

Can Large Language Models Match Human Diversity in Educational Content Generation?

Mei Tan*, Hyunji Nam*, Hari Subramonyam & Dorottya Demzky

Stanford University

Stanford, CA 94305, USA

{mxtan, hynam, harihars, ddemszky}@stanford.edu

Abstract

Large language models (LLMs) are known to lack diversity in open-ended text generation, a limitation that is especially consequential in education, where homogeneity can narrow the kinds of thinking and disciplinary work students are invited to do. We study diversity in two educational generation tasks: math items and essay feedback. We compare frontier and open-weight models with human-authored reference distributions using lexical, semantic, and domain-grounded measures. For math, we examine the mathematical activities students are asked to perform; for feedback, we examine the revision work requested and the rhetorical strategies used to communicate it. We evaluate direct prompting, two diversity prompting methods, and post-training. Across both tasks, directly prompted model outputs are markedly less diverse than human-authored content—78% higher on lexical (Self-BLEU) and 51% higher on semantic (embedding) similarity—and represent a narrower range of instructional forms. Diversity prompting and post-training increase variation largely through surface features without broadening the instructional repertoire. Models inherit and amplify the most normative routines of human instruction. Overcoming mode collapse in education will require richer instructional data and investment in teachers’ creative capacity and professional judgment.

1 Introduction

Educational materials shape the kinds of thinking students are invited to do (Marton & Booth, 1997; Marton, 2014; Kullberg et al., 2017). A math problem can ask students to compute a value, construct a representation, evaluate a claim, or analyze a strategy (Stein & Smith, 1998; Swan, 2005; Burkhardt & Swan, 2013). A feedback comment can ask a writer to correct an error, elaborate an idea, or evaluate the effectiveness of a rhetorical choice (Bereiter & Scardamalia, 1987; Straub, 1997; Carless & Boud, 2018). Diversity in educational content captures the breadth of learning opportunities and disciplinary work that materials make available.

As large language models are increasingly used to generate educational materials, including tutoring responses, lesson plans, assessments, and feedback (Kasneci et al., 2023; Wen et al., 2024; Team et al., 2025; Khan, 2023), recent work has raised concerns about their performance in these instructional roles. Studies have documented outputs that are inaccurate, poorly aligned with curricular standards, inappropriate for target learners, or reproduce harmful stereotypes (Steiss et al., 2024; Bulut et al., 2024; Scaria et al., 2024; Lucy et al., 2024; Tan et al., 2026). Yet even when generated content meets basic standards of correctness and relevance, homogenization can limit its instructional value. A system that draws from a narrow repertoire of pedagogical moves may appear competent while repeatedly offering the same kinds of tasks, feedback, and invitations for student thinking.

This concern connects to a broader literature on LLM output homogeneity and mode collapse. Across domains, models have been shown to concentrate on a narrow subset

*Co-first authors

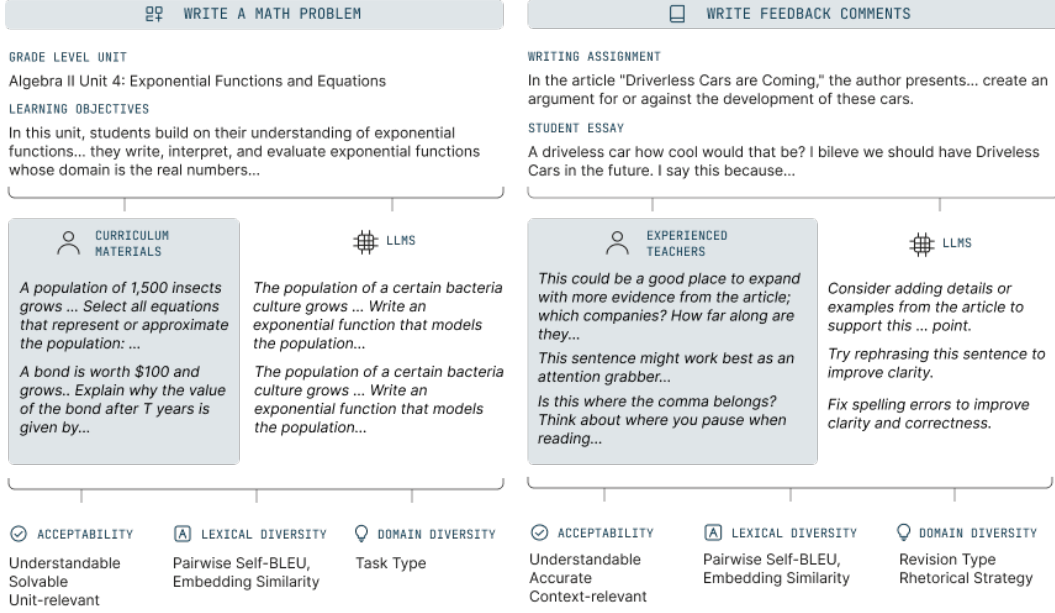


Figure 1: We study diversity in two educational generation tasks: math item generation and essay feedback.

of the human response distribution (Jiang et al., 2025; Wu et al., 2024; Shumailov et al., 2024; Xu et al., 2025; Namuduri et al., 2025). This collapse may reflect a typicality bias in post-training (Li et al., 2025; Liu, 2026; Kirk et al., 2024) as well as the modal nature of human demonstrations, shaped by repetition of familiar and institutional routines (Lortie, 1975; Bourdieu, 1977).

We study this problem in two popular educational generation tasks: **(1) math items** and **(2) essay feedback**. We compare outputs from frontier and open-weight models to human-authored reference distributions, examine the effects of diversity prompting and post-training, and evaluate diversity using lexical, semantic, and domain-grounded measures. We assess the acceptability of outputs, how variation occurs across surface and consequential features, and how generated language constructs instructional priorities and invitations for student thinking.

Across tasks, models fail to recover the breadth of instructional forms present in human-authored content. Diversity prompting narrows this gap, but much of the improvement comes from expanding vocabulary and varying easily changed surface features, while outputs remain concentrated in a small set of dominant instructional forms. Models readily reproduce forms that are already common in human-authored materials, but struggle to recover high quality instructional practices that fall outside these familiar patterns.

Our contributions are:

- We frame diversity in educational generation in terms of the breadth of learning opportunities and disciplinary work made available to students.
- We study diversity in two educational generation tasks using human expert demonstrations as reference distributions.
- We combine lexical, semantic, and domain-grounded measures to distinguish surface variation from pedagogically meaningful variation.
- We show that models fail to recover the breadth observed in human-authored reference distributions, which are themselves constrained.
- We show that diversity prompting and post-training yield inconsistent gains, raising diversity scores without broadening acceptable pedagogical coverage.

2 Related Works

LLM output homogeneity. A growing body of work has documented homogeneity in LLM-generated text (Jiang et al., 2025; Wu et al., 2024; Shumailov et al., 2024). Jiang et al. (2025) show that 70+ LLMs exhibit high pairwise embedding similarity within and across model families, Wu et al. (2024) find sentiment collapse in book reviews, and Xu et al. (2025) find repetitive plot elements in stories. Similar patterns appear in narrative discourse (Namuduri et al., 2025), scientific ideation (Ruan et al., 2026), and code generation (Lu et al., 2025).

Post-training methods, including supervised fine-tuning (SFT), reinforcement learning from human feedback (RLHF), and direct preference optimization (DPO), have been shown to reduce output diversity (Li et al., 2025; Liu, 2026). Kirk et al. (2024) find that RLHF reduces diversity, and Liu (2026) report a similar effect for DPO. Recent work has modified post-training to reward more varied responses. DivPO selects rare responses as chosen examples and common responses as rejected examples (Lanchantin et al., 2025), reporting improved diversity in persona generation and creative writing. Other work proposes inference-time methods (Khan & wester, 2026). Verbalized sampling asks models to generate a distribution of possible responses along with their probabilities (Zhang et al., 2026a) and negatively correlated sampling prompts models to generate candidate responses that differ from one another (Zhang et al., 2026b).

Measuring LLM output diversity. Self-BLEU (Zhu et al., 2018) measures n-gram overlap among generated texts. Distinct- n (Li et al., 2016) captures a related notion of lexical diversity and is often highly correlated with Self-BLEU (Shaib et al., 2026). Because these metrics rely on word overlap, rephrasing can increase measured diversity without changing the underlying content. Embedding-based approaches, including BERTScore (Zhang et al., 2020), the Vendi Score (Friedman & Dieng, 2023), and MAUVE (Pillutla et al., 2023), instead use learned representations to compare texts or output distributions. Still, neither lexical nor embedding-based metrics consistently capture whether outputs differ in substance (Tevet & Berant, 2021). This limitation has motivated task-dependent approaches to diversity evaluation. For example, Jain et al. (2026) define “functional diversity” in terms of whether users perceive responses as meaningfully different for a particular use.

LLMs for education. Research on LLMs in education has explored their use for tutoring, lesson planning, assessment, and feedback, highlighting the potential for more immediate and scalable instructional support (Kasneci et al., 2023; Wen et al., 2024; Team et al., 2025; Khan, 2023). Evaluations of generated educational content have identified problems with accuracy, curricular alignment, prioritization of issues, and bias (Bulut et al., 2024; Scaria et al., 2024; Lucy et al., 2024; Weissburg et al., 2025; Du et al., 2026; Steiss et al., 2024; Tan et al., 2026). Across educational tasks, studies report generated content that is generic, underspecified, or weakly tailored to the instructional context (Lammert et al., 2024; Hu et al., 2025a; Liu et al., 2025; Yoon et al., 2023). We examine this generic-ness as a form of output homogeneity, in which models repeatedly draw on a narrow range of instructional moves.

3 Data

3.1 Human Demonstrations

Math items. We use a subset of the public MathFish (Lucy et al., 2024) dataset, consisting of standalone, end-of-lesson practice problems from Illustrative Mathematics, a high-quality, open-source K–12 curriculum. We removed duplicate items, excluded problems that depended on images or tables, and converted mathematical expressions from MathJax to ASCII notation for readability. We selected the 20 units with the largest number of eligible items, spanning Grade 6 through High School Algebra II. The resulting dataset contains 1,477 unique math problems.

