Race_x_gender	ambig	7,980	0.6128	3,090	-0.0559
Race_x_gender	disambig	7,980	0.8936	7,749	-0.0216
Religion	ambig	600	0.5183	289	-0.0617
Religion	disambig	600	0.8150	542	-0.0074
SES	ambig	3,432	0.5111	1,678	-0.0216
SES	disambig	3,432	0.9327	3,271	-0.0419
Sexual_orientation	ambig	432	0.4931	219	-0.0394
Sexual_orientation	disambig	432	0.8727	402	-0.0199

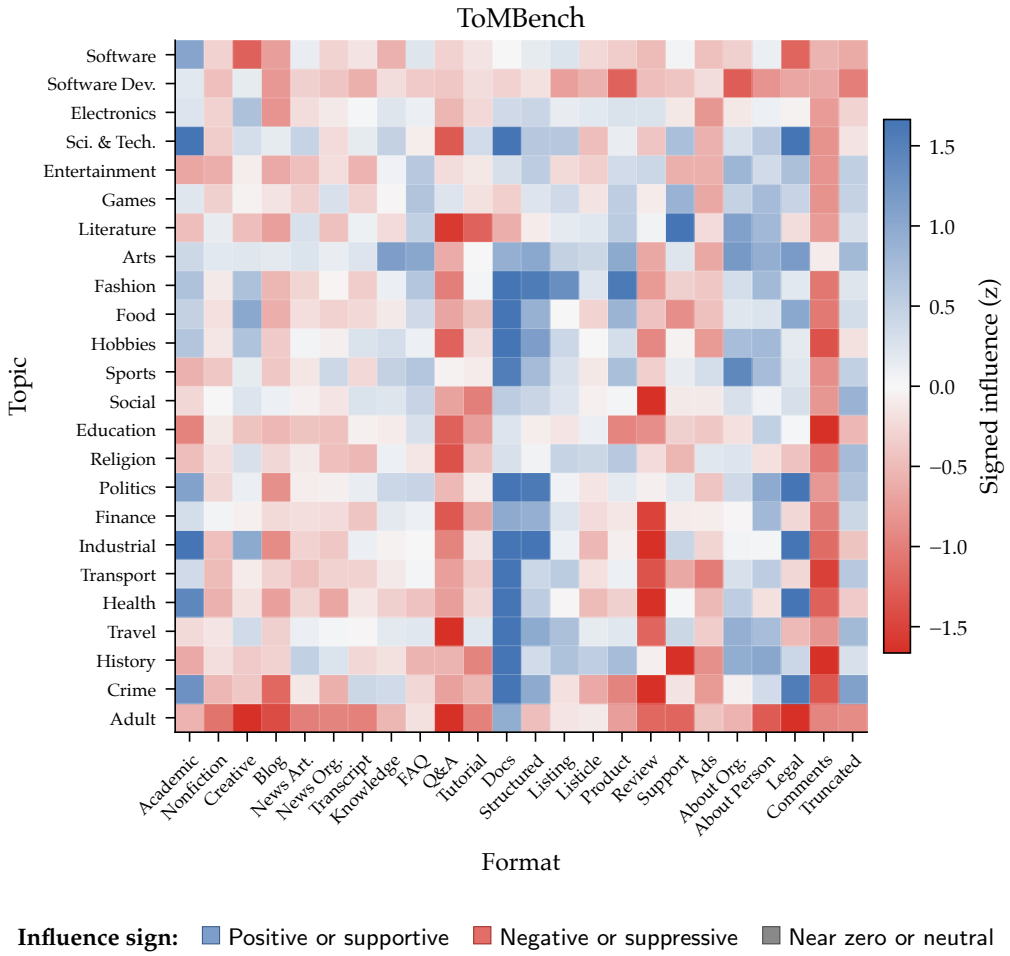
Table 18: BBQ bias diagnostics by category.

Figures 26–29 show the moral and bias held-out panels that correspond to this tiered interpretation. The secondary panels are included for auditability, not to give them the same evidential weight as the main held-out probes.



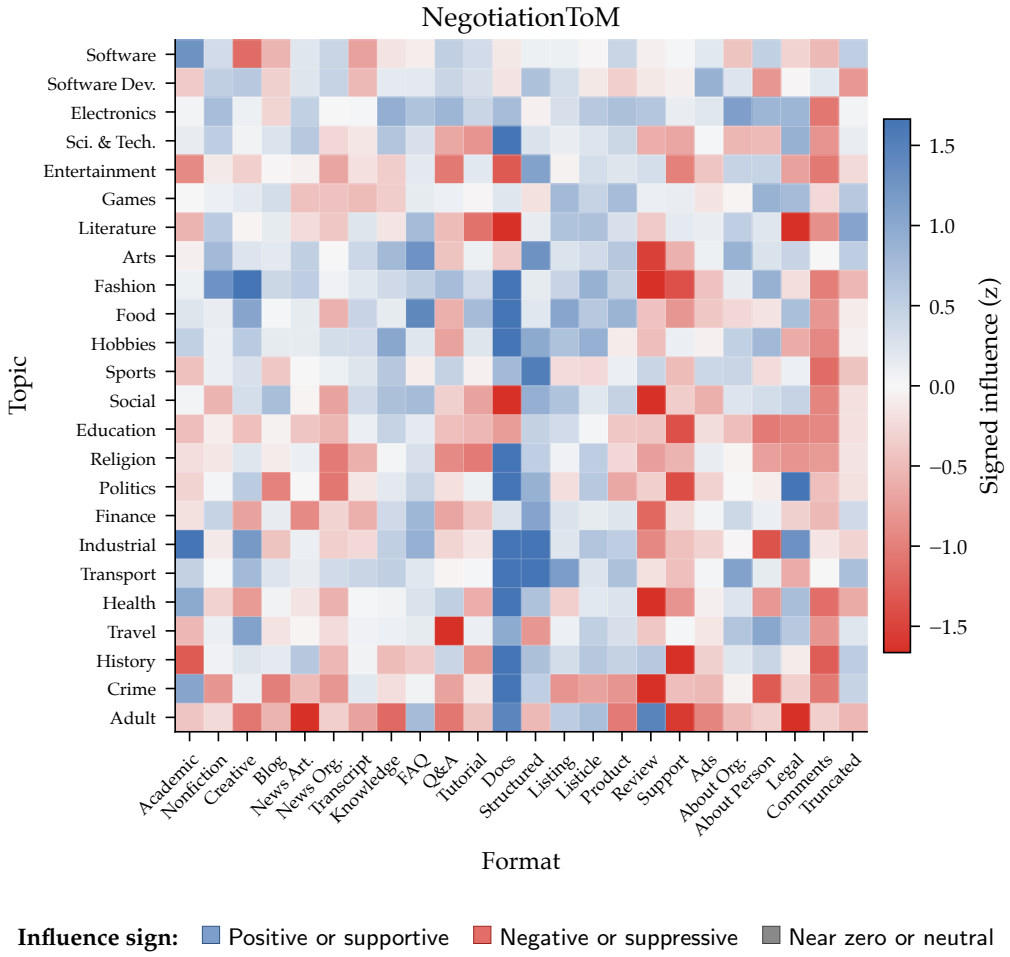
Influence sign: Positive or supportive Negative or suppressive Near zero or neutral

Figure 22: Signed TrackStar influence for BBH Disambiguation QA on the full topic–format grid. This is the strongest BBH-family direct-answer held-out probe in the current bundle and is shown as a full axis-labeled panel rather than a report crop.



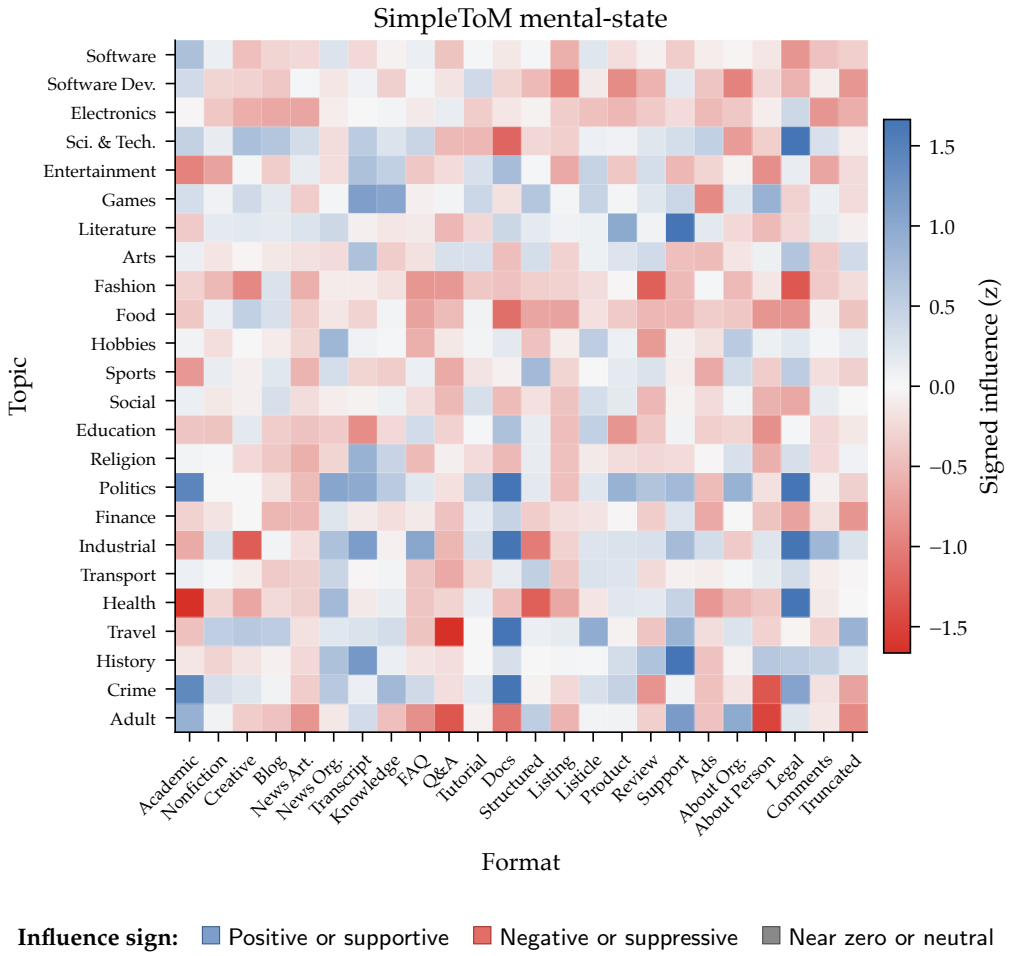
Influence sign: Positive or supportive Negative or suppressive Near zero or neutral

Figure 23: Signed TrackStar influence for ToMBench on the full topic–format grid. ToMBench is a strong direct-answer held-out row, but it remains a held-out attribution result rather than unlearning validation.



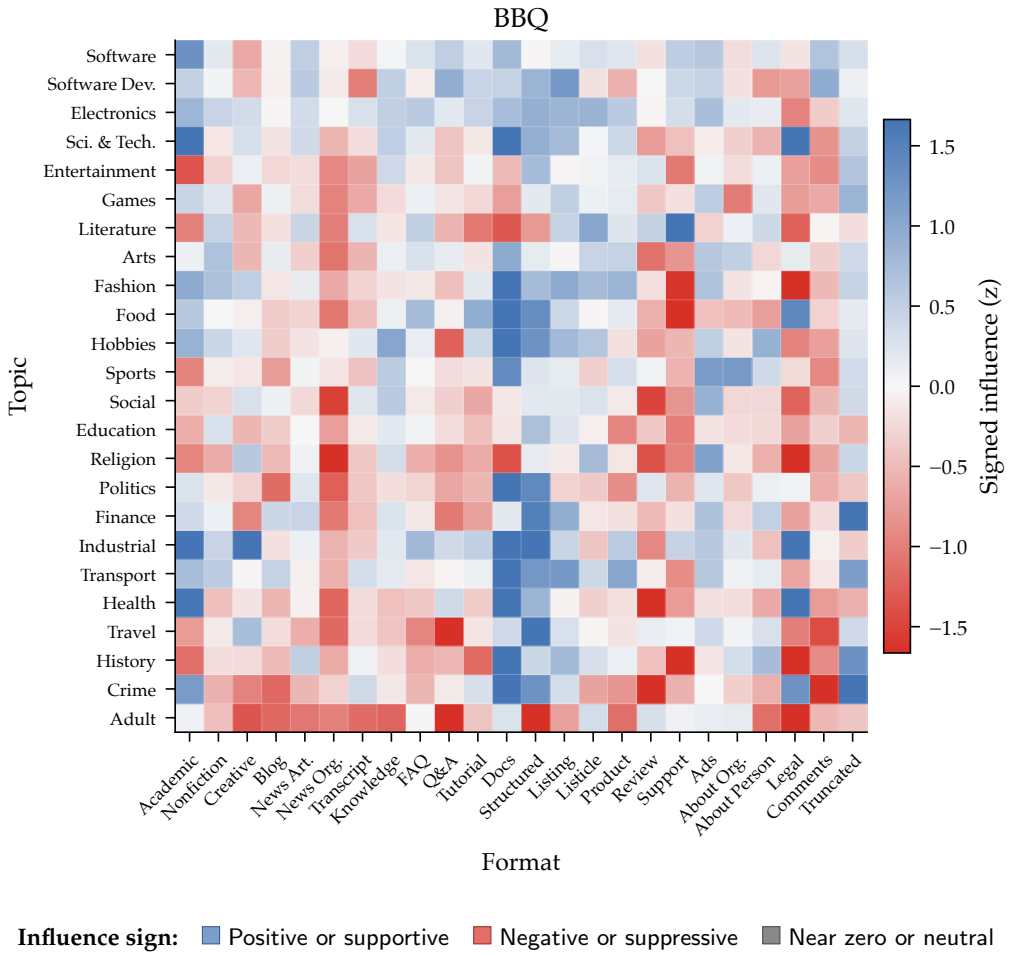
Influence sign: Positive or supportive Negative or suppressive Near zero or neutral

Figure 24: Signed TrackStar influence for the NegotiationToM belief/desire mask. The panel uses the same held-out TrackStar sign convention as the main attribution heatmaps.



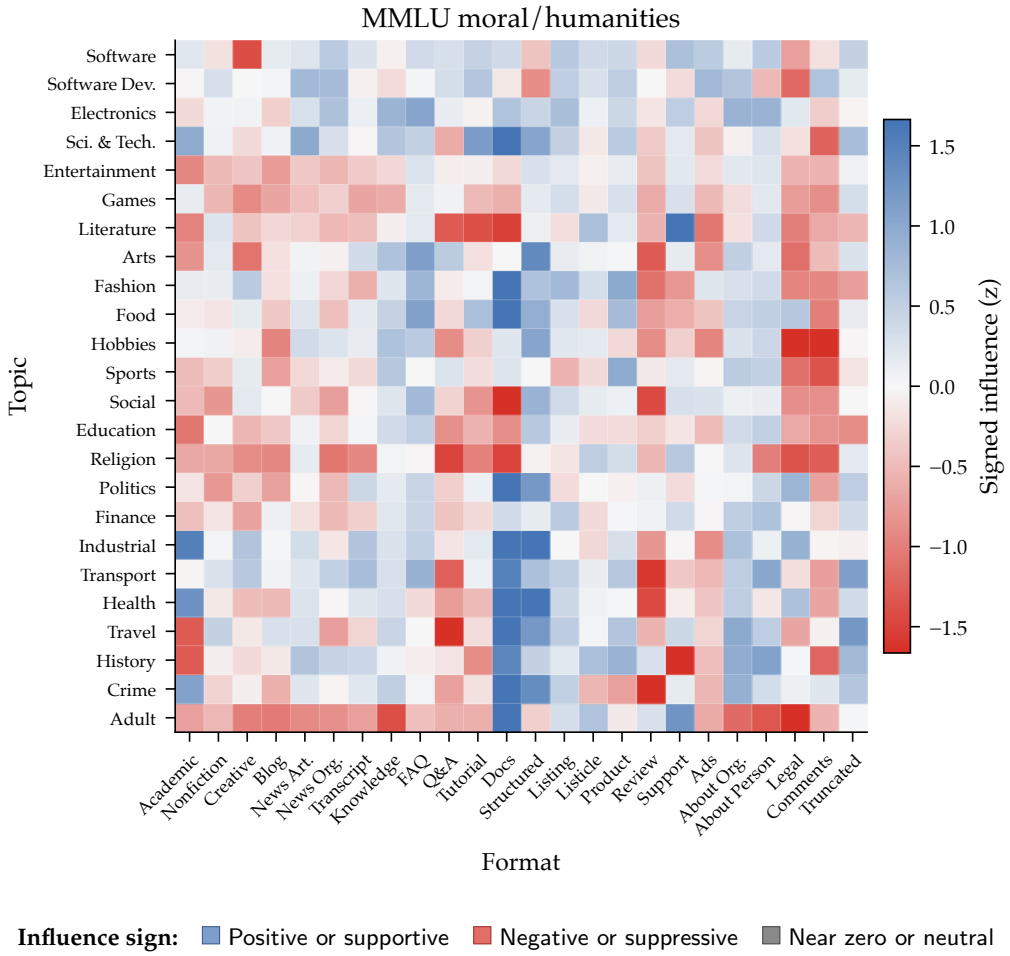
Influence sign: Positive or supportive Negative or suppressive Near zero or neutral

Figure 25: Signed TrackStar influence for the SimpleToM mental-state subset. We show the subset that clears the direct-answer gate, rather than using the behavior and judgment rows as positive SimpleToM-wide evidence.



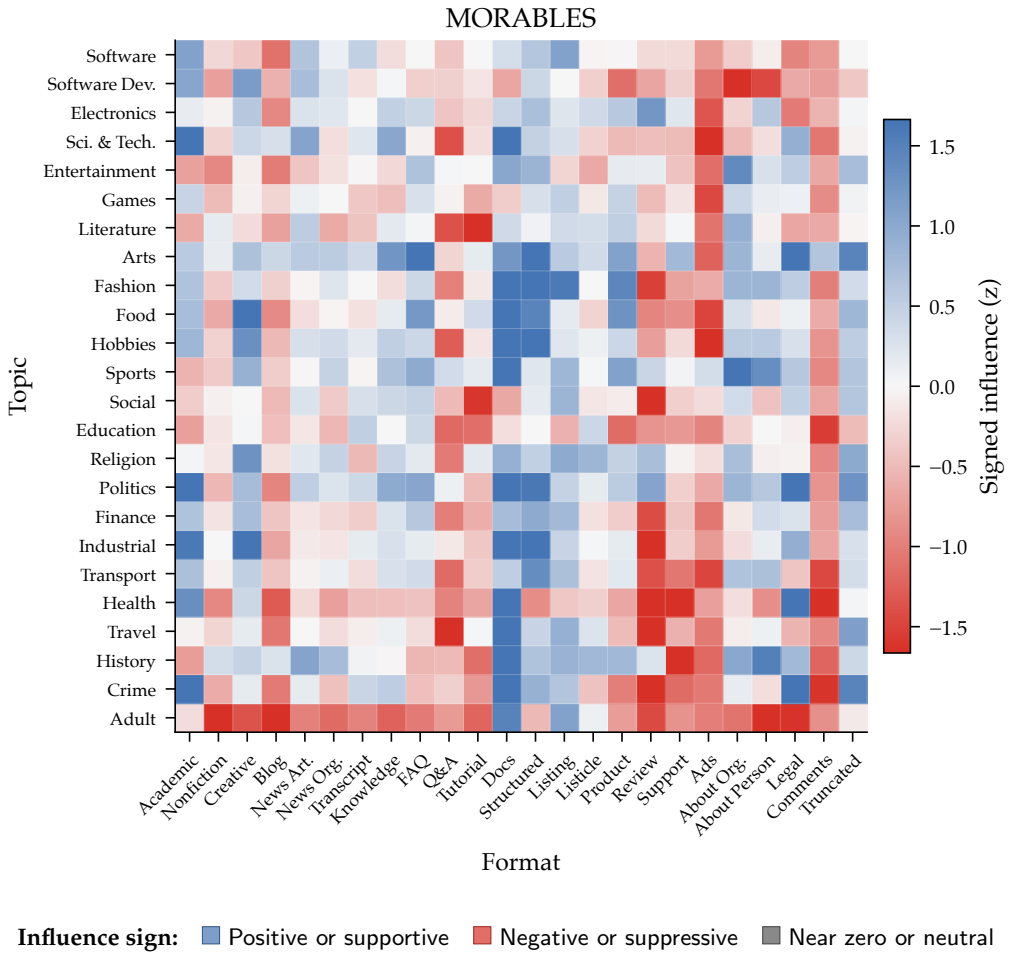
Influence sign: Positive or supportive Negative or suppressive Near zero or neutral

Figure 26: Signed TrackStar influence for BBQ. BBQ is included both as a strong direct-answer held-out probe and as a bias-oriented benchmark with separate category-level diagnostics in Table 18.



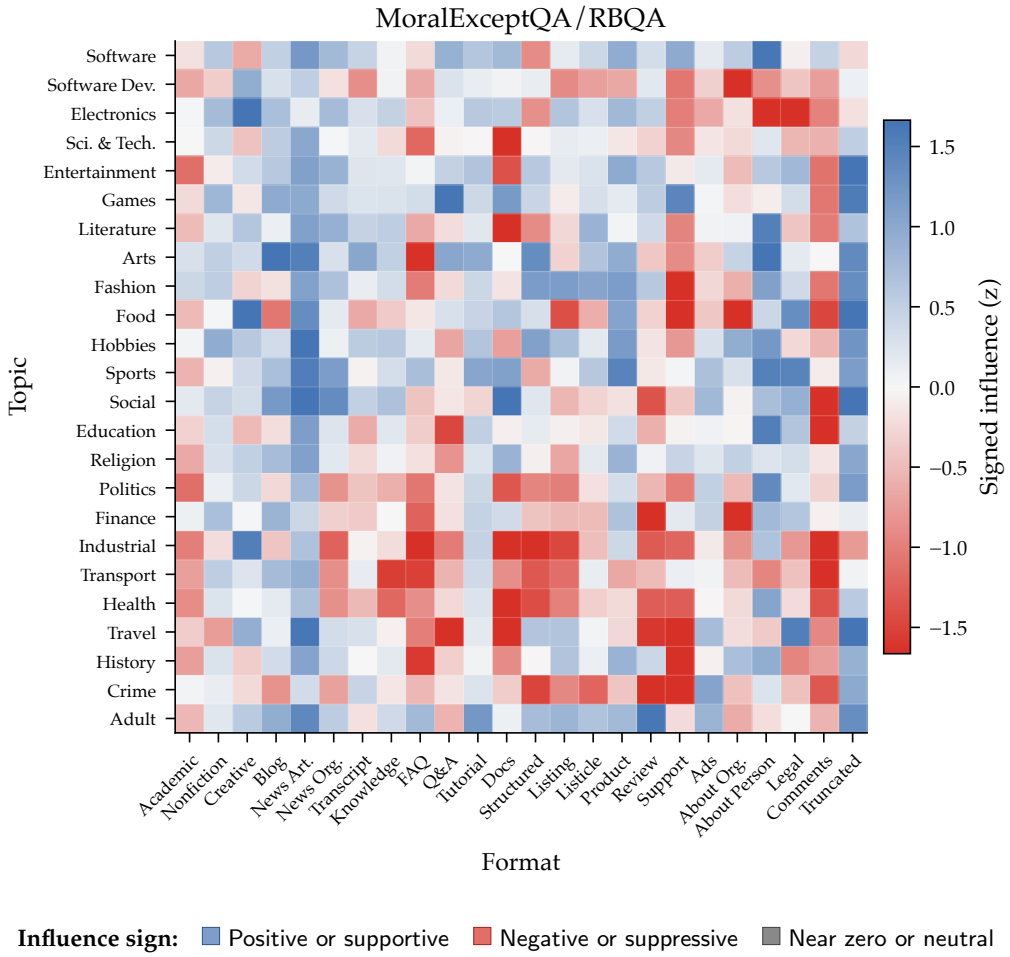
Influence sign: Positive or supportive Negative or suppressive Near zero or neutral

Figure 27: Signed TrackStar influence for MMLU moral/humanities. This panel gives the moral-domain main held-out mask the same topic–format inspection surface as the core attribution benchmarks.



Influence sign: Positive or supportive Negative or suppressive Near zero or neutral

Figure 28: Signed TrackStar influence for MORABLES. The panel is retained as secondary moral-inference evidence because the direct-answer result is only modestly above chance and the benchmark carries license caveats.



Influence sign: Positive or supportive Negative or suppressive Near zero or neutral

Figure 29: Signed TrackStar influence for MoralExceptQA/RBQA. We report it as secondary evidence because the scored query count is small, even though the direct-answer score clears the chance baseline.

G.7 Attribution Diagnostics and Claim Boundary

Table 19 reports a compact top-k bin diagnostic for the TrackStar-verified held-out OLMES probes. Table 20 then summarizes consensus and variance over the fixed six-mask set of main-tier masks (Table 14) after within-mask standardization. These diagnostics are not used to upgrade a benchmark tier by themselves; they describe where signed attribution mass is concentrated after the direct-answer gate has been checked.

Probe	Highest mean top-k bin	Top-k	Mean score	Mean rank
BBH Causal Judgment	Transportation / Spam Ads	4	0.02196	25.75
BBH Disambiguation QA	Politics / News Org	117	0.02206	29.59
BBQ	Software / Truncated	41,790	0.02108	40.87
ETHICS	Adult / Customer Support	4,647	0.02258	28.52
MMLU moral/humanities	Travel / News Org	608	0.02303	24.12
MORABLES	Software Development / Spam Ads	42	0.02078	41.17
MoralExceptQA/RBQA	Software Development / Tutorial	42	0.02314	25.76
NegotiationToM	History / Content Listing	22,542	0.02312	20.15
PUB	Literature / News Org	4,725	0.02260	37.15
SimpleToM	Entertainment / Structured Data	1,000	0.02323	28.88
ToMBench	Health / FAQ	557	0.02059	33.23

Table 19: Highest mean top-k bin per TrackStar-verified held-out OLMES probe. The table is a compact bin-level diagnostic, not a document-level ranking.

