Illusory Truth or Mere Exposure? Model-Dependent Repetition Effects in LLM-Based Social Media Simulations

Azza Bouleimen, Nicolò Pagan & Anikó Hannák
Department of Computer Science
University of Zurich
Andreasstrasse 15, 8050, Zurich, Switzerland
{azza.bouleimen,nicolo.pagan,aniko.hannak}@uzh.ch

Abstract

Generative agent-based models (GABMs) are increasingly used to simulate social media dynamics, including misinformation spread. For such social simulations to be valid proxies of human behavior, LLM agents should replicate established human cognitive biases, among them the Illusory Truth Effect (ITE), whereby repeated exposure to a claim increases its perceived truth value. We investigate whether and how the ITE manifests across four LLMs (`Gemma-3-4b-it`, `Qwen2.5-7B-Instruct`, `Llama-3.1-8B-Instruct`, and `GPT-5-nano`) in a social media simulation context. We propose a two-phase within-context experimental design that embeds the repetition manipulation inside a realistic news feed interaction. Using this design, we collect 336,000 truth, importance, sentiment, and interest ratings across 100 statements, 10 feed variants, and 3 replications. The key comparison is between ratings assigned to repeated statements, seen throughout a simulation phase, and completely unseen ones, rated within the same experimental context window. We distinguish genuine ITE (truth-specific repetition boost) from mere exposure effects (uniform boost across all attributes). We run an OLS regression followed by a Linear Mixed-Effect Model to account for differences across models and ratings. Our results reveal four qualitatively distinct patterns: `Gemma-3-4b-it` exhibits a genuine ITE; `Qwen2.5-7B-Instruct` shows a mere exposure effect; `GPT-5-nano` displays no repetition effect on truth and mild skepticism toward repeated content; `Llama-3.1-8B-Instruct` shows a small truth boost alongside decreases in evaluative dimensions, inconsistent with either account. Crucially, temperature has no effect on these findings, and a variance decomposition highlights the high context-sensitivity of LLM rating behavior. Our findings caution against assuming uniform ITE replication across LLMs in social simulations, while suggesting that `Gemma-3-4b-it` may offer the most behaviorally realistic approximation for misinformation-related simulations.

1 Introduction

Generative Agent Based Modeling (ABM), with the recent advances in LLMs, is the subject of a dynamic and growing scholarly interest. In particular, it carries the promise of simulating social and economic phenomena with an unprecedented level of realism Bail (2024). Several studies already presented promising generative simulations of social behavior, especially in the context of social network simulations Cau et al. (2025); Park et al. (2023) and social media Ferraro et al. (2024); Yang et al. (2024); Gao et al. (2023); Törnberg et al. (2023); Composta et al. (2025); Larooij & Törnberg (2025a). These simulations have a number of relevant applications such as testing algorithms for prosocial behavior, e.g., reducing toxic discourse, polarization, misinformation Törnberg et al. (2023); Larooij & Törnberg (2025a), or modeling opinion dynamics Composta et al. (2025); Cau et al. (2025).

Although, some of the mentioned papers attempt to validate their systems, thoroughly answering the question: "to what extent these simulations are a good proxy to human

behavior" remains challenging and limited. In fact, validating social simulations requires, as stated in Sen et al. (2025), Zhou et al. (2025) and in **?**, careful methodological approaches. For the social simulations to be realistic and reliable enough, we need, among other things, to ensure that the agents display similar cognitive bias to humans in an informational environment.

One widespread and established cognitive bias in humans which manifests also on social media is the Illusory Truth Effect (ITE). Initially observed in Hasher et al. (1977), the ITE simply states that repeated claims are rated as more true than new ones Bacon (1979). This effect is highly relevant to social media usage and misinformation adoption. In fact, it has been shown that prior exposure increases perceived accuracy of fake news Pennycook et al. (2018), even if it contradicts previous knowledge Fazio (2020); Fazio et al. (2015). Repetition has also been shown to increase the perceived truth of conspiracy theories Béna et al. (2023) and rumors DiFonzo et al. (2016). The ITE's impact goes beyond belief, it increases sharing intentions of news headlines (including false ones) and decreases how unethical an act is perceived to be Udry & Barber (2024).