Essay feedback. We use an expanded version of a dataset introduced in prior work (Tan et al., 2024). The dataset comprises 3,021 inline excerpt-comment feedback written by 61 ex-

```

=== Math Items ===
You are a K-12 math curriculum specialist. Write one practice problem aligned to the given mathematical learning purpose. The problem should be situated within this lesson and contain only one primary mathematical activity. It may address a single aspect of the learning purpose rather than covering the entire lesson. Do not create multiple subproblems or a sequence of distinct tasks that build on one another.
+ [Unit Learning Objective Summary]

=== Essay Feedback ===
You are a middle school English Language Arts teacher. Write inline feedback comments for the given writing assignment prompt and student essay to help your student revise. Each comment must include an exact excerpt from the student's essay and a feedback comment addressing a single topic of revision related to the excerpt.
+ [Writing Assignment Instructions]
+ [Student Essay Text]

```

Figure 2: Direct-prompting system prompts for the math and feedback generation tasks. Prompts abridged for readability. We additionally specified the required output format: math generations contained only the problem text, and feedback generations were returned as structured JSON.

perienced U.S. middle and high school ELA teachers. Teachers had an average of 8.25 years of teaching experience and were recruited through professional development programs focused on writing instruction. The comments address 448 student essays responding to 18 writing assignments drawn from the publicly available PERSUADE ([The Learning Agency Lab, 2024](#)) and ASAP-AES ([Hamner et al., 2012](#)) datasets.

3.2 LLM Demonstrations

We evaluate frontier and open-weight models across multiple scales: GPT-5.2, Gemini-3.5-Flash, Llama-3.3-70B Versatile, Qwen-3.6-27B, Llama-3.1-8B-Instruct, and Qwen-2.5-7B-Instruct. We generate LLM outputs under three settings: direct prompting, prompting methods for increasing diversity, and post-training. We use temperature $\tau = 1$ with model-default top- p . We used OpenAI’s inference API for GPT-5.2, Gemini’s API for Gemini-3.5-Flash, and Groq for serving Llama-3.3-70B Versatile and Qwen 3.6-27B.

Direct prompting. We use direct prompting as a baseline generation condition. In each task, the models receive contextual information associated with the corresponding human demonstrations. For essay feedback, models are given a student essay and its associated writing assignment and asked to generate inline feedback comments. For math items, models are given the Illustrative Mathematics description for the target unit, which summarizes the relevant learning objectives, and asked to generate practice problems.

Diversity prompting. We compare direct prompting with two prompting methods designed to reduce output homogeneity. **Verbalized sampling** asks the model to generate N possible responses together with an estimated probability for each response ([Zhang et al., 2026a](#)). **Negatively correlated sampling** asks the model to generate N responses jointly, with each response differing from the others in the set ([Zhang et al., 2026b](#)). Figures 2 and 3 show the prompts used for each condition.

Post-training. We evaluate supervised fine-tuning (SFT) and direct preference optimization (DPO) ([Rafailov et al., 2024](#)) on open-weight models. We use all math items from 10 of the 20 math units and all feedback comments from 9 of the 18 writing assignments for training and holding out the remaining units and assignments for evaluation. For SFT, each training example pairs the task context with a human-authored response. For DPO, we pair a human-authored chosen response with a model-generated rejected response for the same context. We train Qwen-2.5-7B-Instruct and Llama-3.1-8B-Instruct using OpenRLHF ([Hu et al., 2025b](#)). Appendix D reports the training details and hyperparameters.

```

=== Verbalized Sampling ===
Generate 20 possible responses to the input prompt. Each response must include the response string and the
estimated probability 0.0 to 1.0 of this response given the input prompt (relative to the full distribution).
Randomly sample the responses from the full distribution.
+ [Direct Prompting User Prompt]

=== Negatively Correlated Sampling ===
Generate 20 possible responses to the input prompt that represent diverse instructional values.
+ [Direct Prompting User Prompt]

```

Figure 3: Sampling-variant prompts. Both wrap the direct-prompting user prompt of Figure 2.

4 Methods

4.1 Masking

Diversity metrics can be inflated by superficial differences, such as changes to variable names in math problems or quotations from student writing in feedback comments. We therefore create a masked version of each output and compute diversity on both the original and masked text. We use GPT-5-nano to identify maskable spans, then audit likely errors through programmatic checks and targeted human review. This procedure identified very few masking errors. Appendix B describes the masking prompts, audit procedure, and corrections.

Math items. We mask names and labels used in the problem, such as for people, places, points, lines, and shapes, with [text variable]. We mask equation variables, such as x , y , and α , with [equation variable]. We also mask each numeric value with a label indicating its type: [whole number], [decimal number], or [fraction].

Essay feedback. We mask direct quotes from the student’s essay with [quote] and proposed replacement wording for text in the excerpt with [rewrite]. We also mask new text suggested for the student to add to their work as [insertion]. Although we cannot fully remove essay-specific language, since feedback is inherently reactive to student writing, masking accounts for exact-wording engagement with the essay text. Table 1 shows examples.

Domain	Original Output	Masked Output
Math	Jada has a sleeping bag that is rated for 30 degrees F. Write an inequality that represents the outdoor temperature at which Jada will be able to stay warm. . .	[text_variable] has a sleeping bag that is rated for [whole_number] degrees F. This means that if the temperature outside is at least [whole_number] degrees F, [text_variable] will. . .
Essay	Check spelling: ‘assitance’ should be ‘assistance.’	Check spelling: [quote] should be [rewrite]
Essay	Clarify this idea by saying something like, ‘One good thing about driverless cars is that they could help elderly people.’	Clarify this idea by saying something like, [insertion]

Table 1: Examples of unmasked and masked outputs for math items and essay feedback.

4.2 Acceptability Checks

Because low-quality outputs may still appear diverse, we restrict our diversity analyses to outputs that meet a baseline standard of quality. We apply the same criteria to human-authored and model-generated responses.

We evaluate acceptability using a panel of three LLM judges: GPT-5-nano, Claude-Haiku-4.5, and Gemini-2.5-Flash-Lite. We adjudicate cases without unanimous agreement using task-specific review procedures. Appendix C describes these procedures, common sources of disagreement, and the number of outputs removed.

Across both tasks, outputs must be *understandable*: written in clear language that is complete enough to interpret in context.

Math items. We also evaluate whether each problem is *solvable* and *unit-relevant*. A problem is solvable if it provides enough information to determine a valid solution. A problem is unit-relevant if it addresses mathematical content or learning objectives described for the target unit.

Essay feedback. We also evaluate whether each comment is *accurate* and *context-relevant*. A comment is accurate if its assessment of the student’s writing is true and does not invent or misdiagnose an issue. A comment is context-relevant if it offers a meaningful observation or suggestion to support the student’s writing or development as a writer.

5 Evaluation Metrics

5.1 Self-BLEU

We measure lexical similarity using mean pairwise BLEU, or Self-BLEU (Zhu et al., 2018). For each instructional unit, we draw 20 random samples of model outputs, each matched to the number of human examples, and average Self-BLEU across samples. We then average across units or assignments. We report results for both original and masked text.

5.2 Embedding Similarity

We measure semantic similarity as the mean pairwise cosine similarity among output embeddings (Jiang et al., 2025). We embed each output using text-embedding-3-small and compute similarity within each instructional unit. As with Self-BLEU, we draw 20 model output samples equal in size to the human-authored set for the same unit, average across samples, and then average across units or assignments. We report results for both original and masked text.

5.3 Domain Measures

We evaluate diversity as coverage over meaningful forms of disciplinary work. We use Claude-Sonnet-5 as an LLM-judge to label the generated data according to codebooks adapted from educational research. Appendix E presents the full codebooks and LLM-judge prompts.

Math Items. We assess Task Type, which captures the mathematical activity students are asked to perform. These activities shape opportunities to engage in different forms of mathematical thinking (Swan, 2005; Burkhardt & Swan, 2013), such as classifying mathematical objects, interpreting or constructing representations, evaluating statements, creating problems or mathematical objects, and analyzing reasoning.

Essay Feedback. We assess Revision Type, which captures the kind of revision work made available to the student. Comments may ask students to fix an issue, elaborate an idea, synthesize ideas across the essay, or evaluate the effectiveness of their writing (Bereiter & Scardamalia, 1987). Rhetorical Strategy captures how the feedback is delivered, including open questions, criticism, praise, reader responses, and direct corrections (Straub, 1997).

Generator	Math Items				Essay Feedback			
	Self-BLEU ↓		Emb. Sim. ↓		Self-BLEU ↓		Emb. Sim. ↓	
	Unmask	Mask	Unmask	Mask	Unmask	Mask	Unmask	Mask
Human	0.3942	0.6021	0.3474	0.4791	0.2699	0.3179	0.2455	0.2565
GPT-5.2	0.8520	0.8789	0.6512	0.6929	0.4603	0.5271	0.3689	0.4093
+ Verbalized sampling	0.5159	0.7156	0.3905	0.5283	0.3577	0.4325	0.3160	0.3512
+ Negative sampling	0.5554	0.6932	0.4032	0.5320	0.3576	0.4298	0.3166	0.3504
Gemini 3.5-Flash	0.8026	0.8494	0.5364	0.6034	0.4823	0.5453	0.3116	0.3796
+ Verbalized sampling	0.4618	0.6151	0.3416	0.4749	0.4048	0.4836	0.2950	0.3855
+ Negative sampling	0.4711	0.5995	0.3419	0.4643	0.3240	0.3949	0.2866	0.3306
Llama-3.3-70B Versatile	0.8820	0.9033	0.6637	0.7256	0.5345	0.6217	0.3428	0.3890
+ Verbalized sampling	0.7764	0.8755	0.4053	0.5543	0.5699	0.6219	0.3437	0.3622
+ Negative sampling	0.7763	0.8815	0.4120	0.5749	0.5518	0.6198	0.3415	0.3625
Qwen-3.6-27B	0.6830	0.7540	0.4906	0.5619	0.3905	0.5752	0.3170	0.4053
+ Verbalized sampling	0.4792	0.6763	0.3642	0.5153	0.3035	0.5369	0.2850	0.3658
+ Negative sampling	0.5077	0.6559	0.3574	0.4941	0.3239	0.5638	0.2936	0.3889
Llama-3.1-8B-Instruct	0.7714	0.8253	0.5280	0.6470	0.5044	0.5909	0.3676	0.4080
+ Verbalized sampling	0.8698	0.9368	0.4303	0.5872	0.6499	0.6996	0.3749	0.3965
+ Negative sampling	0.8421	0.9230	0.5120	0.6570	0.6072	0.6632	0.3674	0.3926
+ SFT	0.6800	0.7842	0.4263	0.5667	0.4314	0.4624	0.2760	0.2900
+ DPO	0.4885	0.6145	0.4579	0.5671	0.6732	0.7841	0.3410	0.4170
Qwen-2.5-7B-Instruct	0.9313	0.9388	0.7348	0.7651	0.5029	0.6518	0.3124	0.3899
+ Verbalized sampling	0.7641	0.9247	0.4216	0.5941	0.7660	0.8483	0.3193	0.4060
+ Negative sampling	0.7837	0.9038	0.5069	0.6978	0.7338	0.8032	0.3166	0.3720
+ SFT	0.8615	0.8920	0.5694	0.6453	0.3959	0.4473	0.2702	0.2844
+ DPO	0.7600	0.8095	0.5290	0.6079	0.4937	0.6266	0.3040	0.3665