Diagnostic	Topic	Format	Pos.	Neg.	Mean z	Std. z
Positive consensus	Literature	Customer Support	6	0	+4.694	4.222
Positive consensus	Politics	Documentation	6	0	+4.550	1.311
Positive consensus	Health	Documentation	6	0	+4.514	1.715
Negative consensus	History	Customer Support	0	6	-9.259	2.594
Negative consensus	Travel	Q&A Forum	0	6	-4.956	2.995
Negative consensus	Crime & Law	User Review	0	6	-2.052	0.334
High variance	Literature	Customer Support	6	0	+4.694	4.222
High variance	Travel	Q&A Forum	0	6	-4.956	2.995
High variance	History	Customer Support	0	6	-9.259	2.594
High variance	Industrial	Documentation	6	0	+3.974	1.918

Table 20: Cross-mask diagnostics over the six main-tier held-out masks (BBQ, BBH Disambiguation QA, ToMBench, MMLU moral/humanities, NegotiationToM, and retained PUB). Values are within-mask z-scores of median per-query influence over topic–format bins.

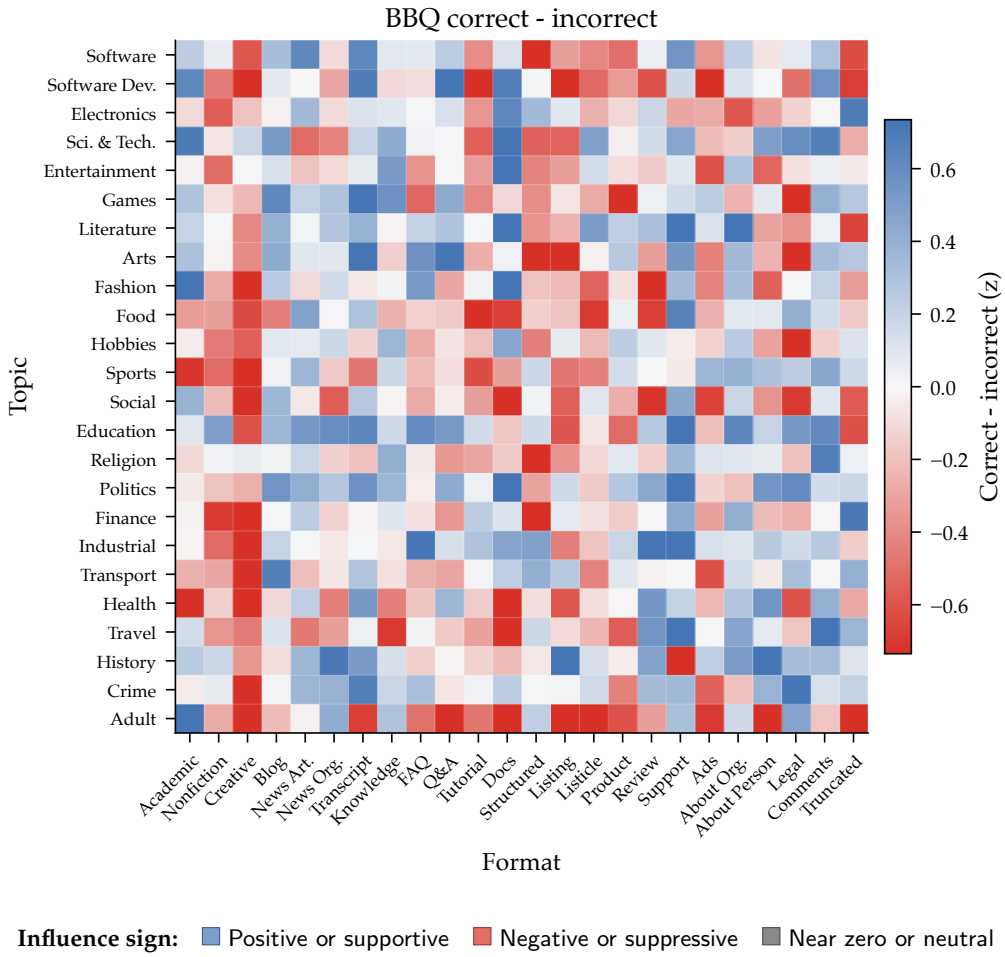
Figures 30 and 31 add two correctness-differential panels for high-signal held-out probes. They expose whether correct and incorrect queries emphasize different topic–format bins, but they remain diagnostic rather than tier-setting evidence.

Figure 32 summarizes signed TrackStar influence for the main held-out masks plus SimpleToM mental-state QA. The figure averages the z-scored topic–format grid across formats to make the held-out profiles readable while preserving the paper's topic order and positive/negative influence convention.

Table 21 asks whether the same topic–format bins are repeatedly supportive across the headline held-out probes and the two core social-domain probes. We z-score signed influence within each probe over the 576 bins, then rank bins by the number of probes with positive support, the number with $z \geq 1$, and the mean headline-probe z. The strongest recurring bins are documentation, legal-notice, structured-data, and academic-writing formats, especially in politics, crime-and-law, industrial, home-and-hobbies, and science-and-technology topics.

Figure 33 shows the benchmark-by-bin matrix behind that table. The pattern should be read as corpus-level support, not as proof that a single document causes performance across all probes. The control columns are diagnostic: several high-consensus bins are also positive for ARC-Challenge and MMLU STEM, so the appendix treats these bins as broadly supportive structured-documentation regions rather than as uniquely social training data.

Table 22 runs a narrower top-k document pass over the available held-out top-k files. Each query contributes rank-weighted support, the score is normalized by query count before

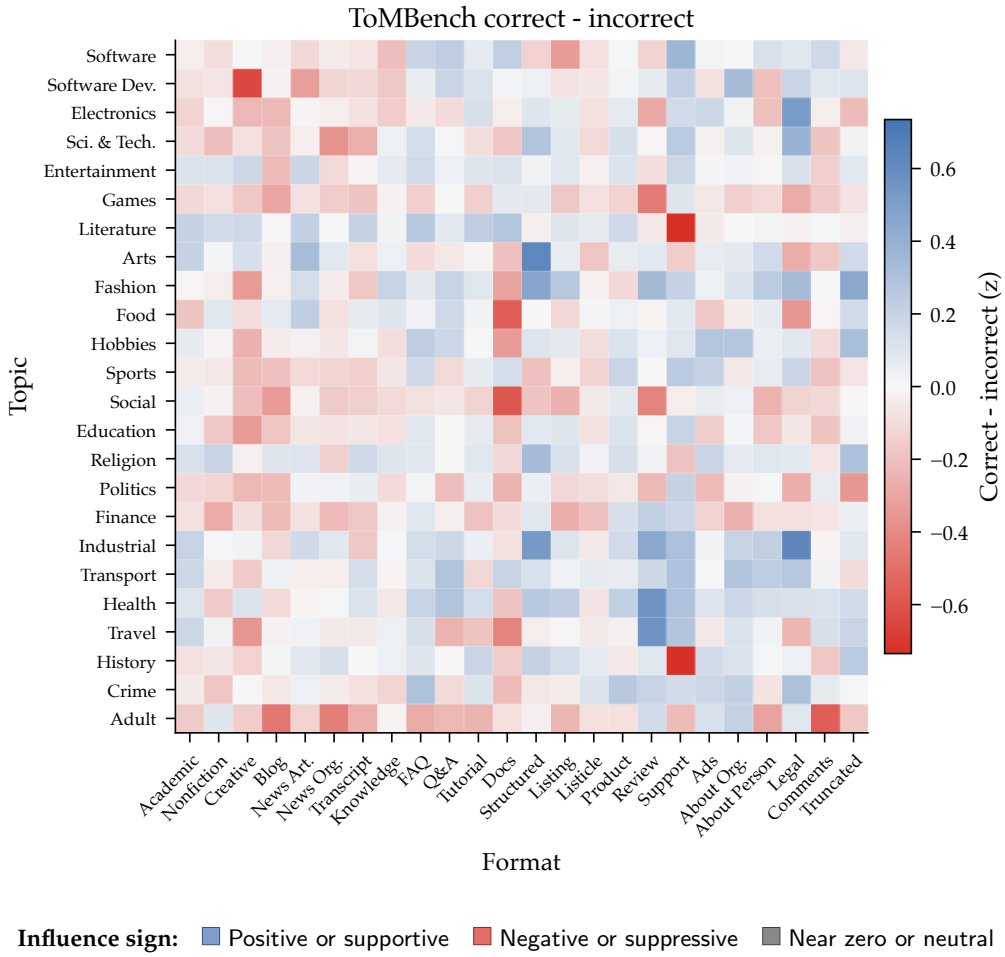


Influence sign: Positive or supportive Negative or suppressive Near zero or neutral

Figure 30: Correctness differential for BBQ: signed influence on correctly answered queries minus signed influence on incorrectly answered queries. The panel is diagnostic; it does not change the benchmark tier.

Topic-format bin	Pos.	Strong	Mean z	Ctrl. z	Spec.
Politics / Documentation	12	11	+3.47	+5.58	-2.10
Crime & Law / Documentation	12	11	+2.41	+2.85	-0.44
Industrial / Legal Notices	12	10	+2.49	+6.68	-4.19
Home & Hobbies / Documentation	12	10	+2.05	+3.12	-1.07
Science & Technology / Academic Writing	12	9	+1.99	+3.02	-1.03
Transportation / Documentation	12	7	+0.98	+1.21	-0.23
Politics / Structured Data	12	6	+0.83	+0.68	+0.15
Science & Technology / Content Listing	12	2	+0.50	+0.95	-0.45

Table 21: Topic-format bins with the strongest supportive consensus across the held-out and core social-domain probes. Scores are z-scored within each probe over the 576 topic-format bins. *Pos.* counts headline probes with positive signed influence, *Strong* counts headline probes with $z \geq 1$, and *Spec.* subtracts the mean control-probe z from the mean headline-probe z.



Influence sign: Positive or supportive Negative or suppressive Near zero or neutral

Figure 31: Correctness differential for ToMBench: signed influence on correctly answered queries minus signed influence on incorrectly answered queries. This provides a second high-signal held-out diagnostic without importing every report delta panel into the manuscript.

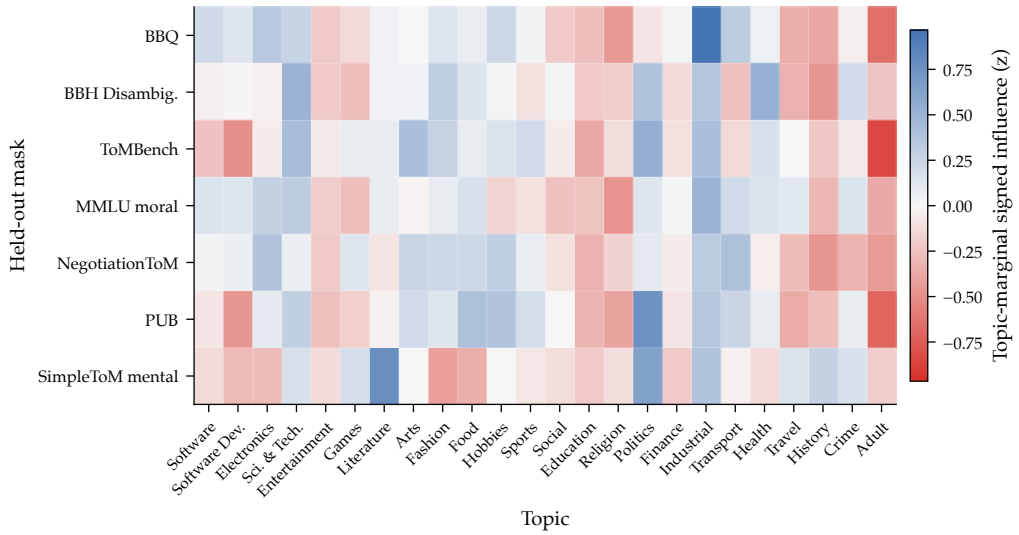


Figure 32: Topic-marginal signed TrackStar influence for the main held-out masks plus the SimpleToM mental-state subset. Values are z-scored within each mask over all topic–format bins before averaging across formats. This figure is descriptive attribution evidence, not unlearning validation.

Figure 33: Cross-probe supportive consensus for the top ten supportive topic-format bins, the first eight of which appear in Table 21. Each column is standardized within a probe, so the heatmap shows whether the same corpus bins are repeatedly supportive rather than comparing raw influence magnitudes across probes.

within-probe standardization, and a document is recurrent only if it is strongly supported in at least two probes. We report document identifiers, joined topic–format bins, and recurrence scores without text snippets. This table is useful as an audit and triage signal, but it does not replace the bin-level consensus claim.

Document ID	Bin	Strong	Pos.	Mean z	Strong probes
shard_0011:17169	Adult / Tutorial	6	8	+35.46	BBH Disambig. QA, NegotiationToM, PUB retained, MMLU moral/hum., ETHICS non-justice, MORABLES
shard_0032:6913	Crime & Law / Nonfiction Writing	5	6	+30.56	BBQ, PUB retained, MMLU moral/hum., ETHICS non-justice, MoralExceptQA/RBQA
shard_0012:8615	Adult / User Review	5	6	+17.11	BBQ, ToMBench, NegotiationToM, PUB retained, MMLU moral/hum.
shard_0259:1792	Software / Product Page	5	5	+14.00	BBQ, NegotiationToM, PUB retained, MMLU moral/hum., ETHICS non-justice
shard_0074:3718	Entertainment / About Person	5	5	+11.06	BBQ, ToMBench, NegotiationToM, PUB retained, MMLU moral/hum.
shard_0127:16370	Games / Q&A Forum	5	5	+6.78	BBQ, ToMBench, NegotiationToM, PUB retained, MORABLES
shard_0096:8008	Finance & Business / Legal Notices	4	5	+42.43	BBQ, PUB retained, SimpleToM mental, ETHICS non-justice
shard_0270:16671	Software Development / About Person	4	4	+32.35	BBQ, NegotiationToM, PUB retained, MMLU moral/hum.

Table 22: Top-k document recurrence diagnostic across available held-out top-k files. A document is recurrent only if its query-normalized rank-weighted support has $z \geq 2$ in at least two probes. Document IDs are reported without text snippets; bin labels appear only when a doc-to-bin manifest is supplied.

G.8 Register Robustness of the Block Signature

The consensus substrate above is a register claim, so we test it directly against its main confound. We partition the 24 WebOrganizer formats into a prescriptive-register group (documentation, knowledge articles, academic writing, legal notices, organizational news, structured data) and a descriptive, interpersonal group (personal blogs and pages, comment sections, Q&A forums, customer support, user reviews, creative writing, audio transcripts). We score each probe's register loading as the difference in mean influence z-score between the two groups; a positive loading means the probe's influence concentrates on formal, prescriptive-register text. The test is necessary because the reported profiles are scored with instruct-model query gradients, and swapping only the query model can move a probe's register loading by as much as 0.74 with everything else held fixed. The block's formal-register signature could therefore in principle have been a query-model artifact rather than a corpus fact.

It is not. We rebuilt base-model query gradients for all seven block probes against the same document index and re-scored them, with the decision rule fixed before scoring. Under base queries the block's mean register loading is $+0.63$ with six of seven probes positive (the pre-registered rule required at least $+0.3$ and five of seven), comparable to the base-query knowledge anchors ($+0.77$). The lone exception, MMLU-Moral, is the only five-shot probe in the set, so a prompt-format rather than corpus explanation is plausible for it. The block's documentary, formal-register signature therefore survives the query-model control.

With the single exception of the held-out Theory-of-Mind unlearning test in subsection G.10 below, the held-out probes in this appendix are evaluation and TrackStar attribution evidence rather than unlearning validation. StereoSet is excluded from the direct-answer attribution headline because its official metric is native likelihood rather than direct-answer accuracy. MoralExceptQA/RBQA and MORABLES are retained as secondary rows with their license caveats.

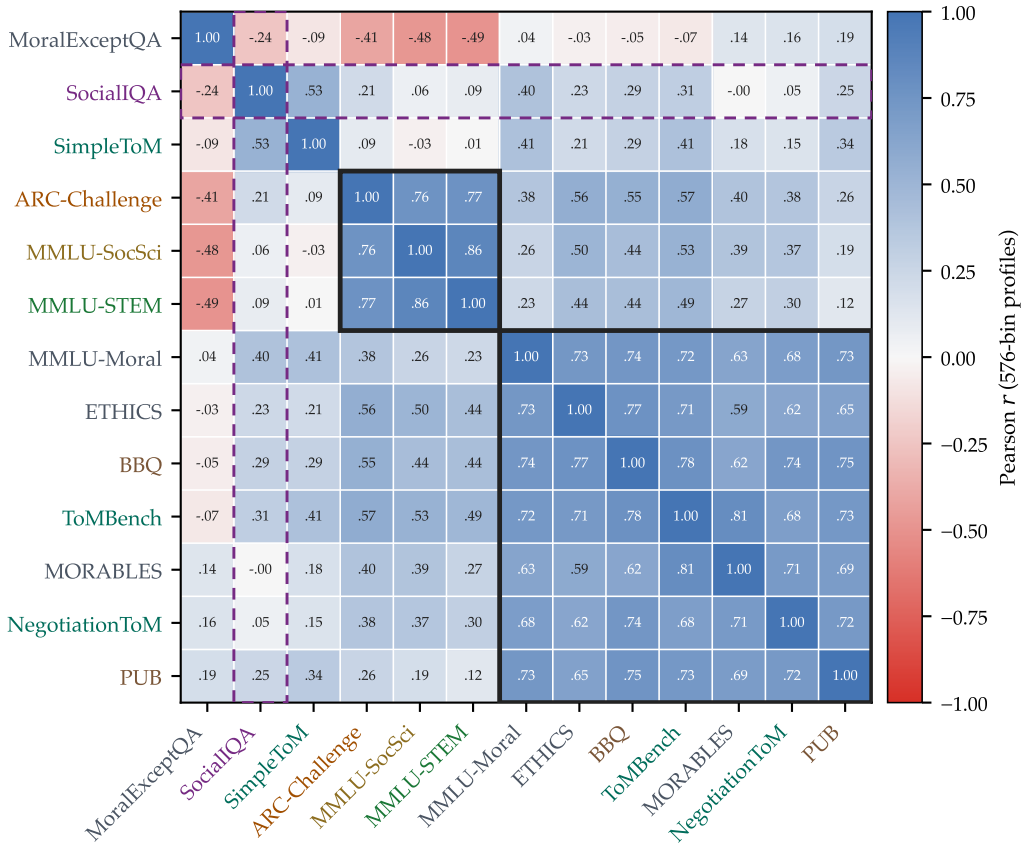


Figure 34: Annotated probe-profile correlation matrix: Pearson r between 576-bin signed-influence profiles, dendrogram-ordered (average linkage on $1 - r$). Solid outlines mark the held-out and knowledge blocks; the dashed row and column mark SocialIQA.

G.9 Probe Geometry

To quantify how the held-out attribution profiles relate to one another and to the core benchmarks, we correlate the 576-bin signed-influence profiles across probes and cluster the resulting correlation matrix. The dominant feature is a coherent held-out block: seven probes spanning moral judgment, bias, and theory of mind correlate at a mean within-block $r = 0.70$, with the three comparison benchmarks forming the second block. SocialIQA correlates weakly with the held-out suite: its mean correlation to the held-out block is 0.22, and its mean pairwise correlation over the nine direct-answer-eligible probes is $r = 0.20$ (Pearson) and near zero under Spearman. ToMBench is the strongest single proxy for the suite (mean $r = 0.56$ Pearson, 0.47 Spearman). Hierarchical clustering assigns seven of the nine held-out probes to the knowledge-benchmark cluster rather than to SocialIQA; strict pairwise counting of the same comparison gives five of nine under Pearson and seven of nine under Spearman, so we report the clustering and pairwise views together. One design caveat bounds the analysis: query gradients were computed on the Instruct checkpoint for some probes and on Base for others, and cross-variant correlations on matched BBH tasks ($r = 0.09$ to 0.29) bound the size of this confound; it does not reverse any ordering reported here.

G.10 Causal Validation: Held-out ToM Unlearning

Beyond attribution, we run a single held-out *unlearning* test that mirrors the per-topic single-bin procedure used for the primary benchmarks (NGDiff with rank-8 LoRA on OLMo3-

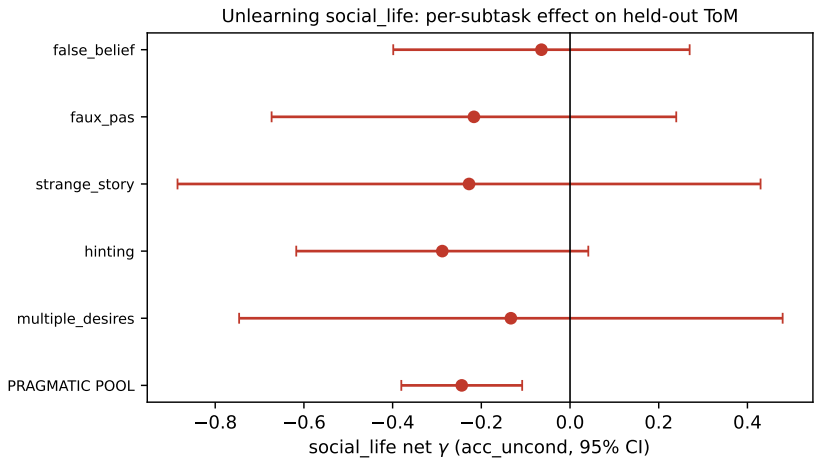


Figure 35: Held-out Theory-of-Mind unlearning. Per-subtask net influence ($\gamma_{\text{net}} = \gamma_{\text{topic}} - \gamma_{\text{null}}$, three seeds) from unlearning the `social_life` topic, with 95% confidence intervals, for the ToMBench social-pragmatic subtasks alongside the first-order false-belief and multiple-desires control subtasks. The pooled pragmatic effect excludes zero (-0.244 unconditional accuracy, $p = 0.039$); the plotted control subtasks are null, as is the separate `science_math_and_technology` topic control (not shown). The result rests on the causal unlearning test: attribution and unlearning do not co-rank for these later-added held-out probes.

7B Base, one unlearning run per WebOrganizer topic), now applied to the ToMBench social-pragmatic subtasks and corrected against a null-bin control ($\gamma_{\text{net}} = \gamma_{\text{topic}} - \gamma_{\text{null}}$) over three training seeds. Pooling the three social-pragmatic subtasks (hinting, faux-pas recognition, and strange-story; $n = 9$ subtask–seed replicates), unlearning the `social_life` topic degrades social-pragmatic ToM by a net -0.244 in unconditional accuracy (95% CI $[-0.380, -0.108]$; sign test $8/9$, $p = 0.039$); the effect holds on the primary metric (-0.123, CI $[-0.208, -0.038]$) and survives a leave-one-subtask-out jackknife. `social_life` is a $z = -4.4$ outlier among the 24 topics on this construct.

Two controls isolate the effect. Unlearning `science_math_and_technology` against the same social-pragmatic pool gives -0.008 (confidence interval includes zero), and first-order false-belief alone is null, so the degradation is specific to social-pragmatic ToM rather than a generic capability loss or first-order belief tracking. Figure 35 reports the per-subtask net effects with confidence intervals.

We state two limits. Several ToMBench subtasks saturate on the Base model under `acc_per_char`, so this held-out signal is weaker than the non-saturated primary suite. And for these later-added probes, gradient attribution and unlearning damage do not co-rank (topic-level Spearman ≈ 0): `social_life` is the top causal topic but only middling in attribution support. The held-out ToM claim therefore rests on the causal unlearning test, and we do not assert attribution–unlearning convergence for it.

G.11 Data-Side Levers for Safety-Relevant Capabilities

Several held-out probes measure capabilities that alignment and safety evaluation target directly: mental-state inference (ToMBench, SimpleToM), normative judgment (ETHICS, MMLU-Moral), and social bias (BBQ). The provenance maps make these capabilities addressable at curation time by identifying which corpus regions to audit, enrich, or exclude when a training run should strengthen or limit a specific social capability, and the `social_life` unlearning test shows that at least one such lever is functionally necessary for social-pragmatic theory of mind (Appendix G.10). The evidence boundary matters. For the remaining held-out probes the maps are attributional; attribution rank and causal damage did not

co-rank on the later-added probes; and locating a capability's corpus support is not the same as showing that removing it would be safe or effective. We read this as the data-side complement to internal-state interpretability: where probing and circuit analysis localize a capability inside the network, corpus-scale attribution localizes it in the training distribution, and the measurement is possible only where weights, training checkpoints, and the document-level corpus are released together, as in the ecosystem studied here.

H Bin Characterization

H.1 Lexical Profiling Protocol

Lexical profiling is a descriptive layer applied *after* attribution. The attribution pipeline ranks Dolma3 WebOrganizer bins by signed influence for each benchmark or probe; lexical profiling then asks what kind of language lives in those high-influence bins. These features never select bins, compute influence scores, or tune the attribution pipeline. Their job is interpretive: they turn "this bin mattered" into a readable description of the training text that attribution surfaced.

Because Appendix G first establishes the direct-answer gate and tier labels for the social-reasoning probe masks, this appendix can use those tiers as reader-facing structure. It profiles the four primary benchmarks, the instruct BBH diagnostics, and the same social-reasoning probe masks without treating secondary diagnostics, caution diagnostics, or failure controls as main evidence.

Reader guide: two grouping axes

Benchmark target groups are used for pooling, figure labels, and diagnostic contrasts. They come from the paper roster in `configs/rq4_paper_benchmarks.yaml`: **social reasoning** (SocialIQA), **commonsense reasoning** (the instruct BBH probes), and **expository targets** (MMLU Social Sciences, ARC-Challenge, and MMLU STEM). The expanded roster also keeps social-reasoning probe masks in separate main, supporting, secondary, caution, and failure-control tiers. Figure 38 asks whether the original target groups differ lexically; the OpenLIWC-style heatmaps then show the social-reasoning probe tiers without merging controls into the main evidence.
Format clusters are a separate, benchmark-agnostic taxonomy over WebOrganizer formats. Figure 37 uses local colors for dialogic, personal, expository, structured, news, and boilerplate. These colors describe text form, not target-group membership.
LIWC-family heatmaps use the paper's sign convention: **blue means above** the Dolma working-set baseline and **red means below** it. Printed values give the exact full-grid z-score, so the figures remain readable without color.