In this work, we contribute to the scholarly efforts on validating generative ABM and focus on investigating the manifestations –or not— of the ITE, in the context of social media simulation in particular. We design an experiment where we instruct an LLM to behave like a social media user and feed it a series of tweets iteratively to interact with. In every iteration of news feeds, we include a statement (in a form of a tweet) which appears in every one of the presented feeds. At the end of the experiment, we task the LLM to rate the truth value of the repeated statement as well as additional completely new statements (unseen statements). We then compare, across several runs of the experiment, the difference between repeated and unseen statements. We push the analysis further to distinguish between an ITE specific to truth or a mere exposure effect Zajonc (1968). Our main research question (RQ) is: "is ITE effect present in social media simulations?" If an effect is observed, we expend our analysis to cover the following questions:

- RQ1: Is the increased rating effect specific to truth or it is rather a mere exposure effect?
- RQ2: How does the ITE and mere exposure effects differ across models and their temperatures?
- RQ3: Does the observed effect depend on the experimental context?

Our findings suggest that LLMs display different levels and types of effects in relation to the ITE in social media simulations. Additionally, we find that the effect is highly dependent on the experimental context as well as the specific statement to rate.

2 Related Work

The Illusory Truth Effect

Hasher et al. (1977) first demonstrated the ITE phenomenon in 1977: college students rated 60 statements across three sessions separated by two-week intervals, with some statements repeated across sessions and others appearing only once. Repeated statements received significantly higher validity judgments than non-repeated ones — for both true and false statements — suggesting that frequency of occurrence serves as a criterion for judging the referential validity of plausible statements. These findings were subsequently replicated in several studies referring to the effect as the Illusion of Validity (Bacon, 1979; Schwartz, 1982; Gigerenzer, 1984; Begg et al., 1985; Arkes et al., 1989).

The ITE is typically explained as a fluency effect: initial exposure makes processing during the test phase more fluent, and fluency is taken as an indicator of truth (Reber & Schwarz, 1999). A related phenomenon, the mere exposure effect (Zajonc, 1968), occurs when increased familiarity increases ratings in other dimensions such as liking or importance. Most ITE studies do not test whether the effect is specific to truth judgments, leaving it conflated with mere exposure. Our study addresses this by comparing the effect on truth ratings versus other dimensions, following the discriminant validity paradigm (Campbell & Fiske, 1959) and the approach of Griffin et al. (2023).

Beyond controlled experiments, the ITE has real-world implications for misinformation. As reviewed in Udry & Barber (2024), it reproduces in digital environments, particularly social media. Pennycook et al. (2018) found that even a single prior exposure increases perceived accuracy, while Vellani et al. (2023) showed that participants were more likely to share previously encountered statements — with this relationship mediated by perceived accuracy, making repetition a driver of misinformation spread.

Generative ABM for Social Simulations

Concurrently, generative agent-based models (ABMs) powered by LLMs have become state-of-the-art in social simulations. Building on Park et al. (2023)'s foundational work on believable generative agents, researchers have rapidly adopted this approach across a range of social simulation studies (Ferraro et al., 2024; Yang et al., 2024; Gao et al., 2023; Törnberg et al., 2023; Composta et al., 2025; Larooij & Törnberg, 2025a; De Marzo et al., 2023; Bouleimen et al., 2025). These studies demonstrate LLMs' promise as surrogates for human participants, reproducing interaction and linguistic patterns (Ferraro et al., 2024; Yang et al., 2024), group dynamics (Bouleimen et al., 2025), coherent content generation (Composta et al., 2025), and emotional and attitudinal responses (Gao et al., 2023). At the network level, LLMs reproduce social phenomena such as homophily (Ferraro et al., 2024), the friendship paradox (Orlando et al., 2025), opinion dynamics (Cau et al., 2025), scale-free network structures (De Marzo et al., 2023), political clustering and echo chambers (Ferraro et al., 2024; Larooij & Törnberg, 2025a), and concentrated influence among elite agents (Larooij & Törnberg, 2025a). That said, limitations remain, including difficulty reproducing toxic behavior and limited content diversity (Composta et al., 2025; Pagan et al., 2025).