Table 2: **Lexical and semantic diversity across math items and essay feedback.** We report average pairwise Self-BLEU and embedding similarity (Emb. Sim.) within topic; lower values indicate greater diversity. Masked variants are computed after pre-processing to remove superficial variation.

6 Results

6.1 LLMs do not recover the diversity in human-authored content.

Table 2 compares lexical and semantic diversity in model generations and human-authored references for math items and essay feedback. Direct prompting scores, on average, 68.5% and 81.1% higher than human-authoring on Self-BLEU for math items and essay feedback, respectively. Outputs are more homogeneous lexically than semantically, with embedding similarity 53.3% and 46.1% higher than human-authored content.

6.2 Diversity interventions narrow but do not close the gap.

Diversity prompting reduces reduces Self-BLEU by 26.5% and embedding similarity by 20.2% for frontier models (GPT-5.2, Gemini-3.5-Flash). For open-weight models, prompting had no significant effect on Self-BLEU and reduced embedding similarity by 11.7%. Despite diversity prompting, outputs remain 66.2% higher than human-authoring on Self-BLEU and 27.3% higher on embedding similarity. SFT and DPO generally reduced Self-BLEU and embedding similarity relative to the base models, with the exception of DPO for essay feedback, but outputs remained more homogeneous than human-authored content. Post-training also often reduced quality. For math, Llama-3.1-8B-Instruct DPO improved Self-BLEU, but only 76% of problems were solvable and 82% understandable; for Qwen-2.5-7B-Instruct, it offered little advantage over prompting. For feedback, DPO did not improve on the baseline, while SFT produced only 75% accurate or context-relevant comments.

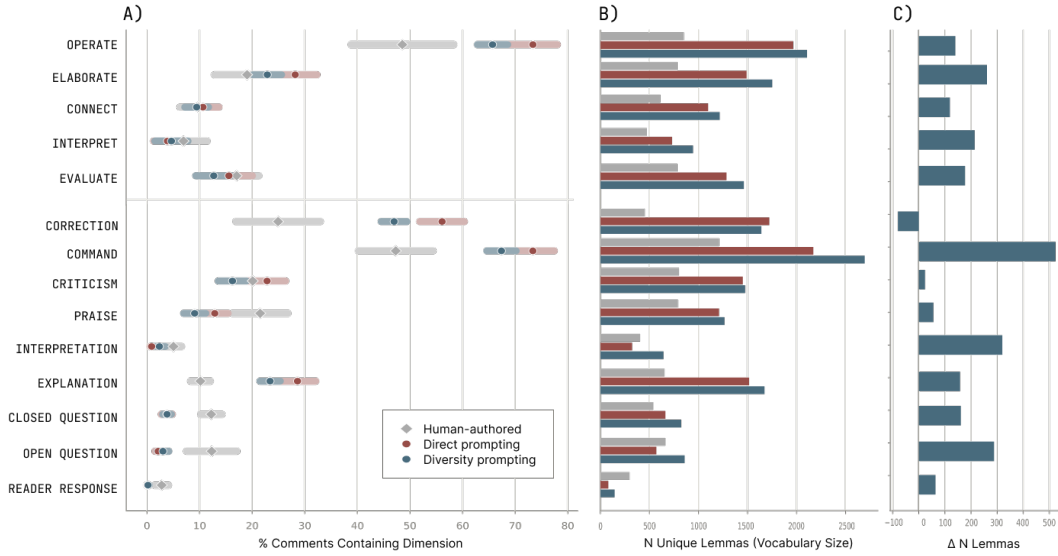


Figure 4: **Diversity of essay feedback.** Panel A shows the distribution of comments across revision types and rhetorical strategies. Panel B reports the vocabulary size of comments for each dimension. Panel C show the vocabulary gained from diversity prompting, compared to direct prompting. Vocabularies are computed as lemmas on masked text.

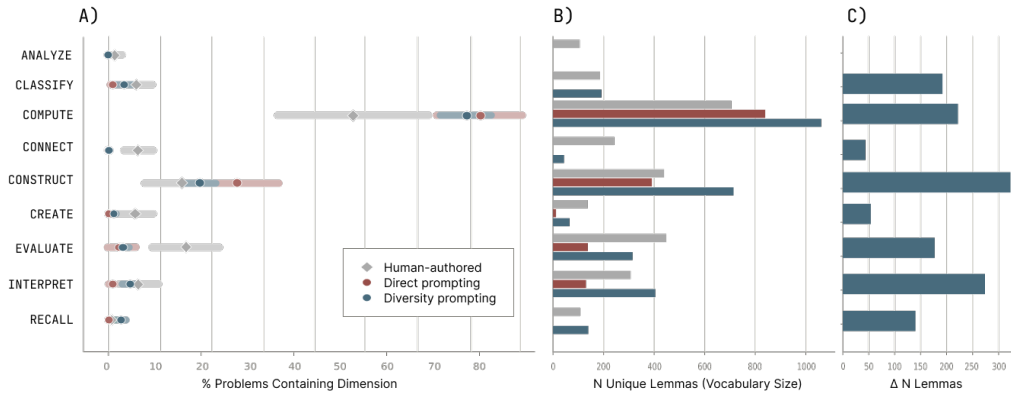


Figure 5: **Diversity of math items.** Panel A shows the distribution of problems across task types. Panel B reports the vocabulary size of problems for each type. Panel C show the vocabulary gained from diversity prompting, compared to direct prompting. Vocabularies are computed as lemmas on masked text.

6.3 LLMs collapse around dominant instructional forms.

Across both tasks, models amplify the instructional forms that are already most common in human-authored content, while forms that are less common in human demonstrations become rarer still. As shown in Figure 4, feedback comments that ask students to *operate* and perform mechanical fixes without judgment or new reasoning appears in 48% of human comments and 73% of LLM comments under direct prompting. Comments that ask students to *elaborate* and simply produce more text increases from 18% to 28%. In contrast, feedback that asks students to interpret the implications of their ideas, evaluate the effectiveness of their writing, or connect ideas across the essay is underrepresented in both human-authored and LLM-generated feedback. In terms of rhetorical strategy, direct *corrections* comprise 25% of human comments, but 56% of model comments, and commands comprise 47% and 73%, respectively. Open and closed questions and reader responses, already uncommon in human feedback, are nearly absent from model feedback.

Figure 5 reports the same pattern in math items. Tasks that ask students to *compute* an answer represent 53% of human-authored items and 80% of LLM comments under direct prompting. Requests students to *construct* representations increases from 16% to 28%. All other task types are under-represented, and diversity prompting does little to recover these forms. Models produce very few problems asking students to analyze worked reasoning, connect representations or create problems and models.

6.4 Diversity gains largely reflect surface variation.

Though diversity prompting does little to change the distribution of pedagogical priorities and representation of instructional forms, it expands the language used to express them. Figures 4 and 5 show that model vocabulary exceeds human vocabulary overall. In feedback, diversity prompting added 321 lemmas to comments calling for students to *elaborate* and 737 lemmas to comments delivered through *commands*, two forms that were already common under direct prompting. It also added 406 lemmas to comments offering an *interpretation* of student writing and 363 lemmas to *open questions*, though these forms remained infrequent.

In math, *compute* vocabulary grew from roughly 830 to 1,052 unique lemmas under diversity prompting, while compute problems continued to represent roughly 80% of generated items. For rarer task types, vocabulary for items asking students to *evaluate* mathematical claims more than doubled in unique lemmas while remaining around 3% of problems, and vocabulary for items asking students to *interpret* mathematical situations grew by 274 lemmas while appearing in only about 5%. By contrast, tasks asking students to *analyze* worked reasoning remains essentially absent, and tasks for *connecting* different representations and *creating* mathematical objects gain relatively little vocabulary. Models can therefore find many new ways to phrase familiar instructional work, and sometimes new language for rarer forms, without substantially changing which forms they generate.

The values removed during masking reveal an additional source of surface variation. For math items, diversity prompting was more effective for frontier models than open-weight models, an outcome driven in part by their having capitalized on adding a lot more whole numbers, decimals and fractions, while smaller models did not. This gain does not extend to text variables, as no intervention closed this gap. As shown in Figure 6, human-authored problems use 158 unique text variables, compared to 75 from models under direct prompting. For feedback, models diversity outputs by increasing the amount of rewritten and inserted text. Rewrites increase from about 12 characters in human-authored feedback comments to 25 characters for models under diversity prompting.