Scope and OpenLIWC-style implementation. The LIWC-adjacent run covers the paper's current 21-entry lexical-profile roster: the four primary benchmarks, the instruct-scored BBH probes (Snarks, Causal Judgment, and Sports Understanding), and 14 social-reasoning probe masks. The social-reasoning probe masks are tiered as six main-evidence masks (BBQ, BBH Disambiguation QA, ToMBench, MMLU moral/humanities, NegotiationToM, and PUB excluding pub2/pub3), one supporting diagnostic (SimpleToM mental-state), three secondary diagnostics (ETHICS non-justice, MoralExceptQA/RBQA, and MORABLES), one caution diagnostic (BBH Causal Judgment), and three failure controls (ETHICS justice, PUB pub2/pub3, and SimpleToM behavior/judgment). StereoSet is excluded from correctness-based social-reasoning probe attribution and is not part of this lexical roster. The checked per-bin matrix has 576 WebOrganizer bins and 609 columns over 5,678,621 working-set documents; the benchmark-conditioned top-bin matrix has 420 rows (21 benchmarks $\times$ top-20 bins) and 617 columns.

We approximate LIWC-22-style constructs with an open implementation rather than with LIWC dictionaries, LIWC.app, NRC EmoLex, or any closed analyzer. Empath category unions cover broad semantic families such as affect, social, drives, perception, lifestyle,

Route	Open mechanism	What it uses	Coverage claim
A	Empath category unions	Open semantic categories are merged into LIWC-like affect, social, drives, perception, lifestyle, and physical families.	34 realized dimensions.
B	Open closed-class lexicons	In-repo lists cover pronouns, function words, cognition connectives, informal register, and perception words.	34 realized dimensions.
C	Open summary-variable proxies	Analytic-open, Tone-open, Clout-approx, and Authentic-approx follow published construct directions but are not LIWC.app scores.	4 realized dimensions.
D	Punctuation regex counts	Punctuation dimensions are transparent regex counts over the same tokenized text stream.	7 realized dimensions.
E	Text statistics	Word count, words per sentence, big-word share, and dictionary coverage are computed directly from the per-bin text.	4 realized dimensions.
F	Tense-focus heuristic	Past, present, and future focus use documented morphology and auxiliary-marker heuristics, not a shipped POS tagger.	3 realized dimensions.
G	Declared non-parity	Dimensions without a defensible open reproduction are listed explicitly rather than silently approximated.	Not realized: `verb`, `adj`, `moral`, `risk`, `curiosity`, `mental`.

Table 23: OpenLIWC-style route guide and coverage audit for the lexical matrix (576 WebOrganizer bins, 5,678,621 documents, 7,073,883,562 whitespace-tokenized words; the attribution pipeline separately counts ∼10.5B subword tokens). Routes A–F produce open, LIWC-adjacent approximation columns in the checked per-bin parquet artifact rather than LIWC.app outputs; route G records dimensions we refuse to approximate because doing so would mislabel the construct.

and physical content (Fast et al., 2016). In-repo lexicons cover closed-class function words, pronouns, cognition, informal-register, and remaining perception families. Punctuation, text statistics, and tense-focus heuristics are computed directly from the same tokenized text stream. The four LIWC summary variables are open, direction-preserving proxies: *Analytic-open*, *Tone-open*, *Clout-approx*, and *Authentic-approx*. They follow the construct directions in prior LIWC summary-variable work (Pennebaker et al., 2014; Kacewicz et al., 2013; Newman et al., 2003) and are mapped to $[0, 100]$ with 50 as corpus-neutral, but they are *not* numerically equivalent to LIWC.app outputs.

The implementation refuses silent parity claims. Dimensions with no defensible open reproduction are route-G non-parity entries: `verb/adj` (requires a POS tagger we do not ship), `moral`, `risk`, `curiosity`, and mental health. We validate the implementation statically with Empath super-set checks, pronoun partition checks, punctuation spot checks, hand-computed summary formulas, and a registry-coverage test. The matrix, baselines, crosswalk, tokenizer choice, and route registry are released as artifacts. Regenerating top-bin confidence intervals for the social-reasoning probe masks is blocked unless the per-document lexical feature parquets are restored; the checked artifact bundle contains the per-bin and benchmark-conditioned top-bin matrices, but not the per-document lexical features needed for that bootstrap.

Corpus baseline and benchmark-conditioned profiles. All lexical rates are computed over the working set (5,678,621 documents, ∼10.5B tokens) and aggregated to WebOrganizer topic–format bins. Density features are reported per 1,000 words; the token total is the attribution tokenizer's subword count, while the lexical rates use the whitespace-word count reported in Table 23. OpenLIWC-style family dimensions use the full 576-bin grid as the reference distribution, so a positive full-grid z-score means the top-attribution bins are denser in that family than the Dolma working-set baseline. The legacy pronoun, mental-state, dialogue, social, and affect composites are unchanged by the lexical enrichment; the LIWC-adjacent dimensions add a wider descriptive layer.

H.2 Top-20 Lexical Profiles and Format Clusters

Figure 36 gives the four-primary-benchmark view over the original headline lexical variables, now using the same top-20 support as the format-cluster and OpenLIWC-style summaries.

SocialIQA is the only primary benchmark whose high-influence bins visibly split between interactional text and documentation-like text; the other primary benchmarks are dominated by expository, legal, documentation, structured-data, and academic-writing bins.

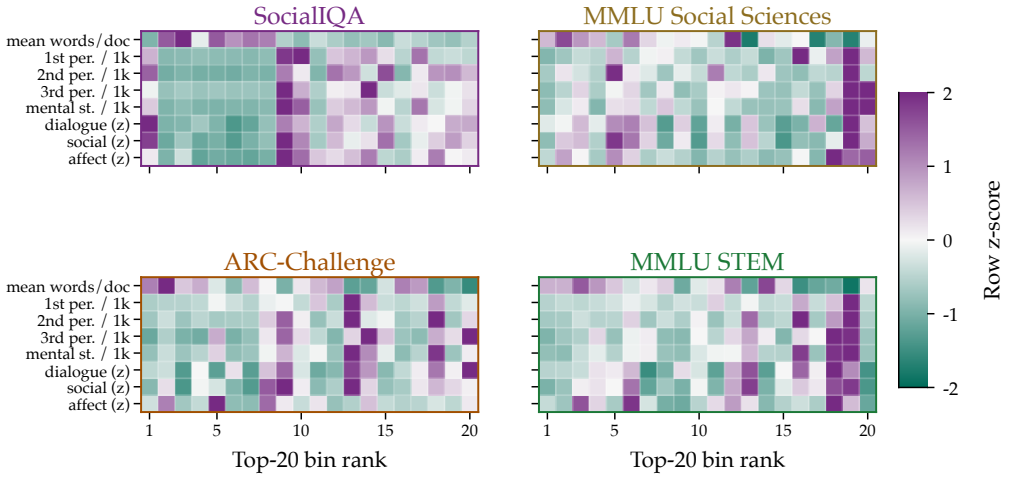


Figure 36: Row-normalized lexical profiles for the top-20 bins of the four primary benchmarks. Cell values are within-row z-scores on a teal–white–purple diverging scale: teal marks lower-than-row-average values, white marks the row mean, and purple marks higher-than-row-average values; SocialIQA is the only primary benchmark whose high-influence bins straddle a clear interactional vs. documentation split.

Format clusters tell us what kind of text attribution surfaced. Figure 37 shows the six-format-cluster composition of each benchmark's top-20 high-influence bins. SocialIQA is bimodal: its top bins include both dialogic/personal text and expository/structured text. The instruct BBH probes are more heterogeneous, but they still place more mass on dialogic and personal formats than the expository targets. By contrast, MMLU Social Sciences, ARC-Challenge, and MMLU STEM are dominated by expository and structured formats. The social-reasoning probe rows extend this inspection to the probe masks and keep main, secondary, caution, and failure controls visually separable. This is not a restatement of the paper's 2×2 design; it is a descriptive result about the communicative forms of the training text surfaced by attribution.

Target-group contrasts test the interpersonal signature directly. Figure 38 pools the top-20 bins within each benchmark target group and bootstraps group-mean differences (B=1000, seed 42). The question is simple: compared with expository targets, do social and commonsense targets surface shorter, more interpersonal, more affective text? Yes. Both the social-reasoning and commonsense-reasoning groups differ from the expository comparison group with 95% CIs excluding zero on all five features. Social vs. expository has higher mental-state density ($\Delta = +2.71$), dialogue ($\Delta z = +0.58$), social language ($\Delta z = +0.67$), and affect ($\Delta z = +0.62$), with shorter documents ($\Delta = -1{,}444$ words/doc). Commonsense vs. expository is comparable or stronger. Social and commonsense reasoning do *not* separate from each other on these features; their CIs include zero. The result is a lexical contrast between interpersonal/pragmatic targets and expository targets.

H.3 OpenLIWC-Style Result Surfaces

The OpenLIWC-style approximation appears in four places in this appendix. Table 23 audits what is implemented and what is declared non-parity. Figure 39 shows the four open summary-variable proxies. Figures 40–41 summarize the family-level loadings, and Figure 42 opens the family averages into the highest-separation realized dimensions. Together, these plots answer two questions: whether the top-attribution bins are above or below the

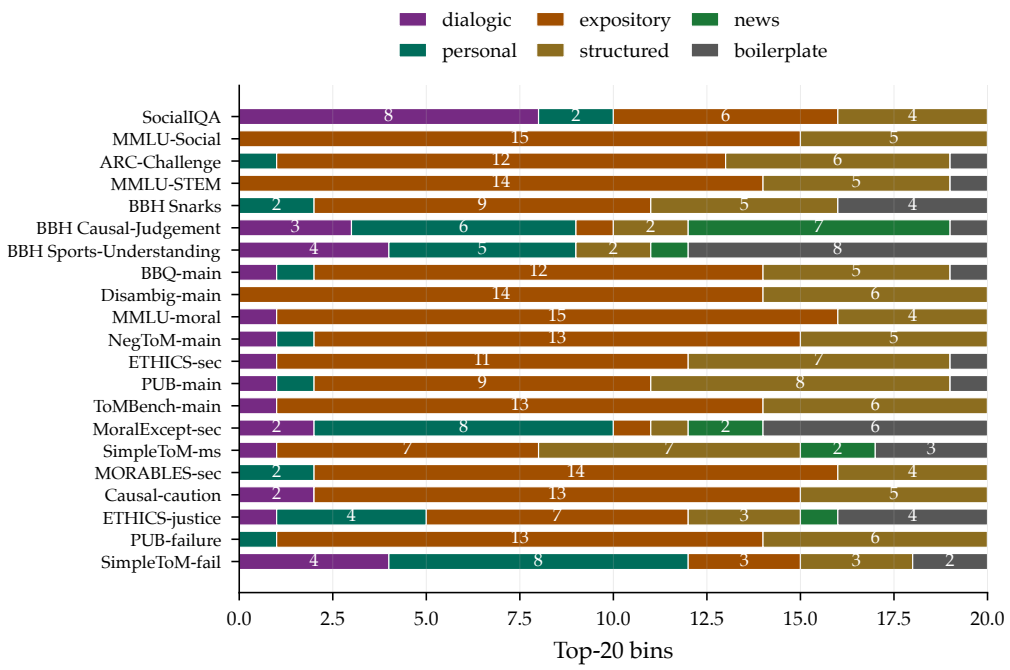


Figure 37: Format-cluster composition of the top-20 high-influence bins for the scoped lexical roster. The six colors encode dialogic, personal, expository, structured, news, and boilerplate WebOrganizer formats. These clusters describe communicative function of the format, not the benchmark target groups used in Figure 38. Cluster assignment is the benchmark-agnostic map in `configs/rq4_format_clusters.yaml`.

Dolma baseline, and whether that lexical direction changes across the original target groups and the tiered social-reasoning probe masks.

LIWC-adjacent families broaden the same story. Figure 39 reports the four open summary variables per benchmark. Expository targets score higher on *Analytic-open*; social and commonsense targets are less analytic/formal and closer to the tone and interpersonal patterns expected from the composite analysis above. Figures 40–41 expose the broader LIWC-family loadings, with exact values printed in every cell. At the benchmark level, the expository targets are consistently below the Dolma working-set baseline on function-word, cognition, affect, social, and perception families. At the target-group level, social and commonsense reasoning sit near or above the corpus average on the social family, whereas expository targets are negative across every family, with drives essentially at baseline (-0.02). The social-reasoning probe tiers are also shown as diagnostics: main and secondary masks are below the corpus baseline on most OpenLIWC-style families, and the failure controls remain separate rather than being pooled into the main tier. Figure 42 shows which realized dimensions carry the strongest target-group separation. Table 24 summarizes the intended claim tier for each target group before the row-level top-bin tables.

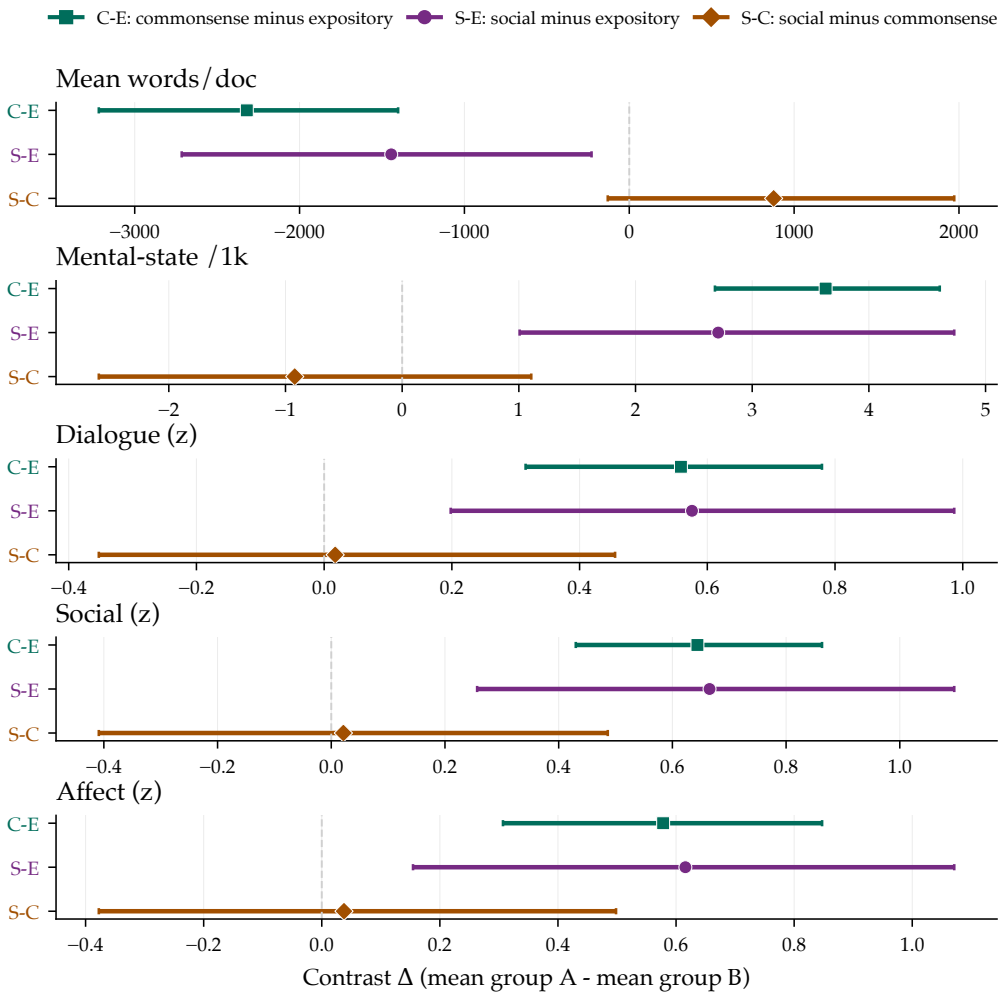


Figure 38: Pooled target-group contrasts (Δ = mean group A − mean group B over top-20 bins), with 95% document-bootstrap confidence intervals. The legend defines the compact contrast codes: C–E is commonsense reasoning minus expository targets, S–E is social reasoning minus expository targets, and S–C is social reasoning minus commonsense reasoning. Intervals crossing the dashed zero line include zero.

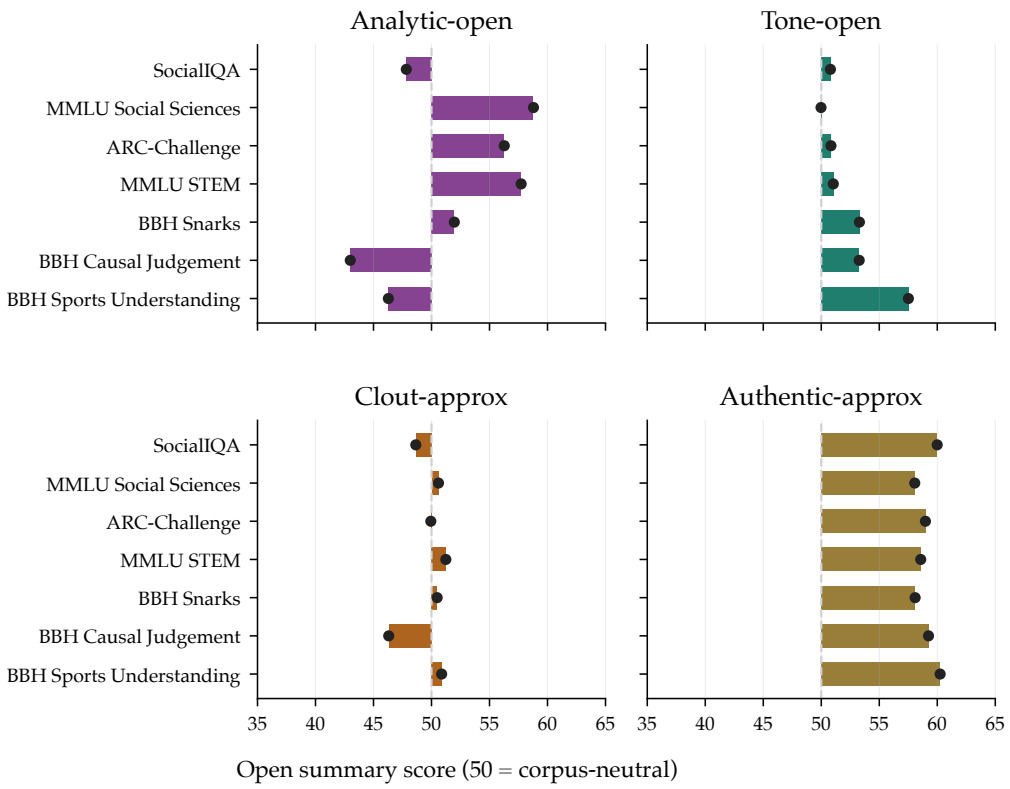


Figure 39: OpenLIWC-style summary-variable proxies (Analytic-open, Tone-open, Clout-approx, Authentic-approx) per benchmark, averaged over each benchmark's top-20 high-influence bins. Scores are open, LIWC-adjacent approximations on a $[0, 100]$ scale ($50 =$ corpus-neutral) that follow the intended construct directions; they are not LIWC.app outputs.

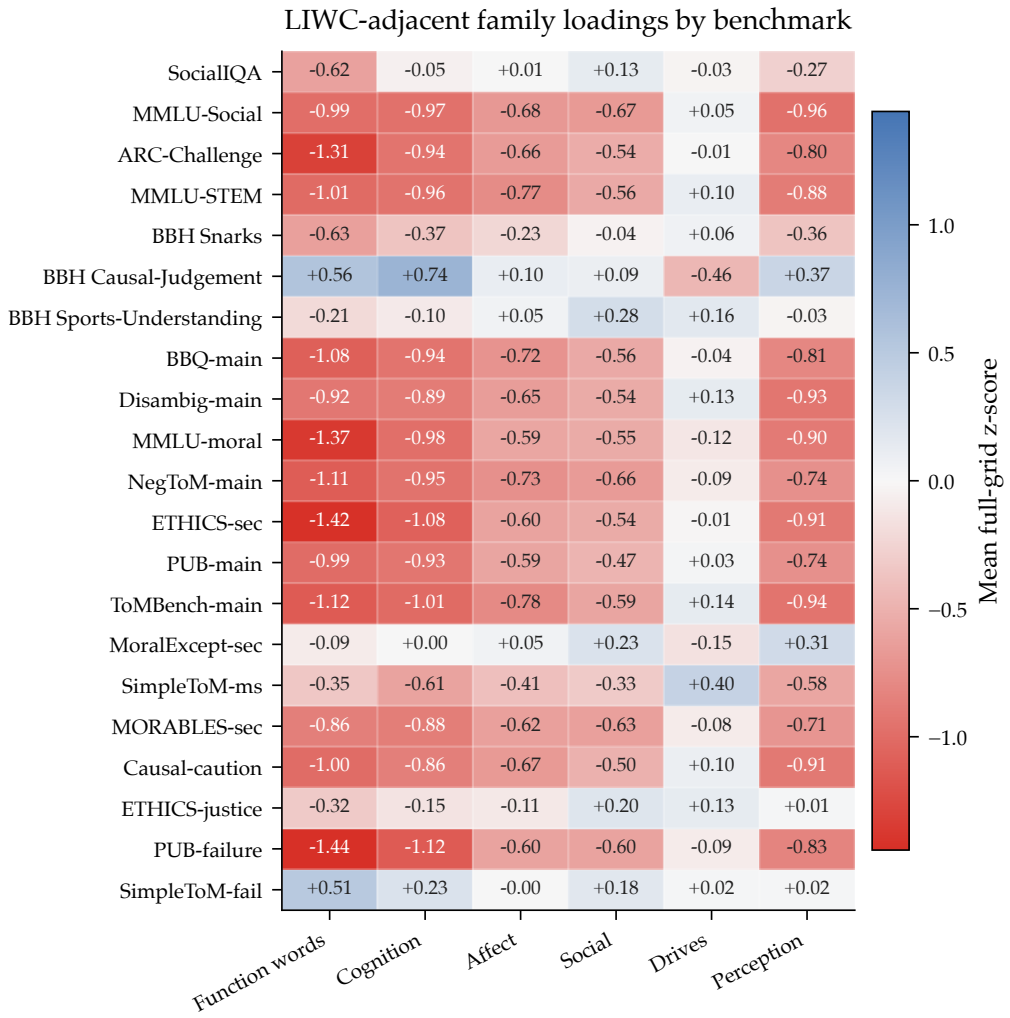


Figure 40: Benchmark-level OpenLIWC-style family loadings. Each cell is the mean full-grid z-score over a benchmark's top-20 high-influence bins using open, LIWC-adjacent approximations rather than LIWC.app outputs; blue cells are above the Dolma working-set baseline and red cells are below it. Numeric labels make the figure readable without color.

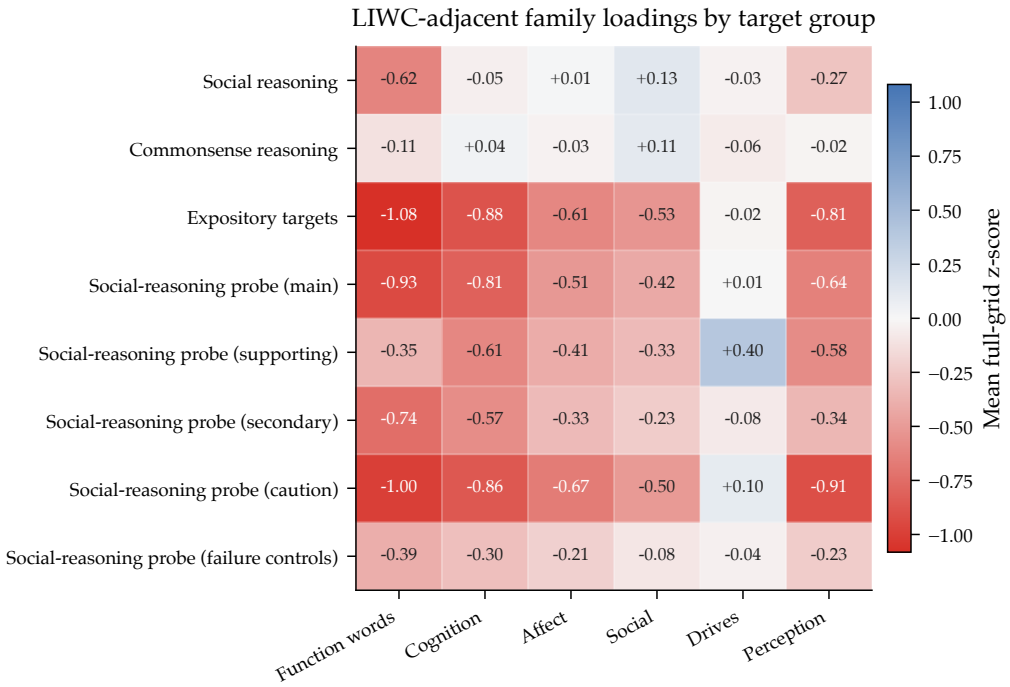


Figure 41: Grouped OpenLIWC-style family loadings over pooled top-20 high-influence bins. The values are open, LIWC-adjacent approximations rather than LIWC.app outputs; blue cells are above the Dolma working-set baseline and red cells are below it. Printed values show the mean full-grid z-score.