ITE and LLMs

While this body of work sheds light on high-level interaction and network dynamics, individual cognitive biases in LLMs remain underexplored. A growing literature addresses LLM belief and knowledge change in this context. Fastowski & Kasneci (2024) introduce the notion of *knowledge drift*, showing that repeatedly exposing a model to incorrect information increases its uncertainty and can shift initially correct answers. Geng et al. (2025) further show that models tend to shift moral or political positions after multi-model discussions or exposure to opposing viewpoints. Whether such susceptibility is useful for realistic social simulation remains debated, but it underscores significant context-dependence in LLM behavior.

To the best of our knowledge, Griffin et al. (2023) is the only study to examine the ITE in LLMs directly. Replicating Hasher et al. (1977) and using `text-davinci-003`, the authors explore the ITE manifestations through a set of true and false claims. They found that repeated statements received significantly higher truth ratings, with no significant effect on interest, sentiment, or importance — ruling out a mere exposure explanation. However, this study is limited to a now-deprecated model, and recent findings (Fastowski & Kasneci, 2024; Geng et al., 2025) suggest that context-sensitivity may affect ITE-related behavior. It is therefore not clear that the effect observed by Griffin et al. (2023) will generalize to newer models or different prompting contexts. This motivates our work, which tests multiple models across varied social simulation contexts, drawing on the experimental design of Griffin et al. (2023) and adapting it accordingly.

3 Experiment

Design. We design a controlled experiment to test whether LLMs exhibit ITE-like behavior when acting as agents in a social media simulation. The experiment is run on four LLMs: `Gemma-3-4b-it` (Kamath et al., 2025), `Qwen2.5-7B-Instruct` (Yang et al., 2025), `Llama-3.1-8B-Instruct` (Dubey et al., 2024), and `GPT-5-nano` (OpenAI, 2025). For the three open-weight models, we run the experiment at two temperature settings (0.1 and 1.0) to assess sensitivity to this parameter, while `GPT-5-nano` has a fixed default temperature of 1.0. This yields 7 model-temperature combinations in total. The statements used throughout the experiment are the 100 statements authored by Griffin et al. (2023), which are not available online, minimizing the risk that they appear verbatim in the training data of the tested LLMs (see Appendix A.1). Each statement is rated along one of four attributes: *truth*, *importance*, *sentiment*, and *interest*. Truth is the primary attribute of interest for our study, directly operationalizing the ITE. The remaining three serve to distinguish a genuine ITE

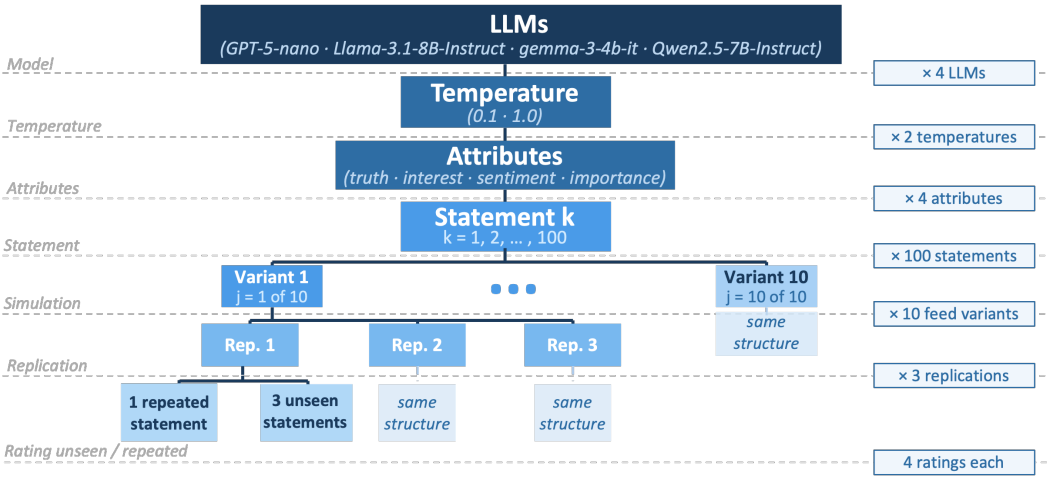


Figure 1: Illustration of the experiment.