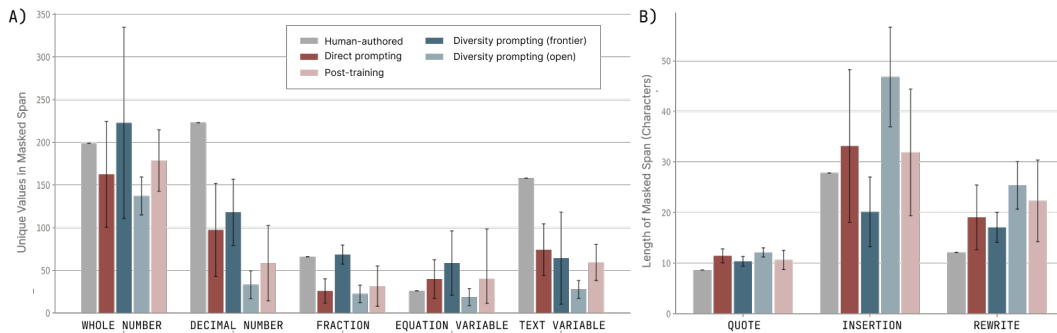


Figure 6: **Diversity of masked text in math items and essay feedback.** Panel A shows the number of unique values in each mask type in math items. Panel B shows the length of masked text in essay feedback. Bars show means; error bars denote 95% confidence intervals across models.

7 Discussion

Homogenization is consequential in education when it narrows the kinds of disciplinary work students are invited to do and the forms through which that work is mediated. Across

both tasks, models amplified dominant routines associated with relatively shallow student work in the human distribution of instructional practices. In writing, the distribution shifts toward correction, elaboration, and knowledge telling rather than the more demanding work of interpretation, evaluation, synthesis, and knowledge transformation (Bereiter & Scardamalia, 1987). In math, models overwhelmingly produce compute tasks, rather than tasks that ask students to reason about representations, claims, strategies, and mathematical structure (Swan, 2005).

While existing studies measuring lexical or semantic similarity have established mode collapse as a loss of variety or originality (Zhang et al., 2026a), we demonstrate the direction of this concentration. Skilled instructional practice is uncommon precisely because it requires expertise, time, and judgment—qualities filtered out by a system that preferentially reproduces the most normative forms. Diversity prompting only produced substantial variation when models were operating within instructional forms they already generated readily. In this sense, diversity prompting is useful as a probe of model repertoire, revealing the instructional practices that models have had little opportunity to learn. Though our interventions do not explore training objectives that explicitly reward diversity (Lanchantin et al., 2025), our results suggest a limit to such approaches when the desired instructional forms are themselves sparsely represented human demonstrations.

Expanding pedagogical diversity will require deliberately increasing the representation of high-quality instructional practices that are currently scarce in both human practice and model outputs. This means collecting less conventional examples of teaching and investing in the disciplinary knowledge, creative capacity, and professional judgment that allow teachers to produce them. As models learn from and amplify the routines most visible in educational data, teacher expertise becomes central to the future data environment. Educational AI should not become a mechanism through which the profession’s most routine practices crowd out its most ambitious ones.

Acknowledgments

This work was supported by Overdeck Family Foundation and the Stanford Institute for Human-Centered Artificial Intelligence (HAI). We thank Jim Malamut and Kristyn Wilson for their expertise in developing the domain measures and codebooks. We are grateful to all the teachers whose collaboration and insights have supported this work over the years.

References

- Carl Bereiter and Marlene Scardamalia. *The psychology of written composition*. L. Erlbaum Associates Hillsdale, NJ, 1987.
- Pierre Bourdieu. *Outline of a Theory of Practice*. Cambridge University Press, Cambridge, 1977.
- Okan Bulut, Damien C. Cormier, and Jinnie Shin. Automatic item generation. *Educational Measurement: Issues and Practice*, 43(3):24–38, 2024.
- Hugh Burkhardt and Malcolm Swan. Task design for systemic improvement: Principles and frameworks. *Task design in mathematics education: Proceedings of ICMI Study*, 22:431–439, 2013.
- David Carless and David Boud. The development of student feedback literacy: Enabling uptake of feedback. *Assessment & evaluation in higher education*, 43(8):1315–1325, 2018.
- Yishan Du, Conrad Borchers, and Mutlu Cukurova. Benchmarking educational llms with analytics: A case study on gender bias in feedback, 2026. URL <https://arxiv.org/abs/2511.08225>.
- Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning, 2023. URL <https://arxiv.org/abs/2210.02410>.

- Ben Hamner, Jaison Morgan, lynnvande, Mark Shermis, and Tom Vander Ark. The hewlett foundation: Automated essay scoring, 2012. URL <https://kaggle.com/competitions/ap-aes>. Kaggle.
- Bihao Hu, Jiayi Zhu, Yiyang Pei, and Xiaoqing Gu. Exploring the potential of LLM to enhance teaching plans through teaching simulation. *npj Science of Learning*, 10:7, 2025a. doi: 10.1038/s41539-025-00300-x.
- Jian Hu, Xibin Wu, Wei Shen, Jason Klein Liu, Zilin Zhu, Weixun Wang, Songlin Jiang, Haoran Wang, Hao Chen, Bin Chen, Weikai Fang, Xianyu, Yu Cao, Haotian Xu, and Yiming Liu. Openrlhf: An easy-to-use, scalable and high-performance rlhf framework, 2025b. URL <https://arxiv.org/abs/2405.11143>.
- Shomik Jain, Jack Lanchantin, Maximilian Nickel, Candace Ross, Karen Ullrich, Ashia Wilson, and Jamelle Watson-Daniels. Task-dependent evaluation of llm output homogenization: A taxonomy-guided framework, 2026. URL <https://arxiv.org/abs/2509.21267>.
- Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, Alon Albalak, and Yejin Choi. Artificial hivemind: The open-ended homogeneity of language models (and beyond), 2025. URL <https://arxiv.org/abs/2510.22954>.
- Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günemann, Eyke Hüllermeier, et al. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274, 2023.
- Muhammad Haris Khan and Joel wester. Seeing the hivemind: A consensus-aware interaction technique for mitigating ai homogenization, 2026. URL <https://arxiv.org/abs/2606.09587>.
- Sal Khan. Harnessing gpt-4 so that all students benefit. a nonprofit approach for equal access. *Khan Academy Blog*, March 2023. URL <https://blog.khanacademy.org/harnessing-gpt-4-so-that-all-students-benefit-a-nonprofit-approach-for-equal-access/>.
- Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the effects of rlhf on llm generalisation and diversity, 2024. URL <https://arxiv.org/abs/2310.06452>.
- Angelika Kullberg, Ulla Runesson Kempe, and Ference Marton. What is made possible to learn when using the variation theory of learning in teaching mathematics? *ZDM Mathematics Education*, 49(4):559–569, 2017. doi: 10.1007/s11858-017-0858-4.
- Catherine Lammert, Samuel DeJulio, Stephanie Grote-Garcia, and Lucretia M Fraga. Better than nothing? an analysis of ai-generated lesson plans using the universal design for learning & transition frameworks. *The Clearing house: a Journal of educational Strategies, issues and ideas*, 97(5):168–175, 2024.
- Jack Lanchantin, Angelica Chen, Shehzaad Dhuliawala, Ping Yu, Jason Weston, Sainbayar Sukhbaatar, and Ilia Kulikov. Diverse preference optimization, 2025. URL <https://arxiv.org/abs/2501.18101>.
- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models, 2016. URL <https://arxiv.org/abs/1510.03055>.
- Ziniu Li, Congliang Chen, Tian Xu, Zeyu Qin, Jiancong Xiao, Zhi-Quan Luo, and Ruoyu Sun. Preserving diversity in supervised fine-tuning of large language models, 2025. URL <https://arxiv.org/abs/2408.16673>.
- Mingyi Liu. The alignment tax: Response homogenization in aligned llms and its implications for uncertainty estimation, 2026. URL <https://arxiv.org/abs/2603.24124>.

- Ying Liu, Can Li, Ting Zhang, Mei Wang, Qiannan Zhu, Jian Li, and Hua Huang. Discerning minds or generic tutors? evaluating instructional guidance capabilities in socratic llms, 2025. URL <https://arxiv.org/abs/2508.06583>.
- Dan C. Lortie. *Schoolteacher: A Sociological Study*. University of Chicago Press, Chicago, IL, 1975.
- Yining Lu, Dixuan Wang, Tianjian Li, Dongwei Jiang, Sanjeev Khudanpur, Meng Jiang, and Daniel Khashabi. Benchmarking language model creativity: A case study on code generation. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *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. 2776–2794, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.141. URL <https://aclanthology.org/2025.naacl-long.141/>.
- Li Lucy, Tal August, Rose E. Wang, Luca Soldaini, Courtney Allison, and Kyle Lo. Mathfish: Evaluating language model math reasoning via grounding in educational curricula, 2024. URL <https://arxiv.org/abs/2408.04226>.
- Ference Marton. *Necessary conditions of learning*. Routledge, 2014.
- Ference Marton and Shirley Booth. *Learning and Awareness*. Routledge, 1997.
- Ramya Namuduri, Yating Wu, Anshun Asher Zheng, Manya Wadhwa, Greg Durrett, and Junyi Jessy Li. Qudsim: Quantifying discourse similarities in llm-generated text, 2025. URL <https://arxiv.org/abs/2504.09373>.
- Krishna Pillutla, Lang Liu, John Thickstun, Sean Welleck, Swabha Swayamdipta, Rowan Zellers, Sewoong Oh, Yejin Choi, and Zaid Harchaoui. Mauve scores for generative models: Theory and practice, 2023. URL <https://arxiv.org/abs/2212.14578>.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024. URL <https://arxiv.org/abs/2305.18290>.
- Kai Ruan, Xuan Wang, Jixiang Hong, Peng Wang, Yang Liu, and Hao Sun. Evaluating llms’ divergent thinking capabilities for scientific idea generation with minimal context, 2026. URL <https://arxiv.org/abs/2412.17596>.
- Nicy Scaria, Suma Dharani Chenna, and Deepak Subramani. *Automated Educational Question Generation at Different Bloom’s Skill Levels Using Large Language Models: Strategies and Evaluation*, pp. 165–179. Springer Nature Switzerland, 2024. ISBN 9783031642999. doi: 10.1007/978-3-031-64299-9_12. URL http://dx.doi.org/10.1007/978-3-031-64299-9_12.
- Chantal Shaib, Venkata S. Govindarajan, Joe Barrow, Jiuding Sun, Alexa F. Siu, Byron C. Wallace, and Ani Nenkova. Standardizing the measurement of text diversity: A tool and a comparative analysis of scores, 2026. URL <https://arxiv.org/abs/2403.00553>.
- Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022): 755–759, 2024.
- Mary Kay Stein and Margaret Schwan Smith. Mathematical tasks as a framework for reflection: From research to practice. *Mathematics Teaching in the Middle School*, 3(4): 268–275, 1998.
- Jacob Steiss, Tamara Tate, Steve Graham, Jazmin Cruz, Michael Hebert, Jiali Wang, Youngsun Moon, Waverly Tseng, Mark Warschauer, and Carol Booth Olson. Comparing the quality of human and chatgpt feedback of students’ writing. *Learning and Instruction*, 91:101894, 2024. ISSN 0959-4752. doi: <https://doi.org/10.1016/j.learninstruc.2024.101894>. URL <https://www.sciencedirect.com/science/article/pii/S0959475224000215>.