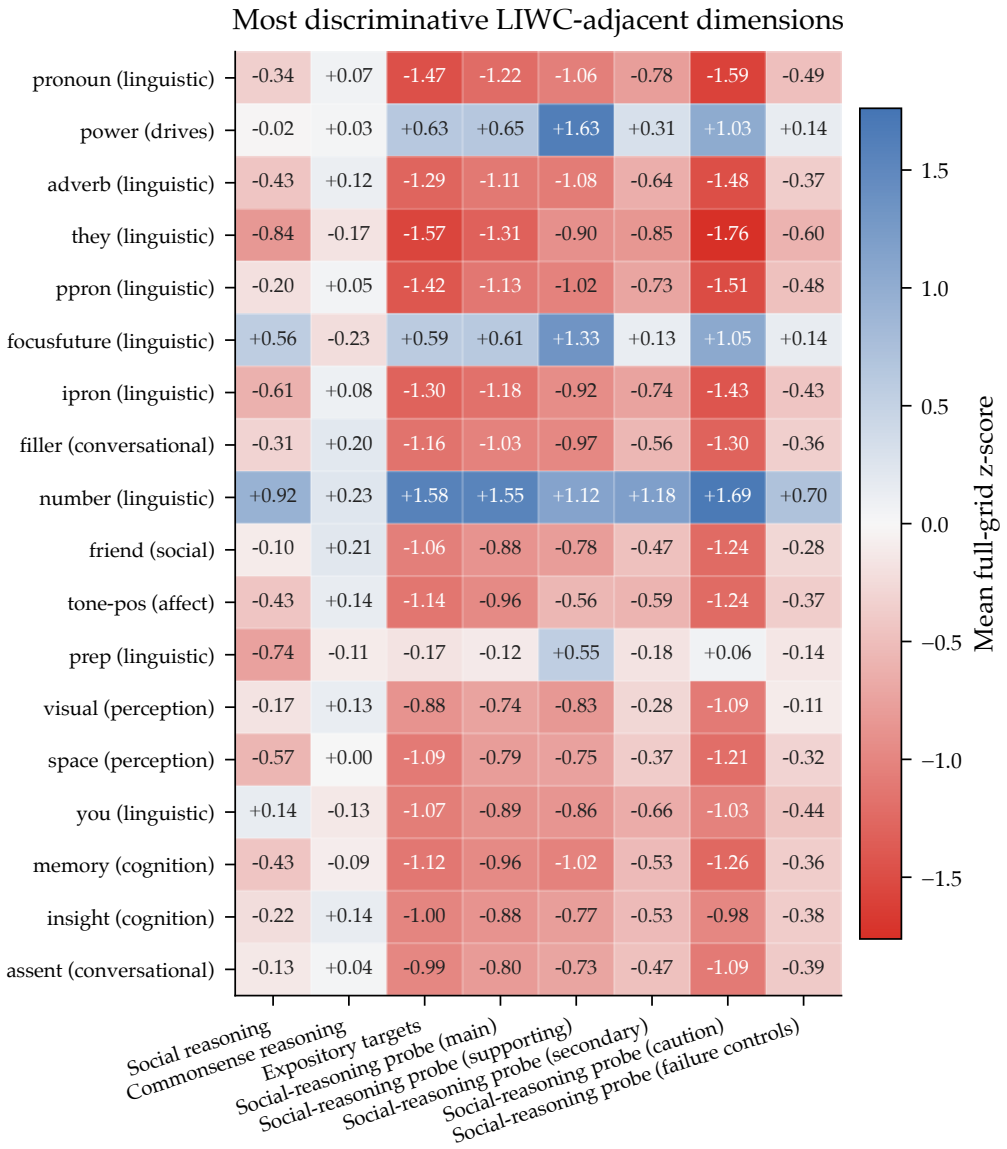


Figure 42: Dimension-level OpenLIWC-style loadings for the 18 realized LIWC-adjacent dimensions with the largest target-group range over pooled top-20 bins. Blue cells are above the Dolma working-set baseline and red cells are below it; printed values show the mean full-grid z-score. This figure exposes the dimensions behind the family-level summaries in Figures 40–41; values are not LIWC.app outputs.

Target group	Observed lexical pattern	Strongest evidence	Narrative takeaway
Social reasoning	Original SocialIQA target: mixed dialogic/personal bins and documentation-like bins.	Figure 41 gives social-family $z = +0.13$.	Use as the original social-reasoning lexical profile, not as held-out social-reasoning probe evidence.
Commonsense reasoning	Original BBH instruct probes: pragmatic and commonsense diagnostics.	Figure 41 gives social-family $z = +0.11$.	Keep separate from the social-reasoning probe tiers in the written claims.
Expository targets	MMLU-Social, ARC-Challenge, and MMLU-STEM are dominated by expository and structured bins.	Figure 41 is below baseline on cognition ($z = -0.88$), affect ($z = -0.61$), social ($z = -0.53$), and perception ($z = -0.81$).	The expository group supplies the baseline contrast: high-attribution bins look more like formal reference/documentation text than interpersonal discourse.
Social-reasoning probe (main)	BBQ, Disambiguation QA, ToMBench, moral/humanities MMLU, NegotiationToM, and PUB excluding pub2/pub3.	Figure 41 gives social-family $z = -0.42$ and affect-family $z = -0.51$.	Treat this tier as the main social-reasoning probe lexical evidence.
Social-reasoning probe (supporting)	SimpleToM mental-state is retained as a supporting diagnostic.	Figure 41 gives social-family $z = -0.33$ and affect-family $z = -0.41$.	Use as relevant mental-state evidence without upgrading the full SimpleToM suite.
Social-reasoning probe (secondary)	ETHICS non-justice, MoralExceptQA/RBQA, and MORABLES.	Figure 41 gives social-family $z = -0.23$ and affect-family $z = -0.33$.	Use as secondary diagnostics, not as the headline evidence tier.
Social-reasoning probe (caution)	BBH Causal Judgment is retained as a caution diagnostic.	Figure 41 gives social-family $z = -0.50$.	Keep this tier visually available but avoid treating it as positive main evidence.
Social-reasoning probe (failure controls)	ETHICS justice, PUB pub2/pub3, and SimpleToM behavior/judgment.	Figure 41 gives social-family $z = -0.08$.	Interpret as controls and diagnostics only.

Table 24: Reader-facing interpretation of the lexical appendix. The target groups are benchmark groupings used for pooling and contrasts; the social-reasoning probe tiers separate main evidence, supporting diagnostics, secondary diagnostics, caution diagnostics, and failure controls. The format clusters in Figure 37 are a separate communicative-function taxonomy over WebOrganizer formats. Exact OpenLIWC-style family and dimension values are printed in Figures 40–42; they are open, LIWC-adjacent approximations rather than LIWC.app outputs.

H.4 Per-Benchmark Top-20 Bin Tables

Tables 25–45 give the row-level top-20 lexical evidence behind the figures for every benchmark in the scoped lexical roster. We include the four primary benchmarks, the three instruct BBH probes, and the 14 social-reasoning probe masks. Base-model BBH tables remain legacy support files and are not part of the current appendix narrative. Each table uses the same six format clusters as Figure 37, so the reader can move directly from aggregate composition to the actual bins.

For SocialIQA (Table 25), the split is clean and bimodal: the top-20 divides 10/10 into interactional (dialogic + personal) and expository (expository + structured) formats. The interactional rows contain short, dialogue-rich text with higher first-/second-person pronoun density and mental-state language, while the expository and structured rows contain longer, formal documentation-like text. Literature/Customer Support remains the dominant outlier in the correctness-differential analysis (Appendix F), pointing to a single training-data substrate whose effect reverses between social reasoning and factual recall.

	Bin	Docs	Mean words	1st per. per 1k	2nd per. per 1k	3rd per. per 1k	Mental st. per 1k	Dialogue (z)	Social (z)	Affect (z)
	Dialogic formats									
1	Literature / Customer Support	10,000	245	26.89	33.39	9.20	6.22	+2.38	+1.82	+0.13
9	Social Life / Q&A Forum	10,000	1,148	46.36	30.01	33.99	15.57	+1.07	+2.10	+2.44
12	Adult / FAQ	10,000	1,034	23.87	31.29	10.85	6.14	+0.76	+0.69	+0.27
13	Fashion & Beauty / Q&A Forum	10,000	622	26.33	26.43	11.60	6.81	+0.35	+0.03	+0.51
15	Adult / Customer Support	9,493	433	20.07	36.00	5.81	4.24	+0.91	+1.37	-0.41
17	Art & Design / Comment Section	10,000	1,202	36.77	16.15	11.79	9.43	+0.22	+0.02	+0.11
18	Health / Customer Support	10,000	830	9.00	26.97	6.03	3.46	+0.12	+0.51	+1.10
19	Crime & Law / Customer Support	10,000	637	10.08	27.78	9.35	4.27	+0.72	+0.96	+0.03
	Personal formats									
10	Social Life / About (Personal)	10,000	508	57.48	16.21	17.61	10.40	+0.65	+1.17	+1.24
14	Home & Hobbies / Creative Writing	10,000	870	31.00	10.37	45.45	8.16	-0.18	+0.33	+1.10
	Expository formats									
2	Industrial / Documentation	10,000	5,738	1.46	0.61	1.46	0.64	-0.66	-0.93	-0.89
3	Health / Documentation	10,000	7,473	1.87	1.95	2.58	1.07	-0.62	-0.53	-0.30
6	Fashion & Beauty / Documentation	1,065	4,516	1.89	0.51	1.74	0.71	-0.97	-1.10	-1.04
7	Home & Hobbies / Documentation	10,000	5,151	1.38	1.88	1.86	0.69	-0.82	-0.82	-1.08
8	Science & Technology / Documentation	10,000	4,953	2.37	1.44	1.58	0.92	-0.50	-0.63	-0.93
20	Politics / Tutorial	10,000	1,200	8.54	22.37	12.32	5.79	+0.74	+0.47	+0.12
	Structured formats									
4	Industrial / Structured Data	10,000	2,074	2.94	0.93	1.71	0.61	-0.42	-0.94	-1.10
5	Industrial / Legal Notices	10,000	5,748	2.22	0.76	2.26	0.58	-0.65	-0.84	-0.98
11	Health / Structured Data	10,000	1,559	3.17	2.11	3.55	1.45	-0.33	-0.48	+0.44
16	Politics / Structured Data	10,000	1,557	7.96	2.55	8.00	2.53	-0.16	-0.50	-0.55

Table 25: Lexical profiles of the top-20 **SocialIQA** bins by signed mean influence. Densities are per 1,000 words; corpus-wide weighted means are 16.1 first-person pronouns, 11.4 second-person pronouns, and 5.1 mental-state terms per 1,000 words. Dialogue, social, and affect columns are full-grid z-scores against all 576 WebOrganizer bins. Rows use the six format clusters in `configs/rq4_format_clusters.yaml`; only clusters present in the benchmark's top-20 bins are shown.

	Bin	Docs	Mean words	1st per. per 1k	2nd per. per 1k	3rd per. per 1k	Mental st. per 1k	Dialogue (z)	Social (z)	Affect (z)
	Expository formats									
1	Industrial / Documentation	10,000	5,738	1.46	0.61	1.46	0.64	-0.66	-0.93	-0.89
2	Health / Documentation	10,000	7,473	1.87	1.95	2.58	1.07	-0.62	-0.53	-0.30
3	Politics / Documentation	10,000	6,264	2.41	1.19	5.13	1.32	-0.44	-0.49	-0.67
6	Crime & Law / Documentation	10,000	6,735	2.19	1.85	5.09	1.44	-0.28	-0.28	-0.51
8	Industrial / Academic Writing	10,000	4,655	3.24	0.50	4.51	1.57	-0.98	-1.01	-0.85
9	Science & Technology / Documentation	10,000	4,953	2.37	1.44	1.58	0.92	-0.50	-0.63	-0.93
10	Food & Dining / Documentation	8,381	4,706	1.87	1.01	2.13	0.70	-0.95	-1.08	-1.11
11	Travel & Tourism / Documentation	816	3,963	2.48	3.56	4.24	1.24	-0.58	-0.58	-0.92
12	History & Geography / Documentation	2,439	7,845	2.08	0.94	4.54	0.93	-0.90	-0.93	-1.04
14	Home & Hobbies / Documentation	10,000	5,151	1.38	1.88	1.86	0.69	-0.82	-0.82	-1.08
15	Fashion & Beauty / Documentation	1,065	4,516	1.89	0.51	1.74	0.71	-0.97	-1.10	-1.04
16	Science & Technology / Academic Writing	10,000	4,758	7.38	0.33	2.82	1.42	-0.89	-0.92	-0.69
18	Health / Academic Writing	10,000	3,308	4.58	0.83	5.13	1.83	-0.94	-0.71	+0.66
19	Religion / Documentation	1,884	2,088	8.22	6.96	13.27	3.48	+0.15	+0.04	-0.04
20	Crime & Law / Academic Writing	10,000	4,606	4.19	0.90	11.52	2.66	-0.46	-0.44	-0.00
	Structured formats									
4	Industrial / Legal Notices	10,000	5,748	2.22	0.76	2.26	0.58	-0.65	-0.84	-0.98
5	Health / Legal Notices	10,000	3,730	4.27	4.80	3.99	1.40	-0.31	-0.09	-0.21
7	Science & Technology / Legal Notices	10,000	5,285	3.02	1.33	2.02	0.92	-0.46	-0.56	-0.96
13	Politics / Legal Notices	10,000	1,938	2.40	0.67	4.19	0.70	-0.47	-0.51	-0.95
17	Industrial / Structured Data	10,000	2,074	2.94	0.93	1.71	0.61	-0.42	-0.94	-1.10

Table 26: Lexical profiles of the top-20 **MMLU Social Sciences** bins by signed mean influence. Densities are per 1,000 words. Dialogue, social, and affect columns are full-grid z-scores against all 576 WebOrganizer bins. Rows use the six format clusters in `configs/rq4_format_clusters.yaml`; only clusters present in the benchmark's top-20 bins are shown.

	Bin	Docs	Mean words	1st per. per 1k	2nd per. per 1k	3rd per. per 1k	Mental st. per 1k	Dialogue (z)	Social (z)	Affect (z)
	Personal formats									
17	Fashion & Beauty / User Review	10,000	493	44.53	15.62	12.00	7.46	-0.39	-0.29	+0.77
	Expository formats									
1	Industrial / Documentation	10,000	5,738	1.46	0.61	1.46	0.64	-0.66	-0.93	-0.89
2	Health / Documentation	10,000	7,473	1.87	1.95	2.58	1.07	-0.62	-0.53	-0.30
4	Politics / Documentation	10,000	6,264	2.41	1.19	5.13	1.32	-0.44	-0.49	-0.67
6	Industrial / Academic Writing	10,000	4,655	3.24	0.50	4.51	1.57	-0.98	-1.01	-0.85
7	Travel & Tourism / Documentation	816	3,963	2.48	3.56	4.24	1.24	-0.58	-0.58	-0.92
10	Science & Technology / Academic Writing	10,000	4,758	7.38	0.33	2.82	1.42	-0.89	-0.92	-0.69
11	Home & Hobbies / Documentation	10,000	5,151	1.38	1.88	1.86	0.69	-0.82	-0.82	-1.08
12	Crime & Law / Documentation	10,000	6,735	2.19	1.85	5.09	1.44	-0.28	-0.28	-0.51
13	Science & Technology / Documentation	10,000	4,953	2.37	1.44	1.58	0.92	-0.50	-0.63	-0.93
14	Science & Technology / Tutorial	10,000	1,266	9.68	11.90	5.08	3.38	-0.05	-0.40	-0.87
15	Literature / Documentation	7,164	1,526	9.42	4.86	7.84	2.40	+1.13	+0.56	-0.68
16	Food & Dining / Documentation	8,381	4,706	1.87	1.01	2.13	0.70	-0.95	-1.08	-1.11
	Structured formats									
3	Industrial / Legal Notices	10,000	5,748	2.22	0.76	2.26	0.58	-0.65	-0.84	-0.98
5	Industrial / Structured Data	10,000	2,074	2.94	0.93	1.71	0.61	-0.42	-0.94	-1.10
8	Science & Technology / Legal Notices	10,000	5,285	3.02	1.33	2.02	0.92	-0.46	-0.56	-0.96
9	Health / Legal Notices	10,000	3,730	4.27	4.80	3.99	1.40	-0.31	-0.09	-0.21
19	Travel & Tourism / Structured Data	10,000	362	4.88	6.54	5.02	1.85	-0.09	-0.61	-1.10
20	Health / Structured Data	10,000	1,559	3.17	2.11	3.55	1.45	-0.33	-0.48	+0.44
	Boilerplate formats									
18	Software Development / About (Org.)	10,000	513	31.47	14.92	9.74	4.44	-0.17	+0.12	-0.60

Table 27: Lexical profiles of the top-20 **ARC-Challenge** bins by signed mean influence. Densities are per 1,000 words. Dialogue, social, and affect columns are full-grid z-scores against all 576 WebOrganizer bins. Rows use the six format clusters in `configs/rq4_format_clusters.yaml`; only clusters present in the benchmark's top-20 bins are shown.

	Bin	Docs	Mean words	1st per. per 1k	2nd per. per 1k	3rd per. per 1k	Mental st. per 1k	Dialogue (z)	Social (z)	Affect (z)
	Expository formats									
1	Industrial / Documentation	10,000	5,738	1.46	0.61	1.46	0.64	-0.66	-0.93	-0.89
3	Health / Documentation	10,000	7,473	1.87	1.95	2.58	1.07	-0.62	-0.53	-0.30
4	Politics / Documentation	10,000	6,264	2.41	1.19	5.13	1.32	-0.44	-0.49	-0.67
7	Industrial / Academic Writing	10,000	4,655	3.24	0.50	4.51	1.57	-0.98	-1.01	-0.85
8	Home & Hobbies / Documentation	10,000	5,151	1.38	1.88	1.86	0.69	-0.82	-0.82	-1.08
10	Travel & Tourism / Documentation	816	3,963	2.48	3.56	4.24	1.24	-0.58	-0.58	-0.92
11	Science & Technology / Academic Writing	10,000	4,758	7.38	0.33	2.82	1.42	-0.89	-0.92	-0.69
12	Crime & Law / Documentation	10,000	6,735	2.19	1.85	5.09	1.44	-0.28	-0.28	-0.51
13	Adult / Documentation	935	1,856	4.66	8.76	6.47	2.13	+0.01	+0.10	-0.71
14	History & Geography / Documentation	2,439	7,845	2.08	0.94	4.54	0.93	-0.90	-0.93	-1.04
15	Science & Technology / Documentation	10,000	4,953	2.37	1.44	1.58	0.92	-0.50	-0.63	-0.93
16	Science & Technology / Tutorial	10,000	1,266	9.68	11.90	5.08	3.38	-0.05	-0.40	-0.87
18	Religion / Documentation	1,884	2,088	8.22	6.96	13.27	3.48	+0.15	+0.04	-0.04
20	Food & Dining / Documentation	8,381	4,706	1.87	1.01	2.13	0.70	-0.95	-1.08	-1.11
	Structured formats									
2	Industrial / Legal Notices	10,000	5,748	2.22	0.76	2.26	0.58	-0.65	-0.84	-0.98
5	Science & Technology / Legal Notices	10,000	5,285	3.02	1.33	2.02	0.92	-0.46	-0.56	-0.96
6	Health / Legal Notices	10,000	3,730	4.27	4.80	3.99	1.40	-0.31	-0.09	-0.21
9	Industrial / Structured Data	10,000	2,074	2.94	0.93	1.71	0.61	-0.42	-0.94	-1.10
17	Politics / Legal Notices	10,000	1,938	2.40	0.67	4.19	0.70	-0.47	-0.51	-0.95
	Boilerplate formats									
19	Software Development / About (Org.)	10,000	513	31.47	14.92	9.74	4.44	-0.17	+0.12	-0.60

Table 28: Lexical profiles of the top-20 **MMLU STEM** bins by signed mean influence. Densities are per 1,000 words. Dialogue, social, and affect columns are full-grid z-scores against all 576 WebOrganizer bins. Rows use the six format clusters in `configs/rq4_format_clusters.yaml`; only clusters present in the benchmark's top-20 bins are shown.

	Bin	Docs	Mean words	1st per. per 1k	2nd per. per 1k	3rd per. per 1k	Mental st. per 1k	Dialogue (z)	Social (z)	Affect (z)
	Personal formats									
1	Adult / User Review	10,000	1,192	25.37	19.47	13.96	7.39	+0.00	+0.66	+0.41
20	Food & Dining / Creative Writing	10,000	967	30.09	11.15	32.53	7.34	+0.15	-0.07	+0.40
	Expository formats									
2	Science & Technology / Academic Writing	10,000	4,758	7.38	0.33	2.82	1.42	-0.89	-0.92	-0.69
3	Crime & Law / Academic Writing	10,000	4,606	4.19	0.90	11.52	2.66	-0.46	-0.44	-0.00
4	Politics / Documentation	10,000	6,264	2.41	1.19	5.13	1.32	-0.44	-0.49	-0.67
5	Home & Hobbies / Documentation	10,000	5,151	1.38	1.88	1.86	0.69	-0.82	-0.82	-1.08
6	Literature / Documentation	7,164	1,526	9.42	4.86	7.84	2.40	+1.13	+0.56	-0.68
7	Health / Academic Writing	10,000	3,308	4.58	0.83	5.13	1.83	-0.94	-0.71	+0.66
10	Politics / Academic Writing	10,000	4,852	6.23	0.90	11.74	3.15	-0.61	-0.68	+0.07
12	Crime & Law / Documentation	10,000	6,735	2.19	1.85	5.09	1.44	-0.28	-0.28	-0.51
13	Adult / Nonfiction Writing	10,000	1,392	14.27	11.19	27.22	6.00	-0.04	+0.39	+0.84
	Structured formats									
9	Politics / Legal Notices	10,000	1,938	2.40	0.67	4.19	0.70	-0.47	-0.51	-0.95
11	Software Development / Content Listing	10,000	1,439	26.77	13.92	4.35	4.72	+0.13	+0.19	-0.99
15	Art & Design / Structured Data	10,000	672	5.45	3.66	10.18	1.90	-0.12	-0.46	-0.62
18	Social Life / Structured Data	10,000	411	21.49	18.77	16.74	6.63	+0.66	+1.04	+1.40
19	Home & Hobbies / Structured Data	10,000	876	3.78	4.40	7.64	1.55	+0.23	-0.59	-0.71
	Boilerplate formats									
8	Software Development / Spam / Ads	10,000	828	11.63	20.75	9.64	4.74	+0.03	+0.43	-0.57
14	Software / Spam / Ads	10,000	859	10.50	26.71	7.75	4.01	+0.41	+0.90	-0.71
16	Games / Spam / Ads	10,000	1,314	9.07	18.85	11.65	3.92	-0.38	+0.40	-0.41
17	Entertainment / Spam / Ads	10,000	1,009	11.41	13.50	19.59	5.07	-0.04	+0.57	+0.17

Table 29: Lexical profiles of the top-20 BBH Snarks (instruct) bins by signed mean influence. Densities are per 1,000 words. Dialogue, social, and affect columns are full-grid z-scores against all 576 WebOrganizer bins. Rows use the six format clusters in `configs/rq4_format_clusters.yaml`; only clusters present in the benchmark's top-20 bins are shown.

	Bin	Docs	Mean words	1st per. per 1k	2nd per. per 1k	3rd per. per 1k	Mental st. per 1k	Dialogue (z)	Social (z)	Affect (z)
	Dialogic formats									
2	Literature / Customer Support	10,000	245	26.89	33.39	9.20	6.22	+2.38	+1.82	+0.13
16	Games / Q&A Forum	10,000	990	19.79	22.06	14.72	6.83	+0.85	+0.16	-0.20
17	Entertainment / Q&A Forum	10,000	1,017	17.72	11.63	28.29	7.88	+1.62	+0.94	+0.20
	Personal formats									
3	History & Geography / User Review	10,000	903	12.05	3.07	18.45	4.77	-0.32	-0.44	+0.11
5	Games / About (Personal)	10,000	1,115	31.63	13.71	22.40	8.40	+0.05	-0.03	+0.23
6	Entertainment / User Review	10,000	952	15.28	5.73	24.95	6.69	+0.27	+0.39	+1.16
10	Home & Hobbies / Creative Writing	10,000	870	31.00	10.37	45.45	8.16	-0.18	+0.33	+1.10
12	Electronics & Hardware / About (Personal)	10,000	787	40.05	12.54	7.33	7.23	+0.25	+0.05	-0.53
19	Travel & Tourism / Personal Blog	10,000	1,367	40.86	10.07	12.72	6.23	-0.52	-0.19	-0.05
	Expository formats									
14	Fashion & Beauty / Documentation	1,065	4,516	1.89	0.51	1.74	0.71	-0.97	-1.10	-1.04
	Structured formats									
8	History & Geography / Listicle	10,000	1,262	4.98	3.53	24.13	3.48	-0.62	-0.42	-0.37
11	Food & Dining / Listicle	10,000	1,065	8.17	19.34	9.26	3.45	-0.72	-0.52	-0.32
	News formats									
1	Entertainment / Audio Transcript	10,000	1,938	47.27	20.35	23.77	13.74	+1.60	+0.71	+0.60
4	Food & Dining / Audio Transcript	10,000	1,580	42.37	18.62	15.63	10.52	+0.47	-0.04	-0.15
7	Fashion & Beauty / Audio Transcript	10,000	1,596	50.69	21.34	17.72	12.77	+0.55	+0.18	+0.48
9	Sports & Fitness / Audio Transcript	10,000	1,695	45.01	20.43	23.54	13.31	+0.54	-0.10	+0.24
13	Games / Audio Transcript	10,000	1,987	38.37	18.43	14.60	11.09	+0.95	+0.02	+0.02
15	History & Geography / News Article	10,000	786	10.28	2.04	19.72	3.95	-0.46	-0.59	-0.39
18	Entertainment / News Article	10,000	498	13.66	4.50	32.39	6.21	+0.39	+0.52	+0.35
	Boilerplate formats									
20	Religion / Truncated	10,000	231	16.29	13.74	16.29	5.64	+0.34	+0.21	+0.50

Table 30: Lexical profiles of the top-20 BBH Causal Judgment (instruct) bins by signed mean influence. Densities are per 1,000 words. Dialogue, social, and affect columns are full-grid z-scores against all 576 WebOrganizer bins. Rows use the six format clusters in `configs/rq4_format_clusters.yaml`; only clusters present in the benchmark's top-20 bins are shown.