from a mere exposure effect (Zajonc, 1968): if repetition boosts ratings uniformly across all attributes, the effect is better attributed to familiarity than to a truth-specific bias.

Procedure. For each combination of model, temperature, statement, and rating attribute, the experiment unfolds in two sequential phases within the same context window. In the *simulation phase*, the LLM is prompted to act as a social media user scrolling through a news feed and commenting, resharing, or not interacting with the tweets (see Appendix A.2). It is shown 5 feeds sequentially, each containing 4 tweets. One tweet, the *repeated statement*, appears in every feed, while the remaining 15 are unique filler statements drawn randomly from the pool of 99 (excluding the repeated statement in this experiment), each appearing exactly once. The position of the repeated tweet within each feed is randomized. In the following *rating phase*, without resetting the context window, the LLM is asked to rate 4 claims on a 1-to-7 Likert scale along the target attribute. One of the 4 claims is the repeated statement; the other 3 are completely new, i.e., so far unseen by the LLM in the experiment. The position of the repeated claim in the rating list is again randomized. Scale levels for each attribute are described in Appendix A.3. With this setting, the LLM interacts with or rates a total 24 statements. Given the experiment with humans rating 32 statements each done in Griffin et al. (2023), we believe that our setting is suitable to replicate (if any exists) the ITE of humans in social media. To account for variability in the filler context and following the approach in Griffin et al. (2023), we generate 10 feed variants per statement, each with a different random draw of filler tweets. We replicate each variant 3 times identically to assess the determinism of model responses. This yields 30 ratings per statement in the *repeated* condition and 90 in the *unseen* condition, for each (attribute, model, temperature) combination, and 336,000 ratings in total: 7 model-temperature combinations × 4 attributes × 100 statements × 10 feed variants × 3 replications × 4 ratings per pair of phases. An illustration of the experiment is presented in 1.

4 Statistical Analysis

To address the RQs, we first implement an Ordinary Least Squares Regression (OLS) replicating the model used in Griffin et al. (2023). We then gain additional insights from the data through a Linear Mixed-Effects Model (LMEM). We detail both statistical models in the following subsections.

4.1 Descriptive Analysis: OLS Regression

To characterize the magnitude and direction of repetition effects for each model, attribute, and temperature combination, we fit separate OLS regressions of the form:

$$r' = r + \text{offset} + (r - 4) \times \text{tilt}, \tag{1}$$

where r denotes the mean rating a statement received across its unseen presentations, r' denotes the mean rating the same statement received across its repeated presentations, and

4 is the midpoint of the 1-to-7 rating scale. This parametrization yields two interpretable quantities. The *offset* estimates the predicted shift in rating due to repetition for a statement rated neutrally (4) when unseen. The *tilt* estimates how the repetition effect scales with the baseline (unseen) rating: a tilt of zero implies a uniform additive shift across all statements; a negative tilt implies a smaller effect for statements with high baseline ratings; a positive tilt implies a larger effect for statements with high baseline ratings. We fit each regression on 100 data points, one per statement, computing r and r' as means across the 90 unseen and 30 repeated experimental runs, respectively. We fit one regression per combination of model (4 levels), attribute (4 levels), and temperature (2 levels for open-weight models; 1 level for `GPT-5-nano`), yielding 28 regressions in total.