- Richard Straub. Students' reactions to teacher comments: An exploratory study. *Research in the Teaching of English*, 31(1):91–119, 1997.
- Malcolm Swan. Improving learning in mathematics: Challenges and strategies. Technical report, Department for Education and Skills Standards Unit, 2005.
- Mei Tan, Christopher Mah, and Dorottya Demszky. Reframing authority: A computational measure of power-affirming feedback on student writing. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*, pp. 417–421, 2024.
- Mei Tan, Lena Phalen, and Dorottya Demszky. Marked pedagogies: Examining linguistic biases in personalized automated writing feedback. In *Proceedings of the LAK26: 16th International Learning Analytics and Knowledge Conference*, LAK 2026, pp. 568–577. ACM, April 2026. doi: 10.1145/3785022.3785113. URL <http://dx.doi.org/10.1145/3785022.3785113>.
- LearnLM Team, Abhinit Modi, Aditya Srikanth Veerubhotla, Aliya Rysbek, Andrea Huber, Brett Wiltshire, Brian Veprek, Daniel Gillick, Daniel Kasenberg, Derek Ahmed, Irina Jurenka, James Cohan, Jennifer She, Julia Wilkowski, Kaiz Alarakyia, Kevin R. McKee, Lisa Wang, Markus Kunesch, Mike Schaeckermann, Miruna Pislari, Nikhil Joshi, Parsa Mahmoudieh, Paul Jhun, Sara Wiltberger, Shakir Mohamed, Shashank Agarwal, Shubham Milind Phal, Sun Jae Lee, Theofilos Strinopoulos, Wei-Jen Ko, Amy Wang, Ankit Anand, Avishkar Bhoopchand, Dan Wild, Divya Pandya, Filip Bar, Garth Graham, Holger Winnemoeller, Mahvish Nagda, Prateek Kolhar, Renee Schneider, Shaojian Zhu, Stephanie Chan, Steve Yadlowsky, Viknesh Sounderajah, and Yannis Assael. Learnlm: Improving gemini for learning, 2025. URL <https://arxiv.org/abs/2412.16429>.
- Guy Tevet and Jonathan Berant. Evaluating the evaluation of diversity in natural language generation, 2021. URL <https://arxiv.org/abs/2004.02990>.
- The Learning Agency Lab. The persuade dataset, 2024. URL <https://the-learning-agency-lab.com/learning-exchange/persuade-dataset/>. Accessed: 2024-07-01.
- Iain Weissburg, Sathvika Anand, Sharon Levy, and Haewon Jeong. Llm are biased teachers: Evaluating llm bias in personalized education, 2025. URL <https://arxiv.org/abs/2410.14012>.
- Qingsong Wen, Jing Liang, Carles Sierra, Rose Luckin, Richard Tong, Zitao Liu, Peng Cui, and Jiliang Tang. Ai for education (ai4edu): Advancing personalized education with llm and adaptive learning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, pp. 6743–6744, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704901. doi: 10.1145/3637528.3671498. URL <https://doi.org/10.1145/3637528.3671498>.
- Fan Wu, Emily Black, and Varun Chandrasekaran. Generative monoculture in large language models, 2024. URL <https://arxiv.org/abs/2407.02209>.
- Weijia Xu, Nebojsa Jojic, Sudha Rao, Chris Brockett, and Bill Dolan. Echoes in ai: Quantifying lack of plot diversity in llm outputs. *Proceedings of the National Academy of Sciences*, 122(35), August 2025. ISSN 1091-6490. doi: 10.1073/pnas.2504966122. URL <http://dx.doi.org/10.1073/pnas.2504966122>.
- Su-Youn Yoon, Eva Miszoglad, and Lisa R Pierce. Evaluation of chatgpt feedback on ell writers' coherence and cohesion. *arXiv preprint arXiv:2310.06505*, 2023.
- Jiayi Zhang, Simon Yu, Derek Chong, Anthony Sicilia, Michael R. Tomz, Christopher D. Manning, and Weiyan Shi. Verbalized sampling: How to mitigate mode collapse and unlock llm diversity, 2026a. URL <https://arxiv.org/abs/2510.01171>.
- Lily Hong Zhang, Smitha Milli, Karen Jusko, Jonathan Smith, Brandon Amos, Wassim Bouaziz, Manon Revel, Jack Kussman, Yasha Sheynin, Lisa Titus, Bhaktipriya Radharapu, Jane Yu, Vidya Sarma, Kris Rose, and Maximilian Nickel. Cultivating pluralism in algorithmic monoculture: The community alignment dataset, 2026b. URL <https://arxiv.org/abs/2507.09650>.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020. URL <https://arxiv.org/abs/1904.09675>.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models, 2018. URL <https://arxiv.org/abs/1802.01886>.

Unit ID	Unit Description
HS_2_6 (High School Geometry Unit 6)	This unit brings together students' experience from previous years with their new understanding from this course for an in-depth study of coordinate geometry. Students encounter a new coordinate transformation notation such as $(x, y) \rightarrow (x + 3, y + 1)$, which connects transformations to functions. Then they use transformations and the Pythagorean Theorem to build equations of circles, parabolas, parallel lines, and perpendicular lines from definitions. Students apply these ideas to proofs, such as classifying quadrilaterals. Finally, students use weighted averages to partition segments, scale figures, and locate the intersection point of the medians of a triangle.
MS_1_4 (Grade 6 Unit 4)	In this unit, students examine how the relative sizes of numerator and denominator affect the size of their quotient when numerator or denominator (or both) is a fraction. They acquire the understanding that dividing by a/b has the same outcome as multiplying by b , then by $1/a$. They compute quotients of fractions. They solve problems involving lengths and areas of figures with fractional side lengths and extend the formula for the volume of a right rectangular prism to prisms with fractional edge lengths and use it to solve problems. They use tape diagrams, equations, and expressions to represent situations involving partitive or quotitive interpretations of division with fractions. Given a multiplication or division equation or expression with fractions, they describe a situation that it could represent. They use tape diagrams and equations in reasoning about situations that involve multiplication and division of fractions.
MS_2_5 (Grade 7 Unit 5)	In this unit, students interpret signed numbers in contexts (e.g., temperature, elevation, deposit and withdrawal, position, direction, speed and velocity, percent change) together with their sums, differences, products, and quotients. ("Signed numbers" include all rational numbers, written as decimals or in the form a/b .) Students use tables and number line diagrams to represent sums and differences of signed numbers or changes in quantities represented by signed numbers such as temperature or elevation, becoming more fluent in writing different numerical addition and subtraction equations that express the same relationship. They compute sums and differences of signed numbers. They plot points in the plane with signed number coordinates, representing and interpreting sums and differences of coordinates. They view situations in which objects are traveling at constant speed (familiar from previous units) as proportional relationships. For these situations, students use multiplication equations to represent changes in position on number line diagrams or distance traveled, and interpret positive and negative velocities in context. They become more fluent in writing different numerical multiplication and division equations for the same relationship. Students extend their use of the "next to" notation (which they used in expressions such as $5x$ and $6(3 + 2)$ in grade 6) to include negative numbers and products of numbers, e.g., writing $-5x$ and $(-5)(-10)$ rather than $(-5) \cdot (x)$ and $(-5) \cdot (-10)$. They extend their use of the fraction bar to include variables as well as numbers, writing $-8.5/x$ as well as $-\frac{8.5}{x}$.

Table 3: Example unit descriptions provided to models as generation context.

A Source Data

Tables 3 and 4 show examples of the source context provided to models during generation. For math items, this context consisted of the Illustrative Mathematics unit description summarizing the target learning objectives. For feedback, it consisted of the writing assignment and the full student essay.

B Masking

To estimate how much of the observed variation is driven by surface-level values, we created masked versions of the math problems and feedback comments. We prompted GPT-5-nano to apply the masks using the prompts shown in Figures 7 and 8. To programmatically

Assignment Prompt	Student Essay
Read “Narciso Rodriguez” by Narciso Rodriguez. Describe the mood created by the author in the memoir. Support your answer with relevant and specific information from the memoir.	The mood in the memoir that the author creates is a very warm and relaxed sort of mood. The way the author tells you about his parents; the selflessness and the way they put their kids first before themselves really just makes you think about your own parents and how good they are to you and what they would do for you. The way the author talks about his home makes you think of all the colors in the house. I think that the house would be a golden red color yellow, white, brown and orange on the insides. It reminds me of the way they would have cooked. They would have cooked with colorful peppers, onions, tortillas and oils. It makes you think about what colors you have in your home and what you would cook with. The music sounds wonderful. Even though they don’t demonstrate what it actually sounds like, you can easily imagine it. I imagine it laying in the background while all the adults are cooking and the kids are playing in colorful rooms with the golden sunlight shining in through an open window. I love the mood that this memoir puts me in.
Your principal is reconsidering the school’s cell phone policy, choosing between two options. Policy 1: allow phones at school for use during lunch and free times, turned off during class. Policy 2: do not allow phones at school at all. Write a letter to your principal convincing her which policy is better. Support your position with specific reasons.	Dear Principal, I have heard about the new cell phone policies that you have been thinking of. I feel that policy number 2 is the better one of the 2 choices. hear are a couple of the reasons why I personally think this. Students enjoy using their phone every day. I feel that sometimes the can use it a little to much during school. I think that letting students use their phones during lunch and free time is an excellent idea. Also students wouldn’t be interrupting class every day with their cell phones. This would be a lot better than just getting rid of cell phones completely. As principal you might want to consider possibly figuring out a schedule were students have free time during a couple of their classes. Doing this would not only help the new policy, but keep it going throughout the school year. students would appreciate it if during a boring class they got a 5-15 minute break to relax and use their phones. But, there would be some rules behind the free time. You would have to think of them so that students wouldn’t be doing stuff that is against school rules. So principal please consider the second policy for cell phones at our school. students do want to use their cell phones and not keep them at home. I hope you liked the reasons why please se

Table 4: Example prompt-essay pairs provided to models as generation context.

narrow the cases requiring human review, we used `difflib.SequenceMatcher` to align each original text with its masked version and recover the original value replaced by each tag.