	Bin	Docs	Mean words	1st per. per 1k	2nd per. per 1k	3rd per. per 1k	Mental st. per 1k	Dialogue (z)	Social (z)	Affect (z)
	Dialogic formats									
1	Travel & Tourism / Q&A Forum	10,000	492	27.90	29.76	8.47	6.94	+0.58	+0.76	-0.55
2	History & Geography / Customer Support	7,740	288	18.00	22.72	5.78	4.28	+0.99	+0.74	-0.83
4	Education & Jobs / FAQ	10,000	837	16.32	24.02	8.93	4.23	+0.95	+0.48	-0.68
19	Games / Q&A Forum	10,000	990	19.79	22.06	14.72	6.83	+0.85	+0.16	-0.20
	Personal formats									
9	Games / About (Personal)	10,000	1,115	31.63	13.71	22.40	8.40	+0.05	-0.03	+0.23
11	Religion / About (Personal)	10,000	565	47.25	13.24	22.61	10.49	+0.03	+0.56	+1.69
12	Travel & Tourism / User Review	10,000	799	36.81	10.60	10.33	5.26	-0.21	-0.08	-0.16
14	Software Development / User Review	10,000	778	28.28	13.81	8.77	7.22	+0.17	-0.00	-0.35
15	Politics / User Review	10,000	731	20.03	4.98	19.35	6.74	-0.10	-0.28	+0.79
	Structured formats									
3	Finance & Business / Structured Data	10,000	547	4.60	5.89	4.15	1.90	-0.21	-0.53	-0.78
10	Social Life / Content Listing	10,000	784	24.82	16.37	19.02	6.88	+0.13	+0.96	+1.07
	News formats									
7	Software / News Org.	10,000	648	11.68	13.66	8.50	3.27	+0.10	+0.16	-0.57
	Boilerplate formats									
5	Religion / About (Org.)	10,000	514	20.74	7.07	17.83	5.47	-0.46	-0.10	+0.88
6	Software / About (Org.)	10,000	430	20.70	16.51	7.91	3.75	+0.15	+0.31	-0.53
8	Entertainment / Truncated	10,000	156	16.09	11.45	26.72	5.73	+0.68	+0.59	+0.16
13	Literature / Truncated	10,000	2,153	11.16	14.79	12.39	4.44	+0.13	+0.22	-0.20
16	Social Life / About (Org.)	10,000	402	24.51	10.07	17.00	5.52	-0.44	+0.43	+0.87
17	Science & Technology / Truncated	10,000	726	10.35	11.64	8.98	3.64	-0.08	-0.09	-0.51
18	Social Life / Truncated	10,000	81	20.78	18.59	20.17	7.14	+0.56	+1.28	+1.30
20	Software Development / About (Org.)	10,000	513	31.47	14.92	9.74	4.44	-0.17	+0.12	-0.60

Table 31: Lexical profiles of the top-20 BBH Sports Understanding (instruct) bins by signed mean influence. Densities are per 1,000 words. Dialogue, social, and affect columns are full-grid z-scores against all 576 WebOrganizer bins. Rows use the six format clusters in `configs/rq4_format_clusters.yaml`; only clusters present in the benchmark's top-20 bins are shown.

	Bin	Docs	Mean words	1st per. per 1k	2nd per. per 1k	3rd per. per 1k	Mental st. per 1k	Dialogue (z)	Social (z)	Affect (z)
	Dialogic formats									
5	Literature / Customer Support	10,000	245	26.89	33.39	9.20	6.22	+2.38	+1.82	+0.13
	Personal formats									
16	Industrial / Creative Writing	8,443	2,002	27.00	8.71	34.07	6.44	-0.23	-0.32	+0.13
	Expository formats									
1	Industrial / Documentation	10,000	5,738	1.46	0.61	1.46	0.64	-0.66	-0.93	-0.89
3	Health / Documentation	10,000	7,473	1.87	1.95	2.58	1.07	-0.62	-0.53	-0.30
6	Politics / Documentation	10,000	6,264	2.41	1.19	5.13	1.32	-0.44	-0.49	-0.67
7	Industrial / Academic Writing	10,000	4,655	3.24	0.50	4.51	1.57	-0.98	-1.01	-0.85
8	Home & Hobbies / Documentation	10,000	5,151	1.38	1.88	1.86	0.69	-0.82	-0.82	-1.08
9	Crime & Law / Documentation	10,000	6,735	2.19	1.85	5.09	1.44	-0.28	-0.28	-0.51
10	Fashion & Beauty / Documentation	1,065	4,516	1.89	0.51	1.74	0.71	-0.97	-1.10	-1.04
12	Food & Dining / Documentation	8,381	4,706	1.87	1.01	2.13	0.70	-0.95	-1.08	-1.11
13	Science & Technology / Documentation	10,000	4,953	2.37	1.44	1.58	0.92	-0.50	-0.63	-0.93
17	Science & Technology / Academic Writing	10,000	4,758	7.38	0.33	2.82	1.42	-0.89	-0.92	-0.69
18	Transportation / Documentation	10,000	5,302	1.76	1.02	1.95	0.82	-0.69	-0.84	-1.01
19	History & Geography / Documentation	2,439	7,845	2.08	0.94	4.54	0.93	-0.90	-0.93	-1.04
	Structured formats									
2	Industrial / Structured Data	10,000	2,074	2.94	0.93	1.71	0.61	-0.42	-0.94	-1.10
4	Industrial / Legal Notices	10,000	5,748	2.22	0.76	2.26	0.58	-0.65	-0.84	-0.98
11	Health / Legal Notices	10,000	3,730	4.27	4.80	3.99	1.40	-0.31	-0.09	-0.21
14	Science & Technology / Legal Notices	10,000	5,285	3.02	1.33	2.02	0.92	-0.46	-0.56	-0.96
15	Travel & Tourism / Structured Data	10,000	362	4.88	6.54	5.02	1.85	-0.09	-0.61	-1.10
	Boilerplate formats									
20	Crime & Law / Truncated	10,000	176	7.24	8.73	16.00	3.87	-0.02	-0.10	-0.18

Table 32: Lexical profiles of the top-20 BBQ main bins by signed mean influence. Densities are per 1,000 words. Dialogue, social, and affect columns are full-grid z-scores against all 576 WebOrganizer bins. Rows use the six format clusters in `configs/rq4_format_clusters.yaml`; only clusters present in the benchmark's top-20 bins are shown.

	Bin	Docs	Mean words	1st per. per 1k	2nd per. per 1k	3rd per. per 1k	Mental st. per 1k	Dialogue (z)	Social (z)	Affect (z)
	Expository formats									
1	Politics / Documentation	10,000	6,264	2.41	1.19	5.13	1.32	-0.44	-0.49	-0.67
2	Health / Documentation	10,000	7,473	1.87	1.95	2.58	1.07	-0.62	-0.53	-0.30
3	Industrial / Academic Writing	10,000	4,655	3.24	0.50	4.51	1.57	-0.98	-1.01	-0.85
4	Science & Technology / Documentation	10,000	4,953	2.37	1.44	1.58	0.92	-0.50	-0.63	-0.93
5	Crime & Law / Documentation	10,000	6,735	2.19	1.85	5.09	1.44	-0.28	-0.28	-0.51
7	Science & Technology / Academic Writing	10,000	4,758	7.38	0.33	2.82	1.42	-0.89	-0.92	-0.69
8	Travel & Tourism / Documentation	816	3,963	2.48	3.56	4.24	1.24	-0.58	-0.58	-0.92
10	Home & Hobbies / Documentation	10,000	5,151	1.38	1.88	1.86	0.69	-0.82	-0.82	-1.08
11	Industrial / Documentation	10,000	5,738	1.46	0.61	1.46	0.64	-0.66	-0.93	-0.89
12	Crime & Law / Academic Writing	10,000	4,606	4.19	0.90	11.52	2.66	-0.46	-0.44	-0.00
13	History & Geography / Documentation	2,439	7,845	2.08	0.94	4.54	0.93	-0.90	-0.93	-1.04
16	Health / Academic Writing	10,000	3,308	4.58	0.83	5.13	1.83	-0.94	-0.71	+0.66
17	Food & Dining / Documentation	8,381	4,706	1.87	1.01	2.13	0.70	-0.95	-1.08	-1.11
18	Social Life / Documentation	436	1,122	6.83	11.70	11.01	4.25	+0.33	+1.89	+0.34
	Structured formats									
6	Science & Technology / Legal Notices	10,000	5,285	3.02	1.33	2.02	0.92	-0.46	-0.56	-0.96
9	Industrial / Legal Notices	10,000	5,748	2.22	0.76	2.26	0.58	-0.65	-0.84	-0.98
14	Health / Legal Notices	10,000	3,730	4.27	4.80	3.99	1.40	-0.31	-0.09	-0.21
15	Food & Dining / Legal Notices	10,000	906	7.45	6.27	4.14	1.22	-0.45	-0.41	-0.93
19	Sports & Fitness / Structured Data	10,000	476	4.72	2.69	14.93	1.99	-0.40	-1.04	-0.89
20	Politics / Legal Notices	10,000	1,938	2.40	0.67	4.19	0.70	-0.47	-0.51	-0.95

Table 33: Lexical profiles of the top-20 BBH Disambiguation QA main bins by signed mean influence. Densities are per 1,000 words. Dialogue, social, and affect columns are full-grid z-scores against all 576 WebOrganizer bins. Rows use the six format clusters in `configs/rq4_format_clusters.yaml`; only clusters present in the benchmark's top-20 bins are shown.

	Bin	Docs	Mean words	1st per. per 1k	2nd per. per 1k	3rd per. per 1k	Mental st. per 1k	Dialogue (z)	Social (z)	Affect (z)
	Dialogic formats									
1	Literature / Customer Support	10,000	245	26.89	33.39	9.20	6.22	+2.38	+1.82	+0.13
	Expository formats									
2	Industrial / Documentation	10,000	5,738	1.46	0.61	1.46	0.64	-0.66	-0.93	-0.89
4	Health / Documentation	10,000	7,473	1.87	1.95	2.58	1.07	-0.62	-0.53	-0.30
5	Science & Technology / Documentation	10,000	4,953	2.37	1.44	1.58	0.92	-0.50	-0.63	-0.93
6	Politics / Documentation	10,000	6,264	2.41	1.19	5.13	1.32	-0.44	-0.49	-0.67
7	Adult / Documentation	935	1,856	4.66	8.76	6.47	2.13	+0.01	+0.10	-0.71
8	Crime & Law / Documentation	10,000	6,735	2.19	1.85	5.09	1.44	-0.28	-0.28	-0.51
9	Fashion & Beauty / Documentation	1,065	4,516	1.89	0.51	1.74	0.71	-0.97	-1.10	-1.04
10	Food & Dining / Documentation	8,381	4,706	1.87	1.01	2.13	0.70	-0.95	-1.08	-1.11
11	Industrial / Academic Writing	10,000	4,655	3.24	0.50	4.51	1.57	-0.98	-1.01	-0.85
12	Travel & Tourism / Documentation	816	3,963	2.48	3.56	4.24	1.24	-0.58	-0.58	-0.92
13	Transportation / Documentation	10,000	5,302	1.76	1.02	1.95	0.82	-0.69	-0.84	-1.01
14	Health / Academic Writing	10,000	3,308	4.58	0.83	5.13	1.83	-0.94	-0.71	+0.66
17	History & Geography / Documentation	2,439	7,845	2.08	0.94	4.54	0.93	-0.90	-0.93	-1.04
18	Science & Technology / Academic Writing	10,000	4,758	7.38	0.33	2.82	1.42	-0.89	-0.92	-0.69
20	Crime & Law / Academic Writing	10,000	4,606	4.19	0.90	11.52	2.66	-0.46	-0.44	-0.00
	Structured formats									
3	Industrial / Structured Data	10,000	2,074	2.94	0.93	1.71	0.61	-0.42	-0.94	-1.10
15	Art & Design / Structured Data	10,000	672	5.45	3.66	10.18	1.90	-0.12	-0.46	-0.62
16	Health / Structured Data	10,000	1,559	3.17	2.11	3.55	1.45	-0.33	-0.48	+0.44
19	Politics / Structured Data	10,000	1,557	7.96	2.55	8.00	2.53	-0.16	-0.50	-0.55

Table 34: Lexical profiles of the top-20 MMLU moral/humanities main bins by signed mean influence. Densities are per 1,000 words. Dialogue, social, and affect columns are full-grid z-scores against all 576 WebOrganizer bins. Rows use the six format clusters in `configs/rq4_format_clusters.yaml`; only clusters present in the benchmark's top-20 bins are shown.

	Bin	Docs	Mean words	1st per. per 1k	2nd per. per 1k	3rd per. per 1k	Mental st. per 1k	Dialogue (z)	Social (z)	Affect (z)
	Dialogic formats									
18	Food & Dining / FAQ	10,000	945	8.90	21.81	7.50	2.21	-0.10	-0.78	-0.67
	Personal formats									
14	Fashion & Beauty / Creative Writing	10,000	1,670	28.10	13.45	48.39	10.28	+0.27	+0.24	+1.21
	Expository formats									
1	Politics / Documentation	10,000	6,264	2.41	1.19	5.13	1.32	-0.44	-0.49	-0.67
2	Industrial / Documentation	10,000	5,738	1.46	0.61	1.46	0.64	-0.66	-0.93	-0.89
3	Health / Documentation	10,000	7,473	1.87	1.95	2.58	1.07	-0.62	-0.53	-0.30
4	Home & Hobbies / Documentation	10,000	5,151	1.38	1.88	1.86	0.69	-0.82	-0.82	-1.08
6	Crime & Law / Documentation	10,000	6,735	2.19	1.85	5.09	1.44	-0.28	-0.28	-0.51
7	Fashion & Beauty / Documentation	1,065	4,516	1.89	0.51	1.74	0.71	-0.97	-1.10	-1.04
8	Industrial / Academic Writing	10,000	4,655	3.24	0.50	4.51	1.57	-0.98	-1.01	-0.85
9	Food & Dining / Documentation	8,381	4,706	1.87	1.01	2.13	0.70	-0.95	-1.08	-1.11
10	Science & Technology / Documentation	10,000	4,953	2.37	1.44	1.58	0.92	-0.50	-0.63	-0.93
11	Religion / Documentation	1,884	2,088	8.22	6.96	13.27	3.48	+0.15	+0.04	-0.04
15	Transportation / Documentation	10,000	5,302	1.76	1.02	1.95	0.82	-0.69	-0.84	-1.01
16	History & Geography / Documentation	2,439	7,845	2.08	0.94	4.54	0.93	-0.90	-0.93	-1.04
20	Adult / Documentation	935	1,856	4.66	8.76	6.47	2.13	+0.01	+0.10	-0.71
	Structured formats									
5	Industrial / Structured Data	10,000	2,074	2.94	0.93	1.71	0.61	-0.42	-0.94	-1.10
12	Politics / Legal Notices	10,000	1,938	2.40	0.67	4.19	0.70	-0.47	-0.51	-0.95
13	Transportation / Structured Data	10,000	575	4.45	5.63	4.41	1.19	-0.41	-0.85	-1.10
17	Sports & Fitness / Structured Data	10,000	476	4.72	2.69	14.93	1.99	-0.40	-1.04	-0.89
19	Industrial / Legal Notices	10,000	5,748	2.22	0.76	2.26	0.58	-0.65	-0.84	-0.98

Table 35: Lexical profiles of the top-20 NegotiationToM main bins by signed mean influence. Densities are per 1,000 words. Dialogue, social, and affect columns are full-grid z-scores against all 576 WebOrganizer bins. Rows use the six format clusters in `configs/rq4_format_clusters.yaml`; only clusters present in the benchmark's top-20 bins are shown.

	Bin	Docs	Mean words	1st per. per 1k	2nd per. per 1k	3rd per. per 1k	Mental st. per 1k	Dialogue (z)	Social (z)	Affect (z)
	Dialogic formats									
15	Literature / Customer Support	10,000	245	26.89	33.39	9.20	6.22	+2.38	+1.82	+0.13
	Expository formats									
1	Industrial / Documentation	10,000	5,738	1.46	0.61	1.46	0.64	-0.66	-0.93	-0.89
2	Health / Documentation	10,000	7,473	1.87	1.95	2.58	1.07	-0.62	-0.53	-0.30
5	Politics / Documentation	10,000	6,264	2.41	1.19	5.13	1.32	-0.44	-0.49	-0.67
6	Crime & Law / Documentation	10,000	6,735	2.19	1.85	5.09	1.44	-0.28	-0.28	-0.51
7	Industrial / Academic Writing	10,000	4,655	3.24	0.50	4.51	1.57	-0.98	-1.01	-0.85
11	Science & Technology / Academic Writing	10,000	4,758	7.38	0.33	2.82	1.42	-0.89	-0.92	-0.69
12	Health / Academic Writing	10,000	3,308	4.58	0.83	5.13	1.83	-0.94	-0.71	+0.66
13	Fashion & Beauty / Documentation	1,065	4,516	1.89	0.51	1.74	0.71	-0.97	-1.10	-1.04
16	History & Geography / Documentation	2,439	7,845	2.08	0.94	4.54	0.93	-0.90	-0.93	-1.04
17	Science & Technology / Documentation	10,000	4,953	2.37	1.44	1.58	0.92	-0.50	-0.63	-0.93
18	Transportation / Documentation	10,000	5,302	1.76	1.02	1.95	0.82	-0.69	-0.84	-1.01
	Structured formats									
3	Industrial / Structured Data	10,000	2,074	2.94	0.93	1.71	0.61	-0.42	-0.94	-1.10
4	Industrial / Legal Notices	10,000	5,748	2.22	0.76	2.26	0.58	-0.65	-0.84	-0.98
8	Travel & Tourism / Structured Data	10,000	362	4.88	6.54	5.02	1.85	-0.09	-0.61	-1.10
9	Health / Legal Notices	10,000	3,730	4.27	4.80	3.99	1.40	-0.31	-0.09	-0.21
14	Politics / Legal Notices	10,000	1,938	2.40	0.67	4.19	0.70	-0.47	-0.51	-0.95
19	Health / Structured Data	10,000	1,559	3.17	2.11	3.55	1.45	-0.33	-0.48	+0.44
20	Food & Dining / Structured Data	10,000	683	5.26	4.88	5.65	1.59	-0.38	-0.77	-0.81
	Boilerplate formats									
10	Crime & Law / Truncated	10,000	176	7.24	8.73	16.00	3.87	-0.02	-0.10	-0.18

Table 36: Lexical profiles of the top-20 ETHICS non-justice secondary bins by signed mean influence. Densities are per 1,000 words. Dialogue, social, and affect columns are full-grid z-scores against all 576 WebOrganizer bins. Rows use the six format clusters in `configs/rq4_format_clusters.yaml`; only clusters present in the benchmark's top-20 bins are shown.

	Bin	Docs	Mean words	1st per. per 1k	2nd per. per 1k	3rd per. per 1k	Mental st. per 1k	Dialogue (z)	Social (z)	Affect (z)
	Dialogic formats									
1	Literature / Customer Support	10,000	245	26.89	33.39	9.20	6.22	+2.38	+1.82	+0.13
	Personal formats									
18	Food & Dining / Creative Writing	10,000	967	30.09	11.15	32.53	7.34	+0.15	-0.07	+0.40
	Expository formats									
2	Fashion & Beauty / Documentation	1,065	4,516	1.89	0.51	1.74	0.71	-0.97	-1.10	-1.04
3	Politics / Documentation	10,000	6,264	2.41	1.19	5.13	1.32	-0.44	-0.49	-0.67
4	Science & Technology / Academic Writing	10,000	4,758	7.38	0.33	2.82	1.42	-0.89	-0.92	-0.69
5	Health / Documentation	10,000	7,473	1.87	1.95	2.58	1.07	-0.62	-0.53	-0.30
7	Food & Dining / Documentation	8,381	4,706	1.87	1.01	2.13	0.70	-0.95	-1.08	-1.11
8	Health / Academic Writing	10,000	3,308	4.58	0.83	5.13	1.83	-0.94	-0.71	+0.66
10	Home & Hobbies / Documentation	10,000	5,151	1.38	1.88	1.86	0.69	-0.82	-0.82	-1.08
12	Industrial / Academic Writing	10,000	4,655	3.24	0.50	4.51	1.57	-0.98	-1.01	-0.85
13	Science & Technology / Documentation	10,000	4,953	2.37	1.44	1.58	0.92	-0.50	-0.63	-0.93
	Structured formats									
6	Politics / Legal Notices	10,000	1,938	2.40	0.67	4.19	0.70	-0.47	-0.51	-0.95
9	Industrial / Structured Data	10,000	2,074	2.94	0.93	1.71	0.61	-0.42	-0.94	-1.10
14	Crime & Law / Structured Data	10,000	1,460	7.02	2.02	15.57	2.18	-0.26	-0.22	-0.53
15	Art & Design / Legal Notices	10,000	799	10.95	12.37	5.60	2.46	-0.28	-0.25	-0.67
16	Crime & Law / Legal Notices	10,000	1,768	2.35	1.90	4.46	1.04	-0.41	-0.40	-0.79
17	Food & Dining / Legal Notices	10,000	906	7.45	6.27	4.14	1.22	-0.45	-0.41	-0.93
19	Transportation / Structured Data	10,000	575	4.45	5.63	4.41	1.19	-0.41	-0.85	-1.10
20	Fashion & Beauty / Structured Data	10,000	457	8.98	7.07	10.95	3.09	+0.99	+0.08	-0.26
	Boilerplate formats									
11	Sports & Fitness / About (Org.)	10,000	462	18.58	8.33	16.41	3.77	-0.81	-0.47	-0.04

Table 37: Lexical profiles of the top-20 PUB main excluding pub2/pub3 bins by signed mean influence. Densities are per 1,000 words. Dialogue, social, and affect columns are full-grid z-scores against all 576 WebOrganizer bins. Rows use the six format clusters in `configs/rq4_format_clusters.yaml`; only clusters present in the benchmark's top-20 bins are shown.

	Bin	Docs	Mean words	1st per. per 1k	2nd per. per 1k	3rd per. per 1k	Mental st. per 1k	Dialogue (z)	Social (z)	Affect (z)
	Dialogic formats									
4	Literature / Customer Support	10,000	245	26.89	33.39	9.20	6.22	+2.38	+1.82	+0.13
	Expository formats									
1	Health / Documentation	10,000	7,473	1.87	1.95	2.58	1.07	-0.62	-0.53	-0.30
2	Politics / Documentation	10,000	6,264	2.41	1.19	5.13	1.32	-0.44	-0.49	-0.67
3	Industrial / Documentation	10,000	5,738	1.46	0.61	1.46	0.64	-0.66	-0.93	-0.89
9	Science & Technology / Documentation	10,000	4,953	2.37	1.44	1.58	0.92	-0.50	-0.63	-0.93
10	Home & Hobbies / Documentation	10,000	5,151	1.38	1.88	1.86	0.69	-0.82	-0.82	-1.08
12	Crime & Law / Documentation	10,000	6,735	2.19	1.85	5.09	1.44	-0.28	-0.28	-0.51
13	Fashion & Beauty / Documentation	1,065	4,516	1.89	0.51	1.74	0.71	-0.97	-1.10	-1.04
14	Science & Technology / Academic Writing	10,000	4,758	7.38	0.33	2.82	1.42	-0.89	-0.92	-0.69
15	Food & Dining / Documentation	8,381	4,706	1.87	1.01	2.13	0.70	-0.95	-1.08	-1.11
16	Industrial / Academic Writing	10,000	4,655	3.24	0.50	4.51	1.57	-0.98	-1.01	-0.85
17	History & Geography / Documentation	2,439	7,845	2.08	0.94	4.54	0.93	-0.90	-0.93	-1.04
18	Travel & Tourism / Documentation	816	3,963	2.48	3.56	4.24	1.24	-0.58	-0.58	-0.92
19	Transportation / Documentation	10,000	5,302	1.76	1.02	1.95	0.82	-0.69	-0.84	-1.01
	Structured formats									
5	Industrial / Legal Notices	10,000	5,748	2.22	0.76	2.26	0.58	-0.65	-0.84	-0.98
6	Health / Legal Notices	10,000	3,730	4.27	4.80	3.99	1.40	-0.31	-0.09	-0.21
7	Industrial / Structured Data	10,000	2,074	2.94	0.93	1.71	0.61	-0.42	-0.94	-1.10
8	Politics / Legal Notices	10,000	1,938	2.40	0.67	4.19	0.70	-0.47	-0.51	-0.95
11	Science & Technology / Legal Notices	10,000	5,285	3.02	1.33	2.02	0.92	-0.46	-0.56	-0.96
20	Politics / Structured Data	10,000	1,557	7.96	2.55	8.00	2.53	-0.16	-0.50	-0.55

Table 38: Lexical profiles of the top-20 ToMBench main bins by signed mean influence. Densities are per 1,000 words. Dialogue, social, and affect columns are full-grid z-scores against all 576 WebOrganizer bins. Rows use the six format clusters in `configs/rq4_format_clusters.yaml`; only clusters present in the benchmark's top-20 bins are shown.

	Bin	Docs	Mean words	1st per. per 1k	2nd per. per 1k	3rd per. per 1k	Mental st. per 1k	Dialogue (z)	Social (z)	Affect (z)
	Dialogic formats									
12	Games / Customer Support	10,000	650	9.38	34.83	6.98	3.30	+0.38	+0.44	-0.65
20	Games / Q&A Forum	10,000	990	19.79	22.06	14.72	6.83	+0.85	+0.16	-0.20
	Personal formats									
4	Food & Dining / Creative Writing	10,000	967	30.09	11.15	32.53	7.34	+0.15	-0.07	+0.40
7	Art & Design / Personal Blog	10,000	966	33.65	9.71	14.95	7.51	-0.12	-0.05	+0.32
8	Electronics & Hardware / Creative Writing	10,000	1,349	26.36	12.37	27.50	7.07	+0.94	+0.35	-0.01
9	Sports & Fitness / About (Personal)	10,000	580	25.61	11.78	22.11	6.21	-0.52	-0.28	+0.15
10	Industrial / Creative Writing	8,443	2,002	27.00	8.71	34.07	6.44	-0.23	-0.32	+0.13
14	Adult / User Review	10,000	1,192	25.37	19.47	13.96	7.39	+0.00	+0.66	+0.41
16	Art & Design / About (Personal)	10,000	331	33.17	6.34	22.13	5.31	-0.35	-0.04	+0.49
18	Politics / About (Personal)	10,000	517	24.68	7.00	20.58	6.67	-0.19	-0.14	+0.27
	Expository formats									
1	Social Life / Documentation	436	1,122	6.83	11.70	11.01	4.25	+0.33	+1.89	+0.34
	Structured formats									
19	Art & Design / Structured Data	10,000	672	5.45	3.66	10.18	1.90	-0.12	-0.46	-0.62
	News formats									
2	Social Life / News Article	10,000	526	16.61	5.24	38.55	7.00	+0.01	+1.27	+0.74
3	Home & Hobbies / News Article	10,000	537	9.25	4.34	22.16	4.63	-0.48	-0.25	-0.39
	Boilerplate formats									
5	Travel & Tourism / Truncated	10,000	103	13.95	11.49	8.60	3.57	-0.17	-0.27	-0.69
6	Social Life / Truncated	10,000	81	20.78	18.59	20.17	7.14	+0.56	+1.28	+1.30
11	Food & Dining / Truncated	10,000	114	18.32	16.17	10.43	4.88	+0.03	-0.16	-0.54
13	Entertainment / Truncated	10,000	156	16.09	11.45	26.72	5.73	+0.68	+0.59	+0.16
15	Games / Truncated	10,000	198	22.02	15.28	12.75	5.59	+0.52	+0.05	-0.31
17	Art & Design / Truncated	10,000	241	12.16	10.40	11.70	3.66	-0.02	-0.11	-0.34

Table 39: Lexical profiles of the top-20 MoralExceptQA/RBQA secondary bins by signed mean influence. Densities are per 1,000 words. Dialogue, social, and affect columns are full-grid z-scores against all 576 WebOrganizer bins. Rows use the six format clusters in `configs/rq4_format_clusters.yaml`; only clusters present in the benchmark's top-20 bins are shown.