This model provides a descriptive understanding of the effects observed across every LLM, attribute, and temperature. However, it cannot compare effects across models and attributes, as observations are not independent across groups. In other words, with this model we cannot say whether the effect for truth is higher or lower than for the other attributes (RQ1), nor can we distinguish effects between models (RQ2). Moreover, any insights on the variance of the ratings is made impossible when we consider the mean rating per statement (RQ3).

4.2 Inferential Analysis: Linear Mixed-Effects Models

To address the limitations of the OLS regression, we adopt a LMEM. We first fit a preliminary model to measure the effects of temperature and replication on ratings (see Appendix A.5.2 for details).

Following the preliminary model results (see Section 5.2.1), we fit a main model with fixed effects for a three-way interaction between statement repetition, rating attribute, and LLM. This model lets us simultaneously test differences in the repetition effect across attributes and LLMs, addressing both RQ1 (truth-specificity of the rating effect: ITE vs. mere exposure) and RQ2 (differences in effects across LLMs). We also use the LMEM to conduct a variance component analysis, which helps us understand the potential impact of simulation context on LLM ratings of repeated statements (RQ3). Our data have a nested structure: individual rating observations are grouped within simulation runs (10 runs per statement in the repeated condition, `simulation_num`), which are in turn grouped within statements (100 different statements, `statement_id`). *Statement-level clustering* reflects that some statements receive intrinsically higher or lower ratings across all attributes and conditions, independent of repetition. *Simulation-level clustering* reflects that the surrounding feed context—the specific set of filler tweets presented alongside the repeated statement during the simulation phase—introduces additional variability in ratings beyond statement-level effects. To account for both clustering sources, we include nested random intercepts in the model.

The resulting LMEM is presented in Eq. 2:

$$\texttt{rating} \sim \texttt{repeated} \times \texttt{attribute} \times \texttt{model} + 1|\texttt{statement\_id/simulation\_num}, \quad (2)$$

where `repeated` is a binary variable indicating whether the rated statement was repeated (True) or unseen (False) during the simulation phase, `attribute` is the rating dimension (truth, importance, sentiment, interest), and `model` identifies the LLM (`GPT-5-nano`, `Llama-3.1-8B-Instruct`, `Gemma-3-4b-it`, and `Qwen2.5-7B-Instruct`). `statement_id` and `simulation_num` enter the model as nested random effects. The reference levels are: `repeated` = False, `attribute` = truth, `temperature` = 0.1.

Post-hoc Contrasts To extract interpretable per-model, per-attribute repetition effects from the three-way model, we compute the estimated marginal means (EMMs)—the model's best estimates of the average rating for each combination of repetition status (unseen or repeated), attribute, and LLM (Methodological details can be found in Appendix A.5.3). We then compute pairwise contrasts (`repeated` − `unseen`) within each attribute × model cell, that is, the difference between the estimated means of repeated and unseen ratings. A positive contrast indicates that repeated statements received higher ratings than unseen statements, while a negative contrast indicates the reverse. Comparing these contrasts across attributes and models addresses whether rating effects are truth-specific (RQ1) and whether these effects differ across models (RQ2).

Variance Component Analysis We compute marginal and conditional R^2 values and Intraclass Correlations (ICCs) at both the statement and simulation levels. Marginal R^2 reflects the variance explained by fixed effects alone, while conditional R^2 reflects the variance explained by both fixed and random effects. The ICC quantifies the proportion of total variance attributable to between-statement and between-simulation differences, respectively. Together, these measures reveal the sources of variance in the data and shed light on the potential impact of context on statement ratings, thereby addressing RQ3.

We fit the LMEM using the `nlme` package (Pinheiro & Bates, 2000) in R (R Core Team, 2024). We computed EMMs using the `emmeans` package (Lenth, 2025), and marginal and conditional R^2 values following the method of Nakagawa & Schielzeth (2013) using the `performance` package (Lüdecke et al., 2021).