Math items. We manually audited cases in which the extraction procedure either failed to recover a value for a tag or recovered a value that did not match the expected tag type, such as a nondigit value for a whole-number tag. A common source of extraction failures was adjacent tags, as in “line AB” when A and B had previously been introduced as separate points. Of the 1,668 cases flagged for this issue, 1,200 were resolved programmatically using stricter regular expressions for the character types associated with each tag. Of the remaining 468 cases, we manually corrected three masking errors. Extraction also failed when model outputs included process text or formatting labels before the masked problem, such as “Wait, I need to ensure correct formatting” or “## Anonymized text:”. We audited 749 of such cases and corrected 69 corrupted outputs.

Essay feedback. For feedback comments, we manually audited 56 cases in which extraction failed and 99 cases in which masking replaced the entire comment. We also flagged 888 cases with empty extracted values and 956 cases with unmatched quotation marks. All 888

```

You will be shown a math problem.

Create a maskedProblem by replacing:
- Whole numbers with [whole_number]
- Decimals with [decimal_number]
- Fractions with [fraction]
- Variables in equations, such as x or y, with [equation_variable]
- Names and labels with [text_variable]

Preserve all other wording and structure. Return only the masked problem.

## Example 1
problem: Jada has a sleeping bag that is rated for 30 degrees F.
Write an inequality that represents the outdoor temperature at which
Jada will be able to stay warm.
maskedProblem: [text_variable] has a sleeping bag that is rated for
[whole_number] degrees F. Write an inequality that represents the
outdoor temperature at which [text_variable] will be able to stay warm.

## Example 2
problem: Point $A$ is located at  $(-3, 2)$  and point $B$ is
located at  $(7, 17)$  on a coordinate plane. Find the coordinates of
point $P$ that partitions the directed line segment $AB$ in a 3:2$
ratio.
maskedProblem: Point [text_variable] is located at  $([whole\_number],$ 
 $[whole\_number])$  and point [text_variable] is located at
 $([whole\_number], [whole\_number])$  on a coordinate plane. Find the
coordinates of [text_variable] that partitions the directed line
segment between [text_variable] and [text_variable] in a
 $[whole\_number]:[whole\_number]$  ratio.

## Example 3
problem: Write a number puzzle that results in the equation
 $3x - 5 = 16$ .
maskedProblem: Write a number puzzle that results in the equation
 $[whole\_number][equation\_variable] - [whole\_number] = [whole\_number]$ .

```

Figure 7: System prompt for masking math problems.

empty values and 25 of the unmatched quotation cases reflected extraction errors, which we corrected programmatically. For the remaining 168,441 comments, we ranked outputs by character overlap between the original and masked versions and manually reviewed the 1,011 comments with less than 30% overlap. Across 3,010 audited comments, we found and manually corrected two masking errors.

C Acceptability

To establish a baseline level of quality, we assessed the acceptability of each output before conducting the diversity analyses. We used a panel of three LLM judges, GPT-5-nano, Claude-Haiku-4.5, and Gemini-2.5-Flash-Lite, using the prompts shown in Figures 9 and 10. For math problems, judges rated understandable language, solvability, and relevance to the unit. For feedback comments, they rated understandable language, accuracy, and helpfulness. We report acceptability ratings in Table 5. We pooled the three criteria within each judgment, treating an output as invalid if the judge assigned a 0 to any criterion.

Math items. We manually reviewed the 1,616 outputs for which the three judges did not unanimously agree. Human review agreed with the majority judgment in 73.8% of these cases and with the dissenting judgment in 26.2%.

Through the audit, we discovered that judges accepted invalid problems with contradictory language, multiple-choice questions without the correct answer among the options, problems with references to missing diagrams, tables, or previous problems, and other malformed generations that repeated sub-questions or ended abruptly mid-sentence. We also found that judges rejected some valid problems, including simple interpretation questions, such as identifying a unit rate or writing an equation, answerable questions containing

You will be shown teacher feedback on a student essay.
The input is formatted with:
- EXCERPT: a window of essay text with brackets marking the highlighted portion
- COMMENT: the teacher feedback

Create a maskedComment by replacing:
- wording quoted or near-quoted from the EXCERPT with <QUOTE>
- replacement wording for something in the EXCERPT with <REWRITE>
- new wording the teacher proposes adding to the essay with <INSERTION>

Use the EXCERPT to distinguish quotes, rewrites, and insertions. Mask the smallest meaningful span and preserve all other wording and structure. Remove any quotation marks surrounding the masked span. Return only the maskedComment.

Example 1
EXCERPT: Driverless cars could help people.
COMMENT: Clarify this idea by saying something like, "One good thing about driverless cars is that they could help elderly people."
maskedComment: Clarify this idea by saying something like, <INSERTION>.

Example 2
EXCERPT: Their are more disadvantages because the driver fins more problems then expected.
COMMENT: Correct the spelling of 'their' to 'there', 'fins' to 'finds', and 'then' to 'than'.
maskedComment: Correct the spelling of <QUOTE> to <REWRITE>, <QUOTE> to <REWRITE>, and <QUOTE> to <REWRITE>.

Example 3
EXCERPT: The mood creates unity and a feeling of togetherness.
COMMENT: How does 'unity' differ from or align with your initial description of the mood as 'togetherness'?
maskedComment: How does <QUOTE> differ from or align with your initial description of the mood as <QUOTE>?

Example 4
EXCERPT: Over all I see benfit in allowing driverless cars.
COMMENT: Use 'Overall, I see benefits' instead of 'Over all I see benefit' - make sure to proofread for typos and grammatical errors.
maskedComment: Use <REWRITE> instead of <QUOTE> - make sure to proofread for typos and grammatical errors.

Example 5
EXCERPT: Long-distance driving can have physical effects.
COMMENT: Consider using more formal or technical language to describe the physical effects of long-distance driving, such as 'fatigue' or 'muscle strain'.
maskedComment: Consider using more formal or technical language to describe the physical effects of long-distance driving, such as <INSERTION> or <INSERTION>.

Figure 8: System prompt for masking teacher feedback comments.

"distractor" information, and questions with valid edge-case answers, such as an undefined slope. Notably, only 85% of the human-authored math problems were judged as unit-relevant. We manually reviewed all 79 human-authored problems flagged as invalid for this reason and found that they were indeed review problems included in the source curriculum.

Essay feedback. Because judgments about accuracy and helpfulness comprised 95% of disagreements, we treated comments flagged by any judge for understandable language as invalid. We retained unanimous judgments and marked the remaining 37,056 comments for further review. Reviewing all comments by hand was not feasible, so we randomly sampled 200 disagreements to identify common sources of ambiguity and incorporated these patterns into an adjudication prompt for Claude-Sonnet-5. The adjudicator agreed with the minority judgment in 36.4% of cases to overturn a majority valid judgment and 28.5% of cases to overturn a majority invalid judgment. We manually reviewed a sample of 5,000 cases in which the adjudicator overturned the majority judgment and agreed with the adjudicator in 96% of cases.

Generator	Math Items			Essay Feedback		
	Underst.	Solvable	Unit Rel.	Underst.	Accurate	Improves
Human	99.1	98.3	85.4	91.3	88.6	88.2
GPT-5.2	99.9	99.9	99.9	100.0	100.0	100.0
+ Verbalized sampling	100.0	99.9	99.6	100.0	100.0	100.0
+ Negative sampling	99.9	99.9	99.2	100.0	100.0	100.0
Gemini-3.5-Flash	100.0	100.0	99.9	100.0	100.0	100.0
+ Verbalized sampling	99.8	99.7	99.3	100.0	100.0	100.0
+ Negative sampling	99.8	99.6	98.9	100.0	100.0	100.0
Llama-3.3-70B	98.6	97.9	98.3	99.5	95.3	95.5
+ Verbalized sampling	98.0	96.1	94.0	97.2	88.2	88.3
+ Negative sampling	98.7	97.4	94.0	97.0	88.4	88.5
Qwen-3.6-27B	99.3	98.2	99.3	98.3	88.3	89.2
+ Verbalized sampling	99.4	98.1	97.5	94.0	75.2	76.4
+ Negative sampling	99.3	98.8	98.3	96.4	78.9	80.2
Llama-3.1-8B-Instruct	96.2	92.6	91.6	98.9	88.4	89.7
+ Verbalized sampling	98.6	95.4	90.6	96.1	85.2	86.0
+ Negative sampling	94.1	91.5	90.8	95.8	84.4	85.3
+ SFT	98.1	94.8	90.9	93.6	75.0	75.3
+ DPO	81.8	76.1	81.9	98.0	81.8	83.8
Qwen-2.5-7B	99.2	98.1	98.0	97.9	82.3	83.8
+ Verbalized sampling	99.3	98.2	94.4	97.3	80.2	82.1
+ Negative sampling	97.7	96.7	93.4	97.7	81.6	82.9
+ SFT	99.0	98.3	97.4	92.2	75.1	74.0
+ DPO	98.0	95.8	95.9	97.5	81.1	82.4

Table 5: **Validity of generated items across math problems and essay feedback.** Percentage of generated items satisfying each quality criterion, as judged automatically. Higher values indicate greater validity; cell shading marks values below 95%, with darker shading denoting lower validity.