	Bin	Docs	Mean words	1st per. per 1k	2nd per. per 1k	3rd per. per 1k	Mental st. per 1k	Dialogue (z)	Social (z)	Affect (z)
	Dialogic formats									
1	Literature / Customer Support	10,000	245	26.89	33.39	9.20	6.22	+2.38	+1.82	+0.13
	Expository formats									
4	Politics / Documentation	10,000	6,264	2.41	1.19	5.13	1.32	-0.44	-0.49	-0.67
7	Travel & Tourism / Documentation	816	3,963	2.48	3.56	4.24	1.24	-0.58	-0.58	-0.92
8	Crime & Law / Documentation	10,000	6,735	2.19	1.85	5.09	1.44	-0.28	-0.28	-0.51
10	Industrial / Documentation	10,000	5,738	1.46	0.61	1.46	0.64	-0.66	-0.93	-0.89
11	Politics / Academic Writing	10,000	4,852	6.23	0.90	11.74	3.15	-0.61	-0.68	+0.07
12	Crime & Law / Academic Writing	10,000	4,606	4.19	0.90	11.52	2.66	-0.46	-0.44	-0.00
13	Adult / Academic Writing	5,659	3,764	10.21	3.13	15.81	4.15	-0.36	-0.12	+0.53
	Structured formats									
2	Politics / Legal Notices	10,000	1,938	2.40	0.67	4.19	0.70	-0.47	-0.51	-0.95
3	Industrial / Legal Notices	10,000	5,748	2.22	0.76	2.26	0.58	-0.65	-0.84	-0.98
5	Health / Legal Notices	10,000	3,730	4.27	4.80	3.99	1.40	-0.31	-0.09	-0.21
6	Science & Technology / Legal Notices	10,000	5,285	3.02	1.33	2.02	0.92	-0.46	-0.56	-0.96
9	Crime & Law / Legal Notices	10,000	1,768	2.35	1.90	4.46	1.04	-0.41	-0.40	-0.79
15	Sports & Fitness / Legal Notices	10,000	1,034	9.44	7.45	7.77	2.50	-0.48	-0.31	-0.57
18	Sports & Fitness / Structured Data	10,000	476	4.72	2.69	14.93	1.99	-0.40	-1.04	-0.89
	News formats									
16	Politics / News Org.	10,000	679	11.34	2.73	9.41	2.60	-0.38	-0.44	-0.19
17	History & Geography / Audio Transcript	10,000	2,969	31.18	11.14	30.14	8.72	+0.24	-0.11	-0.10
	Boilerplate formats									
14	Adult / About (Org.)	10,000	422	24.41	16.79	13.70	5.44	-0.14	+0.24	+0.73
19	Travel & Tourism / Truncated	10,000	103	13.95	11.49	8.60	3.57	-0.17	-0.27	-0.69
20	Politics / About (Org.)	10,000	989	13.13	2.05	8.98	2.82	-0.61	-0.51	-0.23

Table 40: Lexical profiles of the top-20 SimpleToM mental-state supporting bins by signed mean influence. Densities are per 1,000 words. Dialogue, social, and affect columns are full-grid z-scores against all 576 WebOrganizer bins. Rows use the six format clusters in `configs/rq4_format_clusters.yaml`; only clusters present in the benchmark's top-20 bins are shown.

	Bin	Docs	Mean words	1st per. per 1k	2nd per. per 1k	3rd per. per 1k	Mental st. per 1k	Dialogue (z)	Social (z)	Affect (z)
	Personal formats									
11	Industrial / Creative Writing	8,443	2,002	27.00	8.71	34.07	6.44	-0.23	-0.32	+0.13
19	Food & Dining / Creative Writing	10,000	967	30.09	11.15	32.53	7.34	+0.15	-0.07	+0.40
	Expository formats									
2	Fashion & Beauty / Documentation	1,065	4,516	1.89	0.51	1.74	0.71	-0.97	-1.10	-1.04
3	Science & Technology / Academic Writing	10,000	4,758	7.38	0.33	2.82	1.42	-0.89	-0.92	-0.69
4	Politics / Documentation	10,000	6,264	2.41	1.19	5.13	1.32	-0.44	-0.49	-0.67
5	Science & Technology / Documentation	10,000	4,953	2.37	1.44	1.58	0.92	-0.50	-0.63	-0.93
6	Crime & Law / Documentation	10,000	6,735	2.19	1.85	5.09	1.44	-0.28	-0.28	-0.51
7	Home & Hobbies / Documentation	10,000	5,151	1.38	1.88	1.86	0.69	-0.82	-0.82	-1.08
8	Health / Documentation	10,000	7,473	1.87	1.95	2.58	1.07	-0.62	-0.53	-0.30
9	Travel & Tourism / Documentation	816	3,963	2.48	3.56	4.24	1.24	-0.58	-0.58	-0.92
10	Food & Dining / Documentation	8,381	4,706	1.87	1.01	2.13	0.70	-0.95	-1.08	-1.11
12	Crime & Law / Academic Writing	10,000	4,606	4.19	0.90	11.52	2.66	-0.46	-0.44	-0.00
14	Sports & Fitness / Documentation	9,177	3,157	2.78	3.07	4.62	1.31	-0.81	-0.89	-0.93
16	Industrial / Documentation	10,000	5,738	1.46	0.61	1.46	0.64	-0.66	-0.93	-0.89
17	History & Geography / Documentation	2,439	7,845	2.08	0.94	4.54	0.93	-0.90	-0.93	-1.04
20	Politics / Academic Writing	10,000	4,852	6.23	0.90	11.74	3.15	-0.61	-0.68	+0.07
	Structured formats									
1	Politics / Legal Notices	10,000	1,938	2.40	0.67	4.19	0.70	-0.47	-0.51	-0.95
13	Fashion & Beauty / Structured Data	10,000	457	8.98	7.07	10.95	3.09	+0.99	+0.08	-0.26
15	Art & Design / Structured Data	10,000	672	5.45	3.66	10.18	1.90	-0.12	-0.46	-0.62
18	Industrial / Structured Data	10,000	2,074	2.94	0.93	1.71	0.61	-0.42	-0.94	-1.10

Table 41: Lexical profiles of the top-20 MORABLES secondary bins by signed mean influence. Densities are per 1,000 words. Dialogue, social, and affect columns are full-grid z-scores against all 576 WebOrganizer bins. Rows use the six format clusters in `configs/rq4_format_clusters.yaml`; only clusters present in the benchmark's top-20 bins are shown.

	Bin	Docs	Mean words	1st per. per 1k	2nd per. per 1k	3rd per. per 1k	Mental st. per 1k	Dialogue (z)	Social (z)	Affect (z)
	Dialogic formats									
6	History & Geography / Customer Support	7,740	288	18.00	22.72	5.78	4.28	+0.99	+0.74	-0.83
7	Literature / Customer Support	10,000	245	26.89	33.39	9.20	6.22	+2.38	+1.82	+0.13
	Expository formats									
1	Industrial / Documentation	10,000	5,738	1.46	0.61	1.46	0.64	-0.66	-0.93	-0.89
3	Politics / Documentation	10,000	6,264	2.41	1.19	5.13	1.32	-0.44	-0.49	-0.67
4	Health / Documentation	10,000	7,473	1.87	1.95	2.58	1.07	-0.62	-0.53	-0.30
9	Industrial / Academic Writing	10,000	4,655	3.24	0.50	4.51	1.57	-0.98	-1.01	-0.85
10	Science & Technology / Academic Writing	10,000	4,758	7.38	0.33	2.82	1.42	-0.89	-0.92	-0.69
11	Crime & Law / Documentation	10,000	6,735	2.19	1.85	5.09	1.44	-0.28	-0.28	-0.51
12	Home & Hobbies / Documentation	10,000	5,151	1.38	1.88	1.86	0.69	-0.82	-0.82	-1.08
13	Health / Academic Writing	10,000	3,308	4.58	0.83	5.13	1.83	-0.94	-0.71	+0.66
14	Science & Technology / Documentation	10,000	4,953	2.37	1.44	1.58	0.92	-0.50	-0.63	-0.93
16	Food & Dining / Documentation	8,381	4,706	1.87	1.01	2.13	0.70	-0.95	-1.08	-1.11
17	Crime & Law / Academic Writing	10,000	4,606	4.19	0.90	11.52	2.66	-0.46	-0.44	-0.00
18	History & Geography / Documentation	2,439	7,845	2.08	0.94	4.54	0.93	-0.90	-0.93	-1.04
19	Transportation / Documentation	10,000	5,302	1.76	1.02	1.95	0.82	-0.69	-0.84	-1.01
	Structured formats									
2	Industrial / Legal Notices	10,000	5,748	2.22	0.76	2.26	0.58	-0.65	-0.84	-0.98
5	Industrial / Structured Data	10,000	2,074	2.94	0.93	1.71	0.61	-0.42	-0.94	-1.10
8	Health / Legal Notices	10,000	3,730	4.27	4.80	3.99	1.40	-0.31	-0.09	-0.21
15	Science & Technology / Legal Notices	10,000	5,285	3.02	1.33	2.02	0.92	-0.46	-0.56	-0.96
20	Politics / Legal Notices	10,000	1,938	2.40	0.67	4.19	0.70	-0.47	-0.51	-0.95

Table 42: Lexical profiles of the top-20 BBH Causal Judgment caution bins by signed mean influence. Densities are per 1,000 words. Dialogue, social, and affect columns are full-grid z-scores against all 576 WebOrganizer bins. Rows use the six format clusters in `configs/rq4_format_clusters.yaml`; only clusters present in the benchmark's top-20 bins are shown.

	Bin	Docs	Mean words	1st per. per 1k	2nd per. per 1k	3rd per. per 1k	Mental st. per 1k	Dialogue (z)	Social (z)	Affect (z)
	Dialogic formats									
1	Literature / Customer Support	10,000	245	26.89	33.39	9.20	6.22	+2.38	+1.82	+0.13
	Personal formats									
4	Fashion & Beauty / User Review	10,000	493	44.53	15.62	12.00	7.46	-0.39	-0.29	+0.77
6	Industrial / Creative Writing	8,443	2,002	27.00	8.71	34.07	6.44	-0.23	-0.32	+0.13
9	Software Development / User Review	10,000	778	28.28	13.81	8.77	7.22	+0.17	-0.00	-0.35
16	Fashion & Beauty / Creative Writing	10,000	1,670	28.10	13.45	48.39	10.28	+0.27	+0.24	+1.21
	Expository formats									
2	Social Life / Documentation	436	1,122	6.83	11.70	11.01	4.25	+0.33	+1.89	+0.34
7	Politics / Documentation	10,000	6,264	2.41	1.19	5.13	1.32	-0.44	-0.49	-0.67
10	Crime & Law / Documentation	10,000	6,735	2.19	1.85	5.09	1.44	-0.28	-0.28	-0.51
13	Software / Nonfiction Writing	10,000	1,214	9.44	16.60	9.90	4.37	+0.24	+0.32	-0.38
17	Literature / Tutorial	10,000	1,116	12.20	27.25	19.08	7.80	+1.54	+1.09	+0.91
18	Crime & Law / Nonfiction Writing	10,000	1,344	9.22	5.09	25.69	4.95	-0.19	-0.01	+0.38
19	Electronics & Hardware / Academic Writing	10,000	1,854	5.36	1.79	4.60	2.08	-0.17	-0.22	-0.87
	Structured formats									
3	Travel & Tourism / Structured Data	10,000	362	4.88	6.54	5.02	1.85	-0.09	-0.61	-1.10
12	Electronics & Hardware / Content Listing	10,000	1,287	9.41	12.91	6.31	3.41	+0.17	+0.15	-0.83
20	Crime & Law / Legal Notices	10,000	1,768	2.35	1.90	4.46	1.04	-0.41	-0.40	-0.79
	News formats									
11	Software / News Org.	10,000	648	11.68	13.66	8.50	3.27	+0.10	+0.16	-0.57
	Boilerplate formats									
5	Social Life / Truncated	10,000	81	20.78	18.59	20.17	7.14	+0.56	+1.28	+1.30
8	Travel & Tourism / Truncated	10,000	103	13.95	11.49	8.60	3.57	-0.17	-0.27	-0.69
14	Fashion & Beauty / Truncated	10,000	63	17.06	16.19	13.53	5.00	+0.28	+0.11	-0.01
15	Food & Dining / Truncated	10,000	114	18.32	16.17	10.43	4.88	+0.03	-0.16	-0.54

Table 43: Lexical profiles of the top-20 ETHICS justice failure control bins by signed mean influence. Densities are per 1,000 words. Dialogue, social, and affect columns are full-grid z-scores against all 576 WebOrganizer bins. Rows use the six format clusters in `configs/rq4_format_clusters.yaml`; only clusters present in the benchmark's top-20 bins are shown.

	Bin	Docs	Mean words	1st per. per 1k	2nd per. per 1k	3rd per. per 1k	Mental st. per 1k	Dialogue (z)	Social (z)	Affect (z)
	Personal formats									
20	History & Geography / About (Personal)	10,000	587	20.45	3.64	18.99	4.23	-0.49	-0.14	-0.45
	Expository formats									
2	Fashion & Beauty / Documentation	1,065	4,516	1.89	0.51	1.74	0.71	-0.97	-1.10	-1.04
3	Politics / Documentation	10,000	6,264	2.41	1.19	5.13	1.32	-0.44	-0.49	-0.67
4	Travel & Tourism / Documentation	816	3,963	2.48	3.56	4.24	1.24	-0.58	-0.58	-0.92
5	Industrial / Academic Writing	10,000	4,655	3.24	0.50	4.51	1.57	-0.98	-1.01	-0.85
6	Science & Technology / Academic Writing	10,000	4,758	7.38	0.33	2.82	1.42	-0.89	-0.92	-0.69
7	Industrial / Documentation	10,000	5,738	1.46	0.61	1.46	0.64	-0.66	-0.93	-0.89
9	Food & Dining / Documentation	8,381	4,706	1.87	1.01	2.13	0.70	-0.95	-1.08	-1.11
10	Health / Academic Writing	10,000	3,308	4.58	0.83	5.13	1.83	-0.94	-0.71	+0.66
11	Health / Documentation	10,000	7,473	1.87	1.95	2.58	1.07	-0.62	-0.53	-0.30
12	Software / Academic Writing	10,000	2,287	7.70	2.94	8.31	2.84	+0.27	+0.46	-0.54
15	Home & Hobbies / Documentation	10,000	5,151	1.38	1.88	1.86	0.69	-0.82	-0.82	-1.08
16	Politics / Academic Writing	10,000	4,852	6.23	0.90	11.74	3.15	-0.61	-0.68	+0.07
17	Crime & Law / Documentation	10,000	6,735	2.19	1.85	5.09	1.44	-0.28	-0.28	-0.51
	Structured formats									
1	Industrial / Structured Data	10,000	2,074	2.94	0.93	1.71	0.61	-0.42	-0.94	-1.10
8	Fashion & Beauty / Structured Data	10,000	457	8.98	7.07	10.95	3.09	+0.99	+0.08	-0.26
13	Art & Design / Structured Data	10,000	672	5.45	3.66	10.18	1.90	-0.12	-0.46	-0.62
14	Industrial / Legal Notices	10,000	5,748	2.22	0.76	2.26	0.58	-0.65	-0.84	-0.98
18	Travel & Tourism / Structured Data	10,000	362	4.88	6.54	5.02	1.85	-0.09	-0.61	-1.10
19	Health / Structured Data	10,000	1,559	3.17	2.11	3.55	1.45	-0.33	-0.48	+0.44

Table 44: Lexical profiles of the top-20 PUB pub2/pub3 failure control bins by signed mean influence. Densities are per 1,000 words. Dialogue, social, and affect columns are full-grid z-scores against all 576 WebOrganizer bins. Rows use the six format clusters in `configs/rq4_format_clusters.yaml`; only clusters present in the benchmark's top-20 bins are shown.

	Bin	Docs	Mean words	1st per. per 1k	2nd per. per 1k	3rd per. per 1k	Mental st. per 1k	Dialogue (z)	Social (z)	Affect (z)
	Dialogic formats									
1	History & Geography / Customer Support	7,740	288	18.00	22.72	5.78	4.28	+0.99	+0.74	-0.83
4	Adult / Customer Support	9,493	433	20.07	36.00	5.81	4.24	+0.91	+1.37	-0.41
10	Art & Design / Q&A Forum	10,000	862	20.83	20.42	8.71	6.78	+0.76	+0.24	-0.28
20	Games / Q&A Forum	10,000	990	19.79	22.06	14.72	6.83	+0.85	+0.16	-0.20
	Personal formats									
7	Adult / User Review	10,000	1,192	25.37	19.47	13.96	7.39	+0.00	+0.66	+0.41
8	Food & Dining / Creative Writing	10,000	967	30.09	11.15	32.53	7.34	+0.15	-0.07	+0.40
11	Art & Design / About (Personal)	10,000	331	33.17	6.34	22.13	5.31	-0.35	-0.04	+0.49
13	Home & Hobbies / Creative Writing	10,000	870	31.00	10.37	45.45	8.16	-0.18	+0.33	+1.10
14	Travel & Tourism / User Review	10,000	799	36.81	10.60	10.33	5.26	-0.21	-0.08	-0.16
15	Sports & Fitness / User Review	10,000	969	30.70	14.36	11.54	6.44	-0.46	-0.52	+0.09
16	Politics / User Review	10,000	731	20.03	4.98	19.35	6.74	-0.10	-0.28	+0.79
17	History & Geography / User Review	10,000	903	12.05	3.07	18.45	4.77	-0.32	-0.44	+0.11
	Expository formats									
2	Social Life / Documentation	436	1,122	6.83	11.70	11.01	4.25	+0.33	+1.89	+0.34
9	Literature / Documentation	7,164	1,526	9.42	4.86	7.84	2.40	+1.13	+0.56	-0.68
12	Adult / Academic Writing	5,659	3,764	10.21	3.13	15.81	4.15	-0.36	-0.12	+0.53
	Structured formats									
3	Sports & Fitness / Legal Notices	10,000	1,034	9.44	7.45	7.77	2.50	-0.48	-0.31	-0.57
6	Crime & Law / Legal Notices	10,000	1,768	2.35	1.90	4.46	1.04	-0.41	-0.40	-0.79
19	Politics / Legal Notices	10,000	1,938	2.40	0.67	4.19	0.70	-0.47	-0.51	-0.95
	Boilerplate formats									
5	Adult / About (Org.)	10,000	422	24.41	16.79	13.70	5.44	-0.14	+0.24	+0.73
18	Games / About (Org.)	10,000	572	20.92	13.49	13.40	4.39	-0.23	+0.08	-0.20

Table 45: Lexical profiles of the top-20 SimpleToM behavior/judgment failure control bins by signed mean influence. Densities are per 1,000 words. Dialogue, social, and affect columns are full-grid z-scores against all 576 WebOrganizer bins. Rows use the six format clusters in `configs/rq4_format_clusters.yaml`; only clusters present in the benchmark's top-20 bins are shown.

I Unlearning Experimental Setup

All unlearning experiments in this paper are performed on OLMo3-7B Base (the unmodified pretrained checkpoint). Baseline benchmark scores against which the accuracy drop γ is measured come from evaluating the same OLMo3-7B Base checkpoint under the OLMES `:mc::olmes` protocol; the post-unlearning evaluation uses the identical protocol on the merged base+adapter checkpoint, so the γ comparison is config-matched. All reported SocialIQA γ values for on-target and random-control conditions use the config-matched OLMES baseline of 0.80334; earlier fast-eval SocialIQA pilots are not used for reported numbers.

I.1 Objective and Optimization

All unlearning experiments use NGDiff (Jin et al., 2025). The update direction normalizes retain and forget gradients independently:

$$g_{\text{update}} = \frac{g_{\text{retain}}}{\|g_{\text{retain}}\|} - \frac{g_{\text{forget}}}{\|g_{\text{forget}}\|},$$

so that neither component dominates regardless of bin size or loss scale. The learning rate is adapted automatically every 10 steps via two additional forward passes on the retain set (Hessian-free schedule). We fine-tune only LoRA parameters (rank 8, applied to Q/K/V/O projections), following the published NGDiff setup for 7B-scale models; the complete hyperparameter configuration appears in Table 46.

Table 46: Unlearning hyperparameters.

Parameter	Value
Method	NGDiff with GeN AutoLR
LoRA Configuration	Rank 8; Q, K, V, O projections
Batch Size	16 (per GPU)
Effective Batch Size	160 (via Gradient Accumulation)
Precision	bf16
Max Steps	5,000

I.2 Data Configuration and Sampling

We standardize the data pipeline across target topics.

Forget Set ($\mathcal{D}_F$) For each single-bin sweep run, we sample exactly 2,000 documents uniformly at random from the target topic. The faithful per-document replication instead uses the top-200 influence-ranked documents per (topic, benchmark) pair, with 1,000-document random in-topic controls.

Retain Set ($\mathcal{D}_R$) The retain set consists of 9,000 documents sampled uniformly at random from outside the target topic (for corpus-global runs, from the remaining corpus disjoint from the forget set). It is sampled once per run and held fixed for the duration of training.

Length-Normalized Loss To prevent long documents from disproportionately biasing the gradient, we apply length-normalization to the loss calculation (SimNPO style; Fan et al., 2025). The per-document loss is averaged over its token count before aggregating into the batch-level mean.

Table 47 summarizes the forget/retain data configuration used for the 2,000-document single-bin sweep; the faithful per-document replication in Appendix J.1 uses top-200 influence-ranked forget sets compared against 1,000-document random in-topic controls.

Table 47: Data and sampling configuration.

Parameter	Value
Forget Set Size (k)	2,000 documents
Retain Set Size	9,000 documents (uniform, out-of-bin)
Retain Refresh Rate	None (fixed for the run)
Forget:Retain Ratio	$\sim$1:4.5

I.3 Stopping Criterion: Randomized-Text Perplexity

We use randomized-text perplexity as an operational stopping rule. For each forget set, we compare the model's perplexity on the original forget documents against shuffled variants that preserve token frequencies while destroying semantic and logical structure:

1. **Baseline Computation:** For each document in $\mathcal{D}_F$, we generate a randomized version by splitting the token sequence into segments and shuffling them.
2. **Target Perplexity ($\text{PPL}_{\text{target}}$):** We calculate the original model's perplexity on this randomized $\mathcal{D}_F$.
3. **Termination:** Unlearning stops when the current model's perplexity on the original (unshuffled) forget set reaches or exceeds $\text{PPL}_{\text{target}}$.

Additionally, we apply a safety guard: if the model's performance on a held-out MMLU subset drops below 90% of the baseline, or if the process reaches 5,000 steps, training terminates immediately.

I.4 Experimental Conditions (Validation)

To test whether attribution scores identify more capability-relevant documents than topic membership alone, we compare two primary conditions within each target topic, measured against a global random baseline:

- **Single-Topic Influence (Ours):** The forget set consists of the $k = 2,000$ documents from the target topic with the highest aggregate influence scores for the target benchmark. The bin-characterization analysis in Appendix H characterizes which bins carry the strongest SocialIQA influence, and the correctness-differential analysis in Appendix F identifies bins whose attribution signal reverses across benchmarks.
- **Single-Topic Random:** The forget set consists of $k = 2,000$ documents sampled uniformly at random from the *same* target topic. This controls for the effect of topic-specific distribution and tests whether specific document selection yields stronger unlearning.

Global Baseline As a secondary control to account for procedural degradation (i.e., the inherent noise introduced by the unlearning process itself), we also conduct **Global Random** experiments. In this condition, the forget set is sampled uniformly from the entire 6T training corpus without regard to topic labels. This baseline allows us to calculate the net effect of topic-specific unlearning: $\gamma_{\text{net}} = \gamma_{\text{topic}} - \gamma_{\text{global}}$.

I.5 Evaluation Protocol

After each unlearning run converges, we merge the LoRA adapter into OLMo3-7B Base and evaluate the resulting base-unlearned checkpoint on the same four benchmarks used for attribution (Table 8). We report raw accuracy damage:

$$\gamma = s_{\text{baseline}} - s_{\text{unlearned}},$$

where s_{baseline} is the score of the unmodified OLMo3-7B Base checkpoint. For accuracy metrics, $\gamma > 0$ indicates capability degradation; $\gamma \approx 0$ indicates the benchmark was

preserved; $\gamma < 0$ indicates an accuracy increase after unlearning. The per-benchmark baseline scores against which γ is computed are reported in Table 48.

Table 48: Baseline scores of the unmodified OLMo3-7B Base checkpoint used as the reference for all γ calculations.

Benchmark	Category	Baseline
SocialIQA	Social / Reasoning	0.80334
MMLU Social Sciences	Social / Knowledge	0.75236
ARC-Challenge	STEM / Reasoning	0.83447
MMLU STEM	STEM / Knowledge	0.59745

J Unlearning Results

Figure 43 summarizes the main validation comparison between influence-targeted and random in-topic unlearning. The remaining appendix figures break this result down by topic, global control, convergence efficiency, and influence–drop diagnostics.