5 Results

We run the experiment on the open weights models on A100 and H100 GPUs. The total running time for the 100 x 10 experiments per rating attribute was about 2.5 hours for `Qwen-2.5-7B-Instruct`, about 3 hours for `Llama-3.1-8B-Instruct`, and about 12 hours for `Gemma-3-4b-it`. We run the `GPT-5-nano` using asynchronous calls from the OpenAI API for which the experiment lasted only few minutes. We collected 336,000 ratings, from which we dropped 155. 151 of them came from `GPT-5-nano` where it either did not abide by the new rating task and kept interacting with the claim as a social media user or refrained from rating this happened in majority for statements 84 and 89. The remaining ratings we dropped were 4 from `Llama-3.1-8B-Instruct` where it refrained from rating statement 24 in four occasions. We detail the findings from 335,845 ratings in the following subsections.

5.1 Descriptive Analysis: Repetition Effects by Model and Attribute

We report in Tab. 1 the results of the OLS regression (Eq. 1). It shows the estimates of tilt and offset[1]. To illustrate some of the numbers in the results, we present in Fig. 2 the scatter plots and the corresponding regression lines for the four models studied when temperature is 1. Fig. 2 shows a difference in the effect across the LLMs. The remaining plots for temperature 0.1 and for the importance, sentiment and interest attributes can be found in Appendix A.4.

A first glance at the table indicates that our experiment shows almost no difference between the regression parameters from temperature 1 and from temperature 0.1. In fact, for truth, the highest difference in absolute value between the offsets and tilts are respectively 0.047 and 0.076. In all cases except one, the estimates are always with similar values, positive / negative sign and same statistical significance level. The only exception is for sentiment in `Llama-3.1-8B` experiment where the tilt is not significant for temperature 1 while it is significant when temperature is 0.1. This is a rather counterintuitive result as higher temperatures are expected to produce more diverse or creative output which may lead to a different behavior of the model towards the ratings.

Focusing on the truth rating attribute, we see that both the offset and the tilt are significant for `Llama-3.1-8B`, `Qwen2.5-7B` and for the `Gemma-3-4b-it`. In all three cases the offset is positive. This indicates an overall increase in the truth rating of the claims when repeated in the simulation phase compared to when unseen. For all these three models, the tilt is negative, which indicates that the observed rating's increase is of a lower magnitude for the highly rated unseen claims and, vice versa, of a higher magnitude for lowly rated unseen claims. As for `GPT-5-nano`, both estimates are not statistically significant highlighting no effect of previous exposure to claim on the truth rating of `GPT-5-nano`. Looking at the sentiment, interest and importance attributes, we see that the offset is always significant and positive for `Qwen2.5-7B` and `Gemma-3-4b-it` which indicates an increase in the average rating of these attributes for `Qwen2.5-7B` and `Gemma-3-4b-it`. For `GPT-5-nano`, the offset for

[1]We assessed the robustness of regression estimates to potential ceiling effects by repeating all regressions on the subset of statements with mean unseen ratings $r \leq 6$. Offset and tilt estimates were not materially affected, indicating that compression at the upper boundary of the rating scale did not substantially distort the regression coefficients.

sentiment and interest are significant but negative instead, suggesting a decrease in the average ratings of these attributes. Finally, `Llama-3.1-8B` shows a significant negative offset only for sentiment.

Considering these results and as detailed in Sec. 4, the OLS model fails in providing statistical inference about the cross level comparison between attributes and models to allow, at this level of the analysis, to make conclusions on the presence of absence of the ITE and whether the observed effect is different from a mere exposure effect. We then present, in the next two subsections, the results from the LMEM.

Table 1: Statistical model estimates for the four models across four rating attributes. Red = significant positive/negative offset; Blue = significant tilt. Bold = $p < 0.05$. Significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. †`GPT-5-nano` temperature is fixed to 1; Temp. 0.1 results not available.