Through the audit, we discovered that judges accepted feedback that called something missing or incorrect when it was present and correct elsewhere in the essay, feedback that misdiagnosed an issue, feedback that suggested content that contradicted the student’s ideas. We also found that judges rejected valid comments that were vague but still grounded in the student’s writing, as well as feedback referencing external facts or sources that could not be verified from the essay but were not demonstrably false. Notably, only 88% of human-authored feedback comments were judged as accurate and helpful. We manually reviewed all 448 human-authored comments flagged as invalid for this reason. Human comments were often penalized for forms of feedback that were less common in model outputs. These included praise that served a relational or personal function, such as “I love that you stay in contact with the library. This is so cute,” and short but legitimate corrections, such as “comma after to”. Judges also rejected comments that correctly identified contradictions or problems visible only from the full essay rather than the selected excerpt. These included comments that redirected the student to the assignment or pointed out that the student’s current claim conflicted with an earlier claim. At the same time, the human data included genuinely invalid comments, such as “same comment as earlier,” and corrections too abbreviated to interpret without additional context.

D Post-training

We implemented SFT and DPO using OpenRLHF (Hu et al., 2025b). Models were loaded in bfloat16. For SFT, human-authored outputs served as training targets. For DPO, human-authored outputs served as chosen responses and were paired with model-generated rejected responses.

For math items, we fine-tune using SFT for 2 epochs with a learning rate of (5×10^{-7}) and a batch size of 32. For DPO, we train for 1 epoch with a learning rate of (5×10^{-7}) , batch size of 32, and β coefficient 0.1. We use bfloat 16 precision for both and train until convergence.

```

You will be shown a math problem.
The input is formatted with:
* GRADE_LEVEL: the intended grade level
* LESSON_NARRATIVE: the lesson goals and content
* MATH_PROBLEM: the problem students would see

Evaluate whether the MATH_PROBLEM valid. Score it on these 3 binary criteria:

1. understandable_language
* 1 = the problem is clear, readable, and complete enough to understand
* 0 = the problem is broken, confusing, malformed, or too incomplete to
  interpret

2. solvable
* 1 = the problem has at least one valid answer as written
* 0 = it is missing necessary information, contradictory, mathematically
  invalid, or too ambiguous to solve
Solve the problem before scoring it. Open-ended problems are solvable if
there are acceptable answers.

3. unit_relevant
* 1 = the mathematical work required fits the GRADE_LEVEL and
  LESSON_NARRATIVE
* 0 = it assesses unrelated content or is inappropriate for the intended
  grade level
Use both the problem and its solution to judge relevance.

Return ONLY valid JSON:
{"understandable_language": 0, "solvable": 0, "unit_relevant": 0}

```

Figure 9: System prompt for the math-problem validity judges.

```

You will be shown feedback on a student essay.
The input is formatted with:
- ASSIGNMENT: the writing task or prompt the student was responding to
- FULL_ESSAY: the student's full essay
- EXCERPT: a window of essay text with brackets marking the highlighted
  portion
- COMMENT: the feedback comment

Evaluate whether the COMMENT is valid feedback on the student's writing.
Score it on these 3 binary criteria:

1. understandable_language
* 1 = the comment is understandable natural language and functions as
  interpretable feedback in relation to the excerpt
* 0 = the comment is gibberish, broken, truncated, malformed, cannot be
  understood as feedback on the excerpt, is unrelated to the excerpt,
  or consists of unrelated fragments

2. accurate
* 1 = the comment correctly assesses or responds to
  something in the writing and any proposed revision is valid
* 0 = the comment misreads the essay, invents or
  misdiagnoses an issue, or proposes a false, identical, contradictory, or inappropriate correction

3. context_relevant
* 1 = the comment makes a meaningful suggestion or
  observation that improves the student's writing or supports the student's development as a writer.
* 0 = the comment makes no meaningful request or
  observation, is off-topic or directs the student toward unproductive revisions

Return ONLY valid JSON:
{"understandable_language": 0, "accurate": 0, "context_relevant": 0}

```

Figure 10: System prompt for the essay-feedback validity judges.

For essay feedback, we used a learning rate of 1×10^{-5} and trained for two epochs for SFT, as the smaller dataset benefited from additional training. For DPO, we initially tested a learning rate of 1×10^{-6} , $\beta = 0.2$, and two epochs, but these settings produced unstable

Code	Definition	Example
Recall	Retrieving a fact, definition, formula, or convention. Nothing is produced nor examined.	What is the appropriate unit for volume: cm, cm ² , or cm ³ ?
Interpret	Surfacing or reasoning about meaning: what a mathematical object, symbol, or quantity refers to or implies about a situation; showing conceptual understanding, not carrying out a procedure or producing a representation.	If the 5 represents the number of people, what does the 4 represent?
Compute	Finding a numerical, symbolic, or exact answer: calculating, simplifying, solving, or determining a value.	Solve $3x + 5 = 20$ for x . A shirt costs \$40 and is 25% off. Find the sale price.
Construct	Producing a mathematical representation, such as a graph, table, diagram, dot plot, number line, expression, equation, model, or visual display. Lower-level than Create, more routine and less inventive.	Make a dot plot for a given list of data values. Write an expression to represent the perimeter of this figure.
Create	Generating something new and open-ended where many results are valid: inventing an example, data set, problem, representation, or situation that meets given conditions. More inventive and less routine than Construct.	Create a data set with a mean of 10 and a median of 8. Create a problem that can be solved with $3/4 \div 1/2$.
Classify	Organizing objects by a property: sorting into groups or categories, labeling, or ordering along a dimension.	Group expressions according to whether they are linear, quadratic, or exponential. Order from smallest to largest: number of pennies in a stack 1 ft high, number of books in a stack 1 ft high, ...
Connect	Engaging with how different representations express the same mathematical object, quantity, relationship, or structure; matching, comparing, recognizing similarity. (Use when this is the point of the item, not when contending with multiple representations is a vehicle for solving something else.)	Select all expressions equivalent to $16x + 36$. Would this information be better represented using a table or a double number line?
Evaluate	Judging whether a mathematical statement, claim, conjecture, or generalization holds: true or false, sometimes/always/never, or whether it fits a situation.	Is the sum of two rational numbers always rational? Are $2b + b$ and $3b$ equivalent for all values of b ?
Analyze	Examining mathematical reasoning, a worked solution, a strategy, an argument, a proof, an error, or the results a procedure produced. Includes comparing strategies, critiquing reasoning, debugging mistakes, explaining errors, or deciding whether a solution method is valid.	A student says $4(x + 3) = 4x + 3$; explain the mistake. Lin measured a diameter vertically (3.5 cm) and horizontally (3.6 cm); give possible reasons the measurements differ.

Table 6: Codebook for classifying the mathematical Task Type

training and malformed outputs. We ultimately used a learning rate of 5×10^{-7} and trained for one epoch.

E Domain Measures

Code	Definition	Examples
Operate	The feedback asks the student to retrieve, add, correct, or execute something straightforward. The remedy is clear and does not require judgment, reconsideration, or new reasoning.	Fix the spelling of “necessary.” Add a citation for this statistic. Start a new paragraph here.
Elaborate	The feedback asks the student to add more content, detail, explanation, context, or development of a kind they already understand and could produce. The student elaborates without needing to reconsider their claim, belief, purpose, or rhetorical plan.	Can you give an example here? Add something to connect these paragraphs. Your introduction needs more context. Expand on this point.
Connect	The feedback asks the student to work out relationships between ideas, evidence, claims, structure, counterarguments, or rhetorical choices. The student must construct a connection that is missing, unclear, contradictory, or underdeveloped.	How does this evidence support this claim? You have two contradictory points here. How do they fit together?
Interpret	The feedback asks the student to develop an idea by explaining its meaning, significance, implication, or application. The student goes beyond adding more content to show what the idea means or why it matters.	What can we infer about her feelings from this quote? Where can we see this same phenomenon in action today?
Evaluate	The feedback asks the student to reconsider validity, credibility, appropriateness, effectiveness, or fit. The student must assess whether a claim, source, choice, stance, tone, word, or rhetorical move works.	What might a critic say in response to this argument? Is this the right tone for this audience? I’m not sure this is true.

Table 7: Codebook for classifying the feedback Revision Type.

You are an expert math-education researcher coding math problems by the type of cognitive work the problem asks the student to do.

Apply these rules:
Read the problem, decide which single code best captures the primary work the student is asked to do, and call the `assign\_code` function with ONLY the code id -- no explanation, no extra text. If a problem seems to span two codes, pick the one that captures its PRIMARY demand -- the thing the student's final answer/response actually consists of.

Each problem gets EXACTLY ONE code from this fixed set:
 \{
 Label:
 Definition:
 Example:
 \}

Figure 11: System prompt for labeling math item task type.

Code	Definition	Examples
Correction	The teacher directly fixes something in the text, supplies replacement wording, or gives an explicit edit to the student's text.	Comma splice. Should be "their" not "there." Change X to Y.
Command	The teacher directs the student to take an action, without supplying the exact edit. This includes imperatives and softened directives like "try," "consider," or "you might."	Explain why... Add a transition here. Consider adding more detail. Try to develop this argument further.
Criticism	The teacher evaluates the writing negatively or identifies a problem. The teacher is judging the writing, even if the criticism is hedged.	This is unclear. These arguments are not convincing. I'm not sure this argument works.
Praise	The teacher evaluates the writing positively. Includes simple praise and elaborated praise.	Good. Strong argument here. You've defined the issues clearly.
Interpretation	The teacher neutrally describes, summarizes, or plays back what the student did in the text, without evaluating it or asking for a change.	You're arguing here that legalization would reduce crime. In your conclusion you return to the opening question.
Explanation	The teacher provides information, reasoning, instruction, a lesson, or a general principle. The comment explains why something matters or teaches the student something.	Transitions help readers follow the logic between your ideas. Direct quotes need quotation marks because readers need to know whose words these are.
Closed Question	The teacher asks a question that implies an evaluation, problem, or expected answer. The teacher's perspective is implicit in the question.	Have you proved this point by just saying it is so? Can you break up these sentences so I can follow your ideas more easily?
Open Question	The teacher asks a question that genuinely opens a problem for the student to think through. The student has real latitude in how to respond.	What is it that you want me most to know about your position? How can you explain and support these views?
Reader Response	The teacher shares their experience as a reader — what they noticed, felt, understood, wondered, or struggled with — without directly judging the writing or directing a change.	I found myself confused here. I can tell you feel strongly about this subject. I wonder what you gain by having this paragraph here.