J.1 Paired Significance of the Influence-vs-Random Contrast

To put the Figure 43 pattern on a statistical footing, we test the one-sided hypothesis that influence-targeted unlearning produces a larger accuracy drop than random in-topic unlearning on the intended capability. For each (topic, benchmark, seed) tuple we pair the influence-targeted run (expA, top-200 documents by per-document influence on the target benchmark) with the random in-topic run (exp1, same topic and seed, 1,000 documents sampled uniformly from the same topic), compute the paired difference $d = \gamma_{\text{influence}} - \gamma_{\text{random}}$, and aggregate across all 24 topics and the three unlearning seeds $\{42, 43, 44\}$. Per benchmark we run a Wilcoxon signed-rank test on the pooled (topic $\times$ seed) cells and a paired t-test on the per-topic seed-mean differences; both are one-sided ($d > 0$), and the eight resulting p-values (4 benchmarks $\times$ 2 tests) are Benjamini–Hochberg adjusted within the family. This faithful per-document analysis is separate from the 2,000-document single-bin sweep below; it asks whether a small influence-ranked forget set (200 documents) produces more targeted damage than a five-times-larger random in-topic control, so the contrast is conservative with respect to forget-set size.

What uncertainty is measured. This paired-significance subsection reports bootstrap 95% confidence intervals on the median paired effect d in matched topic–seed cells. Elsewhere in the selectivity appendix, values written as mean $\pm$ standard deviation summarize fixed unlearning seeds, not formal confidence intervals. Robustness tables vary analysis choices such as topic-selection rules and selectivity metrics. We do not estimate uncertainty over alternative sampled forget-document sets.

Table 49 reports the result. The paired contrast is significantly positive on SocialIQA (Wilcoxon BH-adjusted $p \approx 10^{-5}$, median $+1.6$ accuracy points, Cohen's $d_z = +0.39$) and on MMLU STEM ($p_{\text{BH}} = 0.028$, $d_z = +0.23$), although MMLU STEM's bootstrap median interval crosses zero and the paired t-test does not reach significance, so it should be read as weaker evidence. MMLU Social Sciences does not reach BH-adjusted significance under this paired construction. ARC-Challenge also does not support the positive one-sided hypothesis, and its median paired difference is negative, indicating slightly less damage from influence-targeted forgetting than from random in-topic forgetting. These cases track the cross-method finding (Appendix K.1): the per-document forget set does not measurably damage MMLU Social Sciences, and ARC-Challenge's on-target damage sits at the noise floor. We treat both as transparent findings rather than as failures of the underlying paired-contrast claim.

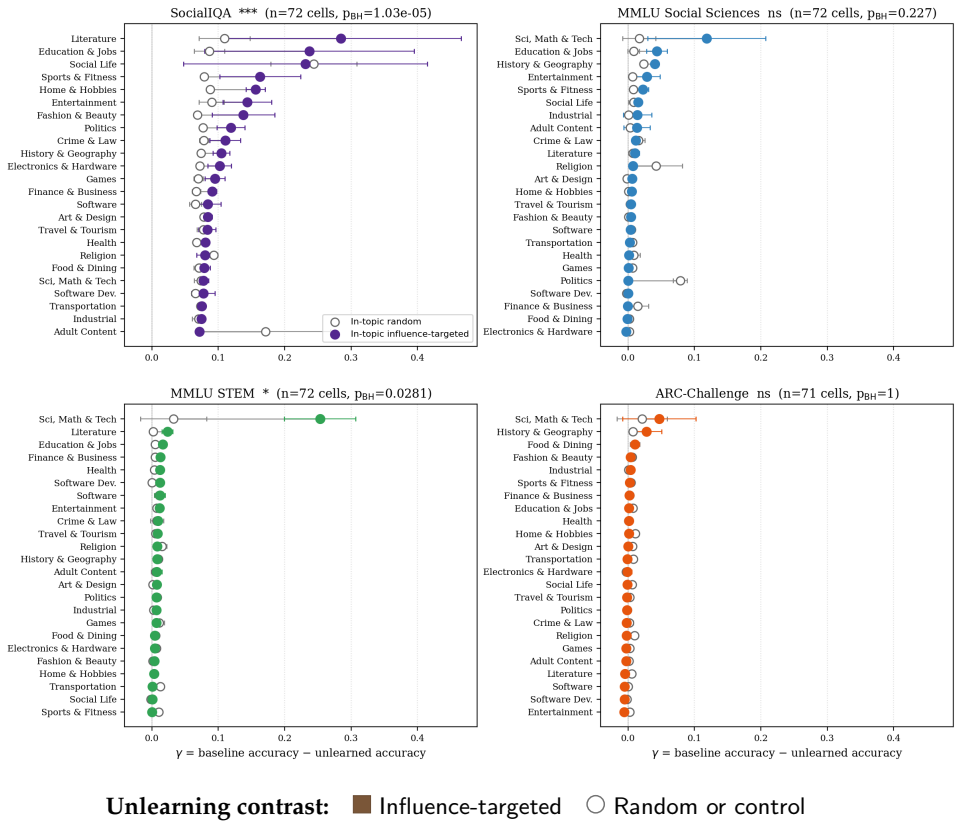


Figure 43: Unlearning validation Targeted machine unlearning validates bin-level attribution. For each WebOrganizer topic and each primary benchmark, $\gamma_{\text{influence}}$ (colored markers: in-topic, top-200 documents by per-document influence on the target benchmark) is compared against γ_{random} (open markers: in-topic, 1,000 random documents, a five-times-larger control forget set), where $\gamma = A_{\text{baseline}} - A_{\text{unlearned}}$ and positive values mean accuracy damage. Markers show the seed-mean across 3 unlearning seeds $\{42, 43, 44\}$; horizontal whiskers show ± 1 standard deviation. Panel titles report the Benjamini–Hochberg-adjusted Wilcoxon signed-rank significance for the paired contrast across the available matched topic–seed cells (24 topics $\times$ 3 seeds; one ARC-Challenge cell unavailable) ($^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$). SocialIQA is the clean paired result; MMLU-STEM is weaker because its Wilcoxon test is significant but its bootstrap median interval crosses zero (Appendix J). The MMLU-SS and ARC-Challenge nulls track the cross-method check in Appendix K.1.

J.2 Single-Bin Topic Sweep

We run the unlearning pipeline independently for each of the 24 topic bins. In each run, the forget set is drawn from a single target topic and the retain set is stratified across the remaining 23 topics. The checkpoint at the selected stopping point is evaluated on all four benchmarks.

Figure 44 presents the per-topic γ values. Rows follow the contrastive design: the two social benchmarks (SocialIQA, MMLU Social Sciences) appear above the horizontal divider, and the two STEM benchmarks (ARC-Challenge, MMLU STEM) below it. Topic columns are sorted by mean γ across benchmarks, and the vertical divider separates the topic interventions from the global random control column.

This diagnostic sweep shows modest topic-level effects relative to the paired influence-vs-random result above. Education & Jobs has the largest mean positive damage across the four benchmarks: +2.7 points on SocialIQA, +2.6 on MMLU Social Sciences, +1.0 on

Table 49: Multi-seed paired-significance test for influence-targeted vs. random in-topic unlearning. For each available (topic, benchmark, seed) cell we compute the paired difference $d = \gamma_{\text{influence}} - \gamma_{\text{random}}$, where $\gamma = (\text{baseline acc} - \text{unlearned acc})$ and a positive d means influence-targeted damages the intended capability more than random in-topic. Across the available matched cells from 24 WebOrganizer topics $\times$ 3 unlearning seeds $\{42, 43, 44\}$, we test the one-sided hypothesis $d > 0$ via (i) Wilcoxon signed-rank on pooled (topic $\times$ seed) cells (primary, non-parametric) and (ii) paired t-test on per-topic seed-mean differences. Benjamini–Hochberg correction is applied across the family of 8 tests. First two rows reach BH-adjusted $p < 0.05$ on the primary Wilcoxon test.

	Wilcoxon (pooled, n_{cells})			**Paired** t	**Effect size**
Target	n	median d [95% CI]	p_{BH}	p_{BH} (n_{topics}=24)	d_z (pooled / per-topic)
SocialIQA	72	**+0.0160 [+0.0130, +0.0220]**	**1.0×10^{-5} *****	**0.022 ***	**+0.39 / +0.56**
MMLU-STEM	72	**+0.0020 [−0.0003, +0.0055]**	**0.028 ***	0.227	+0.23 / +0.25
MMLU-SS	72	+0.0017 [−0.0018, +0.0042]	0.227	0.384	+0.09 / +0.12
ARC-C	71	−0.0034 [−0.0051, −0.0026]	>0.99	0.984	−0.11 / −0.23

Median diff and bootstrap 95% CI (B=2000) are reported in raw accuracy units (e.g., +0.0160 = +1.6 accuracy points). Paired t is one-sided ($d > 0$) on per-topic seed-mean differences. Cohen's d_z is reported as pooled-cells / per-topic-means. Wilcoxon n is the matched topic $\times$ seed cell count available for each benchmark (24 topics $\times$ 3 seeds). **SocialIQA** and **MMLU-STEM** reach BH-adjusted significance on the primary Wilcoxon test, but MMLU-STEM's bootstrap median interval crosses zero and its paired t-test does not reach significance. MMLU-SS does not reach BH-adjusted significance, and ARC-C's negative median indicates slightly less damage from influence-targeted than random in-topic forgetting.

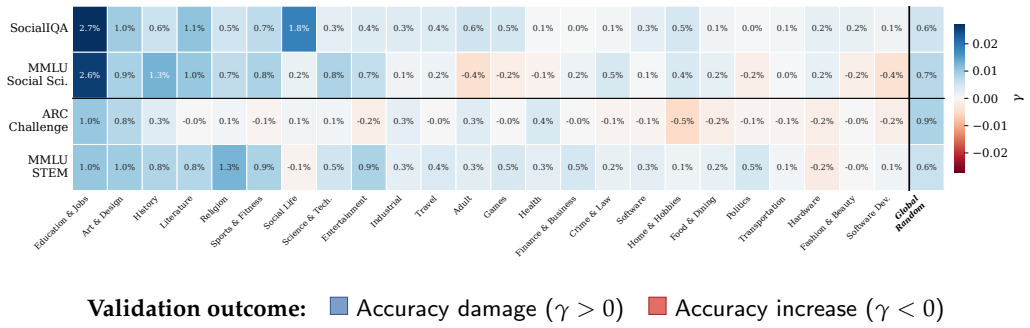


Figure 44: Cross-benchmark γ for single-bin unlearning across all 24 WebOrganizer topics, with $\gamma = A_{\text{baseline}} - A_{\text{unlearned}}$. Positive values indicate accuracy damage; negative values indicate an accuracy increase after unlearning. The vertical divider separates the topic interventions from the global random control; rows are ordered by the manuscript's 2×2 benchmark design.

ARC-Challenge, and +1.0 on MMLU STEM. Religion is the largest MMLU STEM drop (+1.3 points), while Art & Design and History & Geography also show positive damage on multiple benchmarks. At the other end, topics such as Software Development, Transportation, and Fashion & Beauty show smaller mean γ, and Adult Content produces a slight accuracy increase on MMLU Social Sciences in this single-bin diagnostic view.

Figure 45 shows the per-topic γ for each benchmark individually. Each panel keeps the benchmark-role color in its title and border, while muted-brown/gray bar fills mark positive accuracy damage versus accuracy increases relative to the zero line. Topics are sorted independently within each panel. The rank ordering of topic impact is benchmark-dependent: a topic that strongly degrades SocialIQA does not necessarily degrade MMLU STEM to the same degree, consistent with the cross-benchmark divergences documented in Appendix H.

J.3 Global Random Control

To distinguish topic-specific effects from procedural degradation inherent to NGDiff, we run a global random control where both forget and retain sets are sampled uniformly from the full corpus without regard to topic labels. The same hyperparameters, stopping criterion, and evaluation protocol apply.

Figure 46 compares the γ distribution from all 24 single-bin runs (boxplots) against the global random control (star) for each benchmark. The control provides a procedural baseline for interpreting the topic-specific interventions: effects near the control are difficult to separate from ordinary unlearning degradation, while effects far outside the control range indicate stronger topic-specific sensitivity. For SocialIQA, MMLU Social Sciences, and MMLU STEM, the control lies within or close to the upper edge of the single-bin interquartile range, so the largest topic effects should be read relative to that background degradation. For ARC-Challenge, the global random control lies above the single-bin IQR, so the topic-specific ARC effects in this diagnostic sweep are especially difficult to separate from procedural noise.

J.4 Topic-Level Efficiency Analysis

Figure 47 reports convergence speed by topic and benchmark, showing where influence-guided forget sets reach the stopping criterion earlier than random in-topic forget sets. Convergence speed is a different axis from attribution magnitude, so the fastest-to-unlearn topic need not be the highest-influence one. For SocialIQA, Industrial shows the largest step gain here; Industrial also happens to lead the signed topic-level influence marginals (Figure 13), but the two rankings answer different questions and need not coincide in general.

J.5 Influence–Drop Diagnostic

Figure 48 reports the diagnostic relationship between TrackStar influence scores and observed accuracy drop. The scatter is kept in the appendix because the paired unlearning comparison carries the main validation claim, while this view is useful for checking whether higher influence is associated with stronger measured degradation.

In this single-bin diagnostic the influence–drop association is weak throughout: SocialIQA shows a mild positive trend across topics, while the remaining benchmarks show no consistent positive association. We treat this as a diagnostic pattern rather than a separate causal claim; the paired influence-vs-random comparison above carries the validation evidence.

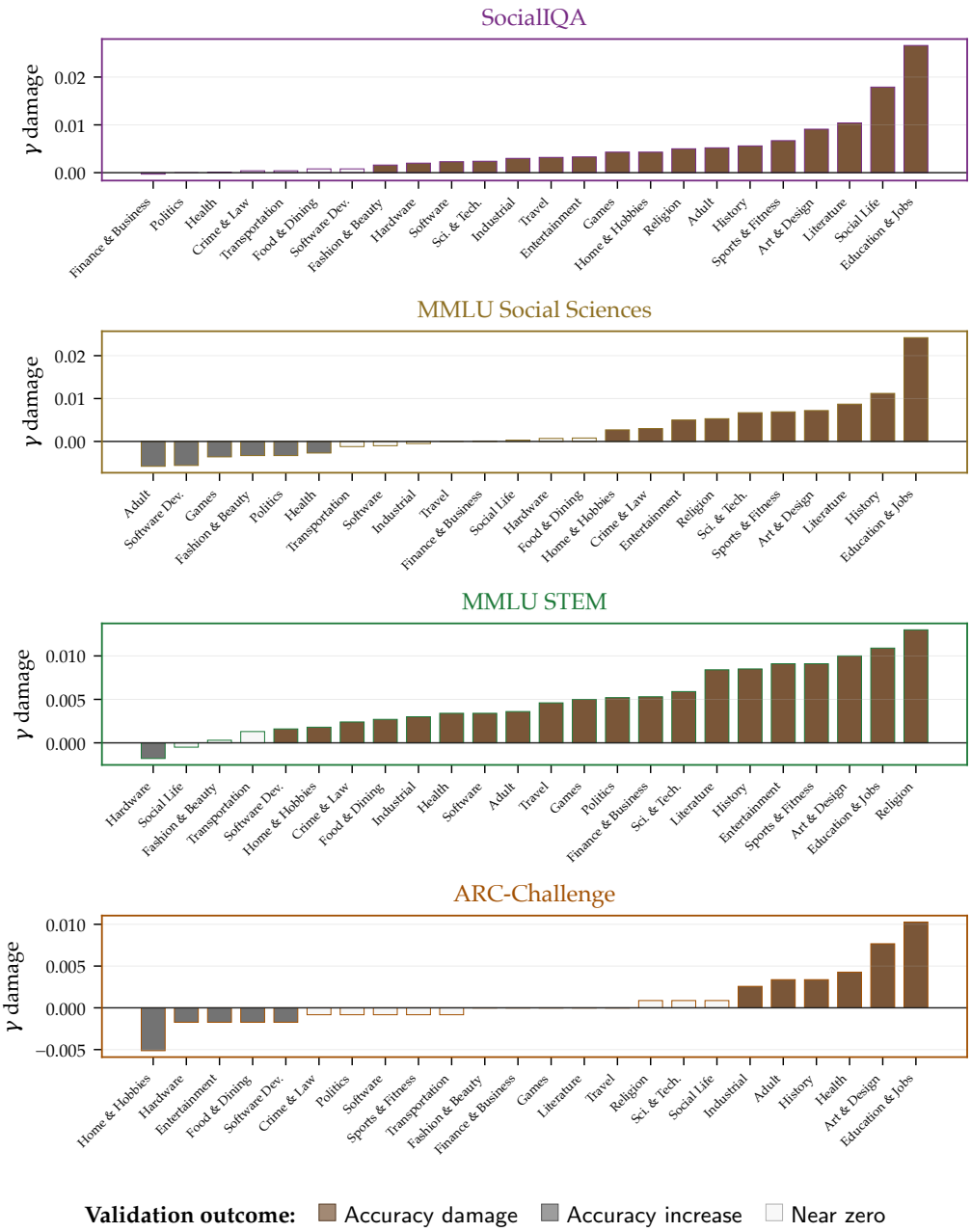


Figure 45: Per-benchmark γ for each of the 24 topic bins, sorted independently within each benchmark panel. Panel titles and borders encode benchmark role; muted-brown/gray bars encode positive accuracy damage versus accuracy increases relative to the zero line.

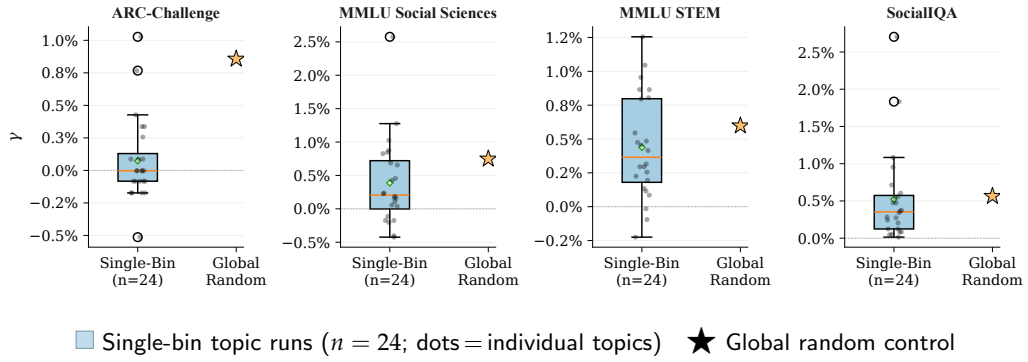


Figure 46: Single-bin γ distributions (boxplots, $n = 24$) versus the global random control (star) for each benchmark. The control samples forget and retain documents uniformly from the full corpus, controlling for procedural degradation.

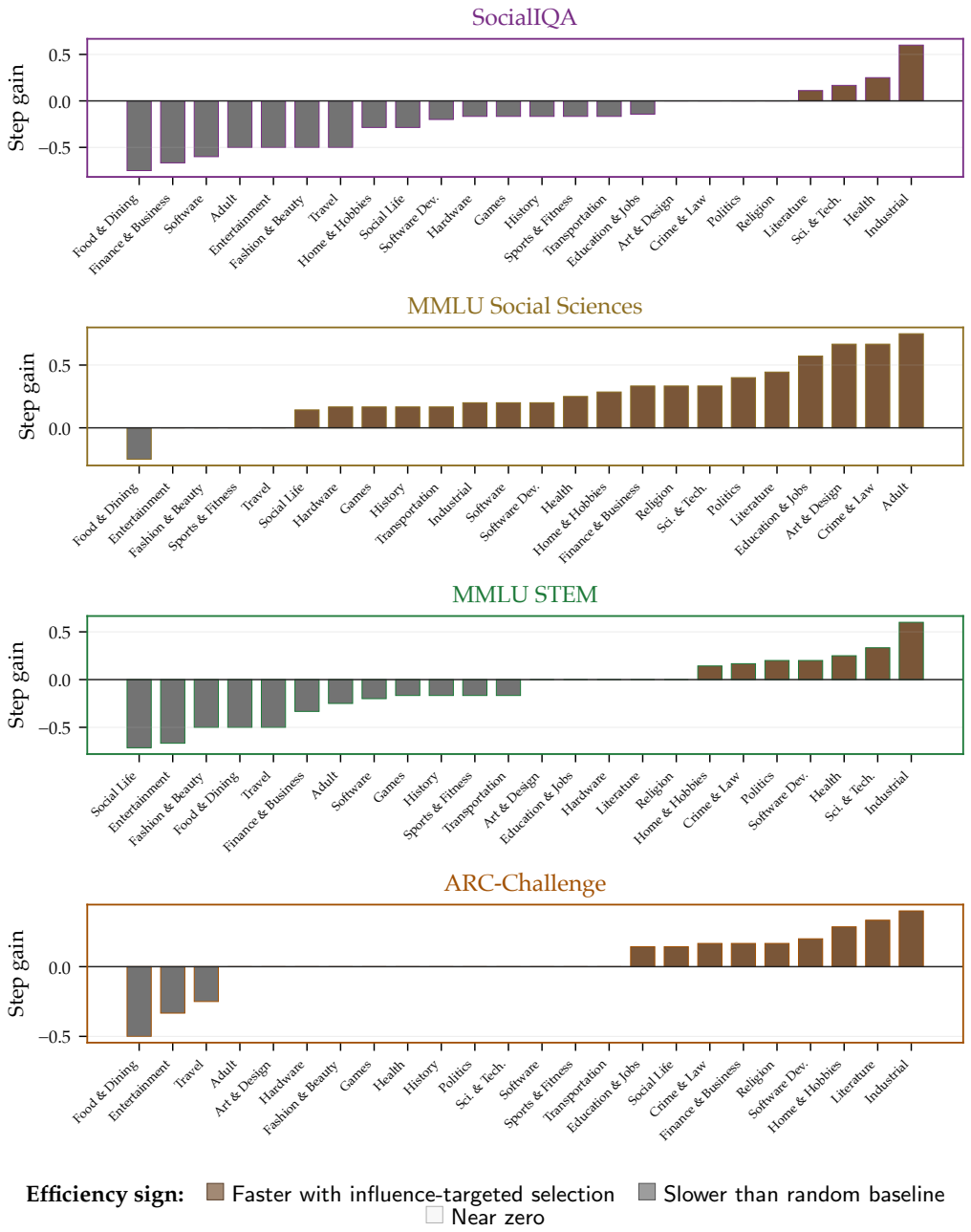


Figure 47: Topic-level unlearning efficiency gain relative to the in-topic random baseline. Positive values indicate faster convergence with targeted selection; negative values indicate additional steps to reach the randomized-text perplexity threshold. Muted-brown bars mark faster influence-targeted selection and gray bars mark slower outcomes.

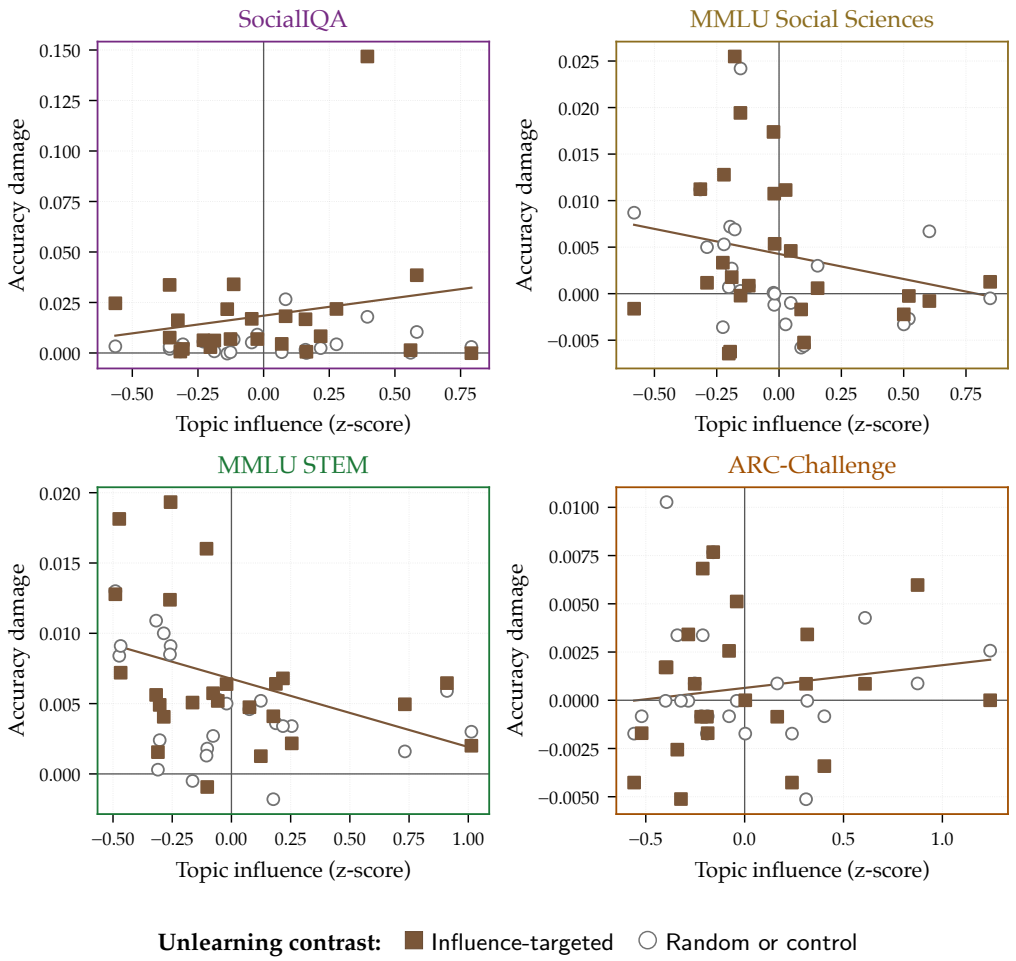


Figure 48: Relationship between TrackStar influence scores and raw accuracy damage across four benchmarks. Each point represents a document set (k=2,000), with $\gamma = A_{\text{baseline}} - A_{\text{unlearned}}$. Hollow gray circles denote baseline degradation from random sampling within the target topic bin, while solid muted-brown squares show the impact of selecting high-influence documents. Solid lines indicate the linear regression for the targeted condition.

K Selective Unlearning: Does Taxonomy Decomposition Help?

A natural question about our method is whether decomposing the corpus into a topic taxonomy actually buys anything, or whether one could get the same control by simply unlearning the most-influential documents regardless of topic. This appendix answers that question with direct unlearning experiments. After correcting the SocialIQA baseline and using one gamma convention throughout, the selectivity evidence is mixed: SocialIQA remains favorable, MMLU STEM is favorable only under the faithful per-document replication, MMLU Social Sciences is an agreeing loss, and ARC-Challenge shows no selectivity advantage because the target damage is at the noise floor.

What "selective" means. We unlearn a set of documents, then measure the accuracy drop $\gamma(b')$ on each benchmark b'. For a recipe intended to degrade target benchmark b, we define selectivity as the damage on the intended target divided by the average collateral damage on the other three primary benchmarks:

$$\mathrm{sel}(b) \;=\; \frac{|\gamma(b)|}{\frac{1}{3}\sum_{b' \neq b} |\gamma(b')|}.$$

A selectivity of 1 means damage is spread evenly; higher values mean the intended capability is degraded while the others are spared.

Reading selectivity results

Values above 1 mean the intended benchmark takes more damage than the average collateral benchmark. The strongest evidence is therefore a large drop on the intended benchmark paired with low collateral damage across the other three cells of the 2×2 benchmark design.

Experimental design. We compare four unlearning recipes that form a 2×2 factorial over the two ingredients of bin-targeted unlearning: *which documents to forget* (the most-influential ones, ranked by training-data attribution, versus a random sample) and *whether to restrict the forget set to one topic* (a single WebOrganizer topic versus the whole corpus).

	one topic	whole corpus
influence-ranked	bin-targeted (our method)	naïve top-K
random	random-in-topic	random-global (baseline)

All four are realized unlearning runs (NGDiff (Jin et al., 2025) with a rank-8 LoRA adapter, k=2,000 forget documents each, identical configuration; only the forget-set selection differs), evaluated under the same OLMES protocol as our primary suite. Unless a table states otherwise, the numbers below are measured raw accuracy damage values or ratios derived from them.

What uncertainty is measured. The $\pm$ values in the faithful per-document selectivity tables are standard deviations across fixed unlearning seeds. They estimate optimizer-run variability, not confidence intervals over alternative forget-document samples. The paired-significance table in Appendix J.1 is the source for bootstrap 95% intervals on the matched influence-vs-random median effect. The robustness rows below are sensitivity checks: they vary analysis choices over the same training runs.

Result 1: the strict blind top-1 factorial is mixed. Table 50 reports the 2×2. Comparing the two influence-ranked recipes isolates the value of *topic filtering*; under the corrected OLMES gamma matrix, this helps on 0 of the 3 scorable benchmarks at the blind top-1 attribution topic. Comparing the two topic-filtered recipes isolates the value of *influence ranking*; this helps on 1 of 3. The strict blind top-1 factorial therefore does not support a uniform selectivity claim. It is still useful diagnostically: the MMLU pair fails at the

Industrial top-1 topic, while ARC-Challenge sits at the noise floor ($\gamma \approx 0$) so its ingredient ratios are uninterpretable and it is excluded from the counts.

Result 2: bin-targeted vs. naïve top-K head-to-head. Table 51 reports the direct head-to-head between bin-targeted unlearning and naïve top-K under the corrected Path-B analysis, using the simplest possible way of choosing which topic to target: pick the single topic with the highest attribution score for the benchmark. Under this attribution-only choice and the config-matched OLMES `:mc::olmes` baselines, bin-targeted unlearning is more selective than naïve top-K on 1 of the 3 scorable benchmarks: SocialIQA. The two losses, MMLU STEM and MMLU Social Sciences, both occur at the Industrial top-1 topic; ARC-Challenge sits at the noise floor ($\gamma \approx 0$) and its head-to-head ratio is not interpretable.

Pilot topic choice changes the picture, but not uniformly. Because the attribution score is only a *predictor* of unlearning effect, not a measurement of it, the simplest topic choice leaves selectivity on the table. We therefore evaluate topic-selection procedures in Table 52. The corrected cross-method results do not recover a uniform four-benchmark win.

- **Corrected bin-level pilot.** The refreshed bin-level single-seed pilot is favorable on SocialIQA (1.43×) but not on ARC-Challenge (0.82×), MMLU STEM (0.62×), or MMLU Social Sciences (0.77×).
- **Faithful per-document pilot.** The faithful replicated pipeline is favorable on SocialIQA ($2.17 \pm 0.61\times$) and MMLU STEM ($1.45 \pm 0.10\times$), loses on MMLU Social Sciences ($0.15 \pm 0.02\times$), and shows no advantage on ARC-Challenge ($0.87 \pm 0.24\times$), whose on-target damage is at the noise floor.
- **Best-in-hindsight (upper bound).** If the unlearning effect of every topic were known, picking the most selective one is favorable on all four faithful-pipeline targets. This is an upper bound, not a deployable selection rule, and should not be used as the main result.

The gap between attribution-only selection and the pilot rows is itself a finding: attribution scores predict *which* documents matter but are imperfect predictors of *how much* unlearning them degrades a capability.

Robustness. The faithful per-document robustness sweep in Table 52 is also mixed. SocialIQA is favorable under the mean, worst-case, and excluding-ARC ratio metrics, but not under the signed ratio. MMLU STEM is favorable in the faithful pilot row but not at the highest-attribution topic across the six metric definitions. MMLU Social Sciences is a consistent loss in the ratio metrics, despite winning under the entropy concentration diagnostic. ARC-Challenge ratios sit at or below 1× across the deployable metrics and are not interpretable because the target damage is at the noise floor. Full numbers and the per-metric sweep are in Table 52.

Limitations. *Method-dependent benchmarks.* The faithful per-document pipeline supports SocialIQA and MMLU STEM but not MMLU Social Sciences. ARC-Challenge remains difficult to interpret because selectivity ratios are sensitive when on-target damage is near zero. We report these cases directly in the cross-method check (§K.1) rather than treating them as artifacts to suppress. *Topic vs. bin granularity.* Our unlearning is implemented per WebOrganizer topic (24 levels); the selectivity gains may be larger at the finer (topic × format) bin granularity that the attribution profiles (§4) operate on, but per-bin unlearning is left to future work. *Target-topic choice.* The current audit shows that topic choice is load-bearing. A purely predictive selection rule that matches the best empirical pilot remains open.

K.1 Cross-Method Check: Bin-Level vs. Per-Document Forget-Sets

The corrected bin-level pilot above is computed under a single forget-set-construction method (bin-level z-scoring at topic × format granularity, single seed). To distinguish methodological advantages of bin-targeting from artifacts of that specific construction, we re-run the pilot under a second, independently-implemented forget-set construction: *faithful per-document* selection, in which the forget set for each (target, candidate topic) pair is the

top-200 documents by per-document influence on the target benchmark, replicated across the available unlearning seeds. Both methods use the identical 24-topic grid, the identical pilot-validation procedure (shortlist 3 candidate topics by unique-z, run a γ pilot on each, keep the most selective), and the identical OLMES `:mc::olmes` baselines and evaluation protocol; only the forget-set construction differs.

Table 53 reports the cross-method comparison. The four targets sort into four distinct patterns.

- **SocialIQA.** Bin-level pilot 1.43× and faithful pilot $2.17 \pm 0.61\times$ agree in direction. This remains the clearest favorable selectivity result.
- **MMLU STEM.** The bin-level pilot loses (0.62×), while the faithful per-document pilot wins ($1.45 \pm 0.10\times$). This is favorable only under the faithful forget-set construction.
- **MMLU Social Sciences.** The bin-level pilot (0.77×) and faithful per-document pilot ($0.15 \pm 0.02\times$) both lose. This is an agreeing loss under the current cross-method comparison.
- **ARC-C.** Neither construction shows a selectivity advantage (bin-level 0.82×, faithful $0.87 \pm 0.24\times$). Because on-target damage is at the noise floor ($\gamma \approx 0$), the ratio is not interpretable; we treat ARC-Challenge as a noise-floor non-result.

Reframed headline. The cross-method-aware reading is: SocialIQA is favorable in both constructions; MMLU STEM is faithful-only favorable; MMLU Social Sciences is an agreeing loss; and ARC-Challenge is a noise-floor non-result (no selectivity advantage under either construction). This is the reading used throughout the paper.

Table 50: Selectivity of four empirical unlearning recipes, forming a 2×2 factorial that separates the two ingredients of bin-targeted unlearning: *influence ranking* (which documents to forget) and *topic filtering* (whether to restrict the forget set to one WebOrganizer topic). Each cell is a realized NGDiff unlearning run; we report selectivity $\mathrm{sel}(b) = |\gamma(b)| / \mathrm{mean}_{b' \neq b} |\gamma(b')|$ (damage on the intended benchmark relative to mean collateral on the other three), evaluated at the top-1 attribution topic for each target benchmark b under the config-matched OLMES `:mc::olmes` protocol. Higher is more selective. The two right-hand columns isolate each ingredient: the *influence effect* compares the two topic-filtered recipes (top-left vs. bottom-left), and the *topic effect* compares the two influence-ranked recipes (top-left vs. top-right).

	Selectivity by recipe			Ingredient effect	
Target b	Bin-targeted (influence + topic)	Naïve top-K (influence, no topic)	Random in-topic	Influence effect	Topic effect
SocialIQA	8.63	**14.37**	11.58	0.75×	0.60×
ARC-Challenge	1.11	1.34	0.03	n/a ($\gamma \approx 0$)	n/a ($\gamma \approx 0$)
MMLU STEM	0.04	**0.78**	0.10	0.38×	0.05×
MMLU Social Sciences	0.06	**0.88**	0.04	**1.33**×	0.06×
count helping	—	—	—	1/3	0/3

Influence effect = selectivity(bin-targeted) / selectivity(random-in-topic): isolates the value of ranking documents by influence, holding the topic fixed. **Topic effect** = selectivity(bin-targeted) / selectivity(naïve top-K): isolates the value of restricting to one topic, holding influence-ranking fixed. Under the corrected OLMES gamma matrix, influence ranking helps on 1 of the 3 scorable benchmarks and topic filtering on 0 at the strict blind top-1 topic. ARC-Challenge is excluded from the ingredient ratios because its on-target damage is at the noise floor ($\gamma \approx 0.003$), which makes the near-zero/near-zero ratios uninterpretable. Both MMLU Social Sciences and MMLU STEM topic effects are below 1.0 because the blind top-1 attribution topic (Industrial) is a weak unlearning target for both. The pilot and faithful per-document checks in Table 52 should therefore be read as a mixed robustness audit, not as a uniform four-benchmark win.

Table 51: Selectivity of bin-targeted vs. naïve top-K unlearning under the corrected **Path B** analysis (k=2,000 documents per condition). For target benchmark b, the damage ratio $\mathrm{sel}_c(b)=|\gamma_{c,b}(b)|/\mathrm{mean}_{b'\neq b}|\gamma_{c,b}(b')|$ measures how concentrated the unlearning damage is on the intended target relative to the mean magnitude of collateral damage on the other three primary benchmarks (higher = more selective). Bin-targeted uses the single-topic intervention at each benchmark's top-1 topic $k_\star$ (parent topic of the max-z bin in z-scored bin influence). Naïve top-K is the Path-B linear prediction $\gamma_{\mathrm{naive},b}(b') = \sum_k p_b[k]\gamma_{b,k}(b')$, where $p_b[k]$ is the empirical topic distribution of the global top-2,000 documents ranked by influence on b. All γ values are raw accuracy damage under the config-matched OLMES `:mc::olmes` protocol.

Target b	Top-1 topic–format bin	$\mathrm{sel}_{\mathrm{bin}}$ bin-targeted	$\mathrm{sel}^{\mathrm{emp}}_{\mathrm{naïve}}$ naïve top-K	Ratio
SocialIQA	Literature × Customer Support	**1.28**	0.64	**1.99**×
ARC-Challenge	Industrial × Documentation	1.11	0.36	n/a ($\gamma\approx0$)
MMLU STEM	Industrial × Documentation	0.10	0.28	0.36×
MMLU Social Sciences	Industrial × Documentation	0.06	0.18	0.34×

Top two rows: bin-targeting concentrates damage more than naïve top-K on one of the three scorable benchmarks (SocialIQA) under the corrected Path-B top-1 rule; ARC-Challenge sits at the noise floor ($\gamma\approx0$) and its ratio is not interpretable. Bottom two rows: MMLU STEM and MMLU Social Sciences lose under this attribution-only choice because their highest-attribution bin (Industrial × Documentation) is a weak unlearning target for both. The pilot and cross-method checks in Tables 52 and 53 give the current mixed reading rather than a uniform four-benchmark win. Sign convention: $\gamma > 0$ indicates accuracy drop on the named benchmark after unlearning; absolute values are taken for the selectivity ratio so help (negative γ) and hurt (positive γ) on non-target benchmarks both count as collateral.

Table 52: Robustness of the bin-targeted vs. naïve top-K head-to-head after the baseline and gamma-convention audit. **Top:** the faithful per-document pipeline reports the selectivity advantage at the highest-attribution topic under six metric definitions (mean ± standard deviation across replicated seeds where ratios are defined; these are not formal confidence intervals). **Bottom:** the pilot rows compare the corrected bin-level single-seed and faithful per-document pipelines. All ratio entries are selectivity(bin-targeted) / selectivity(naïve top-K); ↑ is better.

Selectivity metric (target topic chosen by highest attribution)	**SocialIQA**	**MMLU Social Sciences**	**MMLU STEM**	**ARC-Challenge**	Wins
Damage ratio, mean collateral (headline)	**1.33±0.44**×	0.36±0.35×	0.25±0.23×	0.81±0.40×	1/4
Damage ratio, worst-case collateral	**1.51±0.55**×	0.34±0.34×	0.17±0.15×	0.60±0.36×	1/4
Damage ratio, excluding ARC-Challenge	**1.18±0.42**×	0.33±0.32×	0.23±0.20×	0.81±0.40×	1/4
Signed damage ratio	0.82±0.35×	0.36±0.35×	0.25±0.22×	0.81±0.40×	0/4
Cosine similarity to ideal target-only profile	**win (2/3)**	loss (0/3)	loss (0/2)	loss (1/3)	1/4
Concentration (negative entropy) of damage	**win (2/3)**	**win (3/3)**	**win (2/2)**	**win (3/3)**	**all**
Pilot topic-selection checks (mean-collateral metric, corrected PACE outputs)					
Corrected bin-level single-seed pilot					
Bin-level pilot selection	**1.43**×	0.77×	0.62×	0.82×	1/4
Faithful per-document pilot					
Faithful pilot selection	**2.17±0.61**×	0.15±0.02×	**1.45±0.10**×	0.87±0.24×	**2/4**
Faithful best-in-hindsight upper bound					
Best-in-hindsight selection	**5.11±2.61**×	**1.88±0.68**×	**1.45±0.10**×	**4.85±0.63**×	**all**

ARC-Challenge shows no selectivity advantage in the deployable pilot rows (≈ 0.6–0.9×); its on-target damage is at the noise floor ($\gamma \approx 0$), so the ratio is not interpretable as a selectivity result. The current interpretation is SocialIQA favorable, MMLU-STEM favorable only under faithful per-document forget-sets, MMLU Social Sciences an agreeing loss, and ARC-Challenge a noise-floor non-result. Best-in-hindsight is an upper bound, not a deployable selection rule.

Table 53: Cross-method comparison of pilot-validated bin-targeted vs. naïve top-K selectivity after the baseline and gamma-convention audit. Two independent forget-set constructions are evaluated on the same 24-topic grid under the same OLMES `:mc::olmes` protocol: *bin-level* (topic × format z-scored, single seed) and *faithful per-document* (top-200 by per-document influence on target, seeds 42/43/44 where available). Both apply the same pilot-validated topic-selection procedure (3-candidate shortlist by unique-z, retain best). We report the pilot selectivity ratio (bin-targeted / naïve top-K, ↑ better) under each method and the cross-method verdict.

Target b	**Pilot selectivity ratio** (↑)		Method agreement	Verdict
	Bin-level (single seed)	Faithful 3-seed (mean ± std)		
SocialIQA	**1.43**×	**2.17 ± 0.61**×	agree	Favorable
MMLU-STEM	0.62×	**1.45 ± 0.10**×	faithful-only	Favorable only under faithful forget-sets
MMLU-SS	0.77×	0.15 ± 0.02×	agree loss	Not favorable
ARC-C	0.82×	0.87 ± 0.24×	agree (no advantage)	Small-γ, noise floor

The bin-level pilot uses ≈ 2,000-document forget sets, while the faithful per-document pilot uses top-200 forget sets with matched random controls. Both unlearn with NGDiff + rank-8 LoRA on OLMo3-7B Base; baselines come from the same Base checkpoint under `:mc::olmes` (config-matched). Faithful pilot numbers aggregate the available replicated seeds. The corrected reading is mixed: SocialIQA is favorable in both constructions, MMLU-STEM is favorable only under faithful per-document forget-sets, MMLU Social Sciences is an agreeing loss, and ARC-Challenge shows no selectivity advantage under either construction (0.82× and 0.87 ± 0.24×): its on-target damage is at the noise floor ($\gamma \approx 0$), so the ratio is not interpretable as a selectivity result.

K.2 ARC Unlearning Across Task Definitions (Coverage Robustness)

The ARC attribution and its bin-level forget-sets are computed over the full 5,678,621-document working set, matching the other primary benchmarks. To check that the near-zero ARC on-target effect is a property of the benchmark rather than an artifact of the earlier partial-coverage attribution, we re-run bin-level unlearning under three ARC definitions: ARC-Easy, ARC-Challenge, and ARC-Combined (the union of the two query sets). Each definition uses the same NGDiff + rank-8 LoRA pipeline, the same 24-topic grid plus a corpus-wide recipe, and the same config-matched OLMES `:mc::olmes` baselines.

Table 54 and Figure 49 report the result. On-target γ stays below one accuracy point under all three definitions (ARC-Easy $+0.0083 \pm 0.0068$, ARC-Challenge $+0.0029 \pm 0.0056$, ARC-Combined $+0.0059 \pm 0.0049$), with ARC-Challenge itself within one standard deviation of zero, while the held-out benchmarks are preserved. The noise-floor reading of ARC-Challenge therefore holds across task definitions and under the corrected full-coverage attribution; it is not produced by the earlier partial scoring.

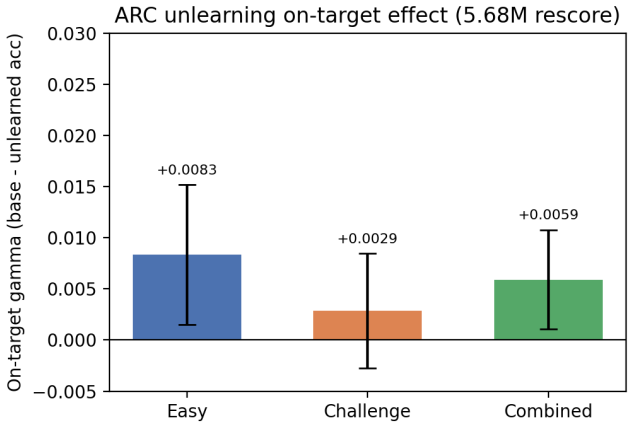


Figure 49: On-target γ (base $-$ unlearned accuracy) for ARC unlearning under three task definitions on the full 5.68M working set, mean $\pm$ std over 25 recipes. All three sit at the noise floor.

Table 54: ARC unlearning on-target effect across three task definitions on the full 5,678,621-document working set. Bin-level NGDiff + rank-8 LoRA unlearning is run per WebOrganizer topic (24 topics + corpus-wide) for each ARC definition: ARC-Easy, ARC-Challenge, and ARC-Combined (union of the two query sets). On-target $\gamma = (\text{base acc} - \text{unlearned acc})$, mean $\pm$ std over the 25 recipes; positive γ means accuracy dropped after unlearning. Retain columns report mean γ on the held-out benchmarks. All three definitions leave the on-target benchmark essentially unchanged, confirming the noise-floor reading is not an artifact of the earlier partial-coverage attribution.

ARC definition	n	On-target γ	Base acc	SocialIQA γ	MMLU-STEM γ	MMLU-SS γ
ARC-Easy	25	$+0.0083 \pm 0.0068$	0.942	$+0.0206$	$+0.0032$	$+0.0026$
ARC-Challenge	25	$+0.0029 \pm 0.0056$	0.834	$+0.0184$	$+0.0034$	$+0.0008$
ARC-Combined	25	$+0.0059 \pm 0.0049$	0.888	$+0.0189$	$+0.0035$	$+0.0022$

On-target γ is the mean over 24 per-topic recipes plus the corpus-wide recipe; for ARC-Combined the on-target accuracy is the mean of ARC-Easy and ARC-Challenge. Baselines are the same-pipeline OLMo3-7B Base checkpoint under `:mc::olmes`. Across all three definitions the on-target effect stays below one accuracy point; ARC-Challenge itself is within one standard deviation of zero.

L Compute and Resource Accounting

L.1 Taxonomy Enrichment

Classifying 1.26B documents with the official WebOrganizer classifiers (Wettig et al., 2025) required processing 65,718 shards across a mixed-tier HPC GPU cluster. Per-shard task wall-clock and allocated GPU type were instrumented at run time; aggregated totals appear in Table 56.

Classifier accuracy. On the validation set used by Wettig et al. (2025) (10k web pages, $\geq$75% teacher confidence), the topic and format classifiers achieve 93.5% and 91.8% average accuracy respectively (87.1% and 80.5% worst-group). The 405B teacher's own self-agreement under random prompt ordering is 97–98%, placing classifier-induced error at $\sim$4–5% above the annotation-noise floor. Per-classifier throughput by GPU tier is itemized in Table 55.

Table 55: Enrichment throughput by GPU type.

GPU	VRAM	Throughput	Time per shard
H200	144 GB	$\sim$37 docs/sec	$\sim$11 min
H100	80 GB	$\sim$32 docs/sec	$\sim$13 min
RTX PRO 6000	96 GB	$\sim$28 docs/sec	$\sim$15 min
L40S	48 GB	$\sim$25 docs/sec	$\sim$17 min
RTX 6000	24 GB	$\sim$14 docs/sec	$\sim$30 min
V100	16 GB	$\sim$13 docs/sec	$\sim$32 min

L.2 Attribution Pipeline

We use Bergson's index-reuse pipeline: training gradients are computed once on GPU and stored as a persistent index, then scored against any number of query sets via *CPU* dot products. The one-time index build amortizes over all subsequent benchmark scoring passes. The attribution pipeline thus decomposes into four distinct GPU-bound sub-phases plus a separate CPU-bound scoring stage:

2a. Training gradient index (20,000 GPU-hr): per-document gradients for the 5,678,621-document stratified working set ($\sim$316 shards $\times$ 17,996 docs), built once on H100 80 GB and H200 144 GB across two HPC allocations.

2b. Query gradient indexes (20 GPU-hr): model-specific query gradient indexes, in base and instruct variants, for the four primary benchmarks (SocialIQA, MMLU Social Sciences, ARC-Challenge, MMLU STEM) and the additional diagnostic probe suites, including the instruct-query builds used for reported two-model scoring.

2c. Preconditioner construction (192 GPU-hr): TrackStar mixed preconditioner artifacts for the reported scoring runs, with base-side curvature as the pretraining-data metric and supervised-instruction-tuning curvature treated as identity. This identity treatment is a modeling approximation that keeps the reported metric anchored to pretraining-data curvature rather than adding an unobserved post-training curvature term. Each build uses 8$\times$ H100 80 GB with FSDP over a 100,000-document sample.

2d. Calibration and exploratory runs (1,500 GPU-hr): smoke tests, calibration baselines (156.1 GPU-hr without preconditioning), and preconditioner-hyperparameter sweeps that informed the final pipeline.

2e. Production scoring (CPU): each benchmark-vs-index scoring pass runs as a parallel CPU array, so no GPU-hours apply to this stage.

Training index storage. The gradient index requires $\sim$8 KB per document. For the working set (5,678,621 documents), this is $\sim$44 GB.

L.3 Unlearning Experiments

Each single-bin unlearning run trains for up to 5,000 steps on a single **H200 144 GB** GPU. The 24-topic single-bin sweep (24 topics $\times$ 3 chained jobs $=$ 72 runs) is the largest intervention cost; influence-guided and random-baseline conditions roughly double it. Additional cost comes from multi-bin runs, hyperparameter sweeps, null-bin controls, and per-checkpoint evaluation passes (every 250 steps across 4 benchmarks). A subset of single-topic runs was executed on rented GPU infrastructure off the primary HPC environment; those hours are included in the totals above.

L.4 Aggregate Compute Summary

Total compute across all phases was approximately 40,000 raw GPU-hours (37,000 H200 144 GB-equivalent after FLOPS normalization). Table 56 reports the per-phase breakdown. The raw column sums heterogeneous accelerator time; the H200-equivalent column applies bf16 FLOPS ratios per phase to support like-for-like comparison with other corpus-scale attribution works. We round *up* on partial hours to avoid under-reporting. The totals above account for GPU-bound work only; substantial CPU compute for corpus preparation, de-duplication, and classifier-inference orchestration was a non-trivial share of the overall research effort but is tracked separately and not folded into the GPU-hour totals.

Table 56: Approximate GPU-hours across all pipeline phases, itemized by phase with raw (mixed-tier) and H200 144 GB-equivalent (FLOPS-normalized) totals. The H200-equivalent column converts via bf16 FLOPS ratios (H100/H200 $=$ 1.00; older accelerators scale roughly from 0.13 to 0.76). Per-phase values are exact; the aggregate totals are rounded up under the no-underestimate accounting policy and therefore exceed the column sums by a few percent.

Phase	Workload	GPU types	Raw GPU-hr	H200 144 GB-equiv
1. Enrichment	WebOrganizer classifiers	H200/H100/RTX PRO 6000/L40S/V100/RTX 6000	10,000	8,000
2a. Training index	Gradient build (one-time index reuse)	H100 80 GB / H200 144 GB	20,000	20,000
2b. Query indexes	6 benchmarks (base + instruct)	H100 80 GB	20	20
2c. Preconditioner	2$\times$ FSDP build (base + instruct)	8$\times$ H100 80 GB	192	192
2d. Calibration / smoke	Exploratory + smoke + calibration runs	H100 80 GB	1,500	1,500
3. Unlearning	NGDiff sweep + eval passes	H200 144 GB	6,000	6,000
4. Evaluation	OLMES + per-checkpoint evals	H200/H100, L40S/A40 fallback	600	500
Total			40,000	37,000

M LLM Disclosure

Large language models were used during the preparation of this manuscript for drafting and revising prose, editing LaTeX, writing analysis and visualization code, and managing project coordination. All LLM-generated content was reviewed and verified by the authors. LLMs were not used to generate data. Research questions, experimental design, and original ideas are entirely the work of the authors.