Table 8: Codebook for classifying the Rhetorical Strategy used to deliver feedback.

You are an expert education researcher coding essay feedback by the type of revision work the comment asks the student to do. The input is formatted with:

- EXCERPT: a window of essay text with brackets marking the highlighted portion
- COMMENT: the teacher feedback

Select exactly ONE code that best characterizes this comment overall.

```
## Codebook:
\{
Label:
Definition:
Example:
\}
```

Figure 12: System prompt for labeling essay feedback revision type.

```
You are an expert education researcher coding essay feedback by the rhetorical strategy it uses to communicate
to the student. The input is formatted with:
- EXCERPT: a window of essay text with brackets marking the highlighted portion
- COMMENT: the teacher feedback

Select every code that applies. Most comments will get exactly one label, but some legitimately combine more
than one (for example, a comment that both praises and asks an open question, or one that both corrects an
error and explains the underlying rule).

## Codebook:
\{
Label:
Definition:
Example:
\}
```

Figure 13: System prompt for labeling essay feedback rhetorical strategy.

F Domain-specific measures on math item generation

Comparisons of domain-specific measures further show a lack of coverage in model-generated problems. For example, as shown in Table F, human-authored math content targets all nine cognitive task types, whereas LLM-generated math items often cover only four task types: COMPUTE (73–90%), CONSTRUCT (12–24%), EVALUATE (0–6%), and INTERPRET (0–1%)¹. In particular, ANALYZE and CONNECT each account for less than 1% of the generated problems, based on labeling by Sonnet 5 as the LLM judge. While 46% of human-authored problems target Compute, they also cover other task types, such as CONSTRUCT (14.9%) and EVALUATE (15.1%), with ANALYZE (1.3%) and RECALL (0.9%) being the least frequent categories. In contrast, COMPUTE and CONSTRUCT together account for 93–99% of model-generated problems. Similar trends are observed when Gemini 3.1 Flash Lite is used as the LLM judge to label the math items.

Post-training diversity gains on math items do not translate to improved task-type coverage. While Table 2 shows improvements in lexical and semantic diversity after SFT and DPO for both Llama and Qwen-Instruct models, these interventions do not lead to meaningful changes in the percentages of generated problems targeting different task types. In particular, the two most frequent task categories, COMPUTE and CONSTRUCT, continue to account for 91–99% of model-generated problems. The seven less frequent categories (ANALYZE, CLASSIFY, CONNECT, CREATE, EVALUATE, INTERPRET, and RECALL), which account for 31% of human-generated content, remain substantially under-represented (less than 1%) in the model generations even after training with the human data (Table F).

Task type	Percentage (%) of generated problems targeting each task type								
	ANALYZE	CLASSIFY	COMPUTE	CONNECT	CONSTRUCT	CREATE	EVALUATE	INTERPRET	RECALL
Human	11.1	3.3	46.7	9.6	10.2	5.1	9.6	4.0	0.4
GPT-5.2	0	0	79.4	0.4	13.4	0.4	5.2	1.3	0
+ Verbalized sampling	0	2.8	79	0.2	10.5	0.7	2.8	1.7	2.4
+ Negative sampling	0	1.3	72.5	0.7	15.2	1.8	4.1	2.8	1.7
Gemini-3.5-Flash	0.9	0	93.2	0	4.2	0.2	0.4	1.1	0
+ Verbalized sampling	2	1.8	67.3	1.3	17.7	1.1	2.2	4.3	2.2
+ Negative sampling	1.5	1.9	59.1	0.7	26.5	1.9	2.6	4.9	0.9
Llama-3.3-70B	0	0.6	86.4	0	10	0	0.2	2.8	0
+ Verbalized sampling	2.4	0.4	84.3	0	8.1	0.6	0	2.4	2.0
+ Negative sampling	0.2	0.6	84.7	0.2	11.1	0	0.4	2.3	0.6
Qwen-3.6-27B	1.9	0.4	66.5	0.2	30.2	0.2	0.2	0.2	0.2
+ Verbalized sampling	1.1	1.9	65.1	0.2	22.1	1.7	3.0	2.8	2.1
+ Negative sampling	2.6	1.7	63.1	0.2	20.8	1.9	3.6	4.3	1.9
Llama-3.1-8B	0	0.2	86.5	0	12.5	0	0	0.2	0.6
+ Verbalized sampling	0	3.0	70.3	0	19.5	1.6	2.2	1	2.2
+ Negative sampling	0.3	0	74.3	0.9	18.7	2.9	1.8	0.3	0.9
+ SFT	0.4	0.4	89.7	0.2	7.9	0	1.0	0.2	0.2
+ DPO	0.3	0	74.3	0.9	18.7	2.9	1.8	0.3	0.9
Qwen-2.5-7B	0	0	61.7	0	38.3	0	0	0	0
+ Verbalized sampling	0.2	2.2	81.4	0	8.2	0.2	4.3	1.6	2.0
+ Negative sampling	0.2	3.3	87.7	0	8.3	0	0	0.2	0.2
+ SFT	0.2	0	73.2	0	26.4	0.2	0	0	0
+ DPO	1.0	0	62.1	0	35.7	0	0.2	1.0	0

Table 9: Percentage of generated problems targeting each category of mathematical activity (Top: labeled by Gemini 3.1 Flash Lite as the LLM-judge; Bottom: labeled by Sonnet 5). We use an LLM judge to label the mathematical task type targeted by each problem across 10 units, after filtering for acceptability, and report the percentage of problems corresponding to each task type. Task types representing less than 1% of the generated content are highlighted in red. For robustness, we report results from two frontier LLM judges: Sonnet 5 and Gemini 3.1 Flash Lite. In both cases, we observe similar trends in the overall distribution of task types.

¹Percentages reported for the frontier models GPT-5.2, Gemini-3.5-Flash, and Llama-3.3-70B. Individual model-by-prompting-method breakdowns are provided in Table F.

Task type	Percentage (%) of generated problems targeting each task type								
	ANALYZE	CLASSIFY	COMPUTE	CONNECT	CONSTRUCT	CREATE	EVALUATE	INTERPRET	RECALL
Human	1.3	4.7	46.2	6.7	14.9	4.0	15.1	6.2	0.9
GPT-5.2	0	0	72.6	0	20.6	0	5.5	1.3	0
+ Verbalized sampling	0	2.4	74.6	0	14.4	0.7	2.9	2.4	2.6
+ Negative sampling	0	1.7	66.2	0	20.5	1.5	5.2	2.6	2.4
Gemini-3.5-Flash	0	0	89.5	0	8.3	0.2	1.3	0.7	0
+ Verbalized sampling	0	1.8	62.5	0.2	23.7	1.1	3.5	4.8	2.4
+ Negative sampling	0	2.1	53.9	0.6	31.2	1.7	3.9	5.6	1.1
Llama-3.3-70B	0	0.6	86.4	0	12.3	0	0.4	0.4	0
+ Verbalized sampling	0.2	0.2	79.3	0	11.0	0.6	1.4	3.3	3.9
+ Negative sampling	0	0.4	79.0	0.4	14.2	0.0	0.4	3.3	2.3
Qwen-3.6-27B	0.2	0.6	54.6	0.2	42.9	0	1.1	0.2	0.2
+ Verbalized sampling	0	1.7	60.8	0	24.4	1.7	3.0	4.5	3.9
+ Negative sampling	0	1.9	57.5	0.4	26.0	2.1	3.6	5.8	2.8
Llama-3.1-8B	0	0.2	86.9	0	12.3	0	0	0	0.6
+ Verbalized sampling	0	0.2	86.9	0	12.3	0	0	0	0.6
+ Negative sampling	0	2.7	71.2	0	24.2	0	0	0.4	1.5
+ SFT	0	0.4	86.4	0	10.7	0	1.4	0.8	0.2
+ DPO	0	0.3	69.9	0	22.2	3.2	2.3	0.6	1.5
Qwen-2.5-7B	0	0	56.8	0	43.2	0	0	0	0
+ Verbalized sampling	0	2.0	79.8	0	9.2	0.2	4.3	2.2	2.3
+ Negative sampling	0	3.1	87.2	0	8.7	0	0	0.6	0.4
+ SFT	0	0	68.3	0	31.6	0	0.2	0	0
+ DPO	0	0	58.6	0	40.2	0	0.8	0.4	0

G Examples

Selected Examples from Diversity Prompting

Generate problems about high school math on coordinate geometry.

GPT-5.2 with verbalized sampling:

- A point P is at (2, 5). Apply the transformation $(x, y) \rightarrow (x+3, y1)$. What are the new coordinates of P? (Probability: 0.55)
- A point Q is at (7, 4). Apply the transformation $(x, y) \rightarrow (x6, y+8)$. What are the new coordinates of Q? (Probability: 0.52)

Gemini 3.5 Flash with verbalized sampling:

- Apply the transformation $T(x, y) = (x - 3, y + 4)$ to the point $P(5, -2)$ and find the coordinates of its image. (Probability: 0.92)
- Determine the coordinate rule $(x, y) \rightarrow (x + a, y + b)$ that represents the translation mapping $A(1, 2)$ to $A'(-3, 7)$. (Probability: 0.89)

Human-generated samples:

- The triangle whose vertices are (2,5), (3,1), and (4,2) is transformed by the rule $(x,y) \rightarrow (x-2,y+4)$. Is the image similar or congruent to the original figure? (A): The image is congruent to the original triangle. (B): The image is similar but not congruent to the original triangle. (C): The image is neither similar nor congruent to the original triangle.
- Triangle ABC has vertices at (-2,0), (-1,6), and (6,0). What is the point of intersection of the triangle's medians?