Model	Rating Attr.	Param.	Temp. 1		Temp. 0.1	
			Est.	95% CI	Est.	95% CI
Llama-3.1-8B	truth	offset	**0.258****	**[0.073, 0.443]**	**0.269***	**[0.042, 0.496]**
		tilt	**−0.149****	**[-0.245, −0.052]**	**−0.225*****	**[−0.327, −0.122]**
	importance	offset	0.078	[−0.068, 0.223]	−0.102	[−0.275, 0.071]
		tilt	**−0.205*****	**[−0.302, −0.108]**	**−0.191*****	**[−0.296, −0.087]**
	sentiment	offset	**−0.231****	**[−0.360, −0.103]**	**−0.251*****	**[−0.385, −0.117]**
		tilt	−0.064	[−0.147, 0.018]	**−0.132****	**[−0.212, −0.052]**
	interest	offset	0.016	[−0.124, 0.155]	−0.010	[−0.195, 0.175]
		tilt	**−0.197*****	**[−0.285, −0.108]**	**−0.236*****	**[−0.345, −0.126]**
Qwen2.5-7B	truth	offset	**0.910*****	**[0.675, 1.145]**	**0.903*****	**[0.660, 1.146]**
		tilt	**−0.225*****	**[−0.334, −0.116]**	**−0.185****	**[−0.297, −0.073]**
	importance	offset	**1.003*****	**[0.812, 1.194]**	**1.034*****	**[0.828, 1.240]**
		tilt	**−0.216*****	**[−0.314, −0.118]**	**−0.215*****	**[−0.321, −0.109]**
	sentiment	offset	**0.451*****	**[0.309, 0.594]**	**0.438*****	**[0.301, 0.575]**
		tilt	−0.039	[−0.140, 0.061]	−0.019	[−0.114, 0.076]
	interest	offset	**0.642*****	**[0.463, 0.821]**	**0.546*****	**[0.380, 0.712]**
		tilt	−0.115	[−0.236, 0.006]	−0.096	[−0.209, 0.016]
Gemma-3-4B	truth	offset	**0.868*****	**[0.644, 1.092]**	**0.915*****	**[0.687, 1.144]**
		tilt	**−0.365*****	**[−0.469, −0.260]**	**−0.404*****	**[−0.510, −0.298]**
	importance	offset	**0.882*****	**[0.687, 1.077]**	**0.822*****	**[0.618, 1.026]**
		tilt	**−0.550*****	**[−0.669, −0.431]**	**−0.495*****	**[−0.618, −0.371]**
	sentiment	offset	**0.593*****	**[0.402, 0.783]**	**0.630*****	**[0.435, 0.825]**
		tilt	**−0.158****	**[−0.267, −0.048]**	**−0.147***	**[−0.261, −0.034]**
	interest	offset	**0.348****	**[0.141, 0.555]**	**0.314****	**[0.103, 0.524]**
		tilt	**−0.352*****	**[−0.486, −0.219]**	**−0.363*****	**[−0.501, −0.226]**
GPT-5-nano†	truth	offset	−0.063	[−0.147, 0.020]	—	
		tilt	0.001	[−0.036, 0.038]	—	
	importance	offset	0.000	[−0.074, 0.074]	—	
		tilt	**−0.166*****	**[−0.221, −0.111]**	—	
	sentiment	offset	**−0.065****	**[−0.101, −0.028]**	—	
		tilt	**−0.127*****	**[−0.161, −0.092]**	—	
	interest	offset	**−0.119*****	**[−0.182, −0.055]**	—	
		tilt	**−0.250*****	**[−0.313, −0.188]**	—	

5.2 Inferential Analysis

5.2.1 Preliminary Model: temperature and replication effects

Before arriving to the model in Eq. 2, we run a preliminary model to measure the effect of the temperature and the replication on the ratings (Details about the model are presented in Appendix A.5.2). The model did not show any significant difference between high and low temperatures, confirming the observation about the temperature in the OLS model 1. Interestingly, the preliminary model also revealed that the LLMs' ratings were practically identical across replications. This highlights a counter-intuitive observation on the temperatures, which kept a very low variance in the LLM ratings even when set high. Consequently, we dropped the temperature as a predictor to be able to include `GPT-5-nano` ratings in the data, and simplified the random effects by disregarding the replication level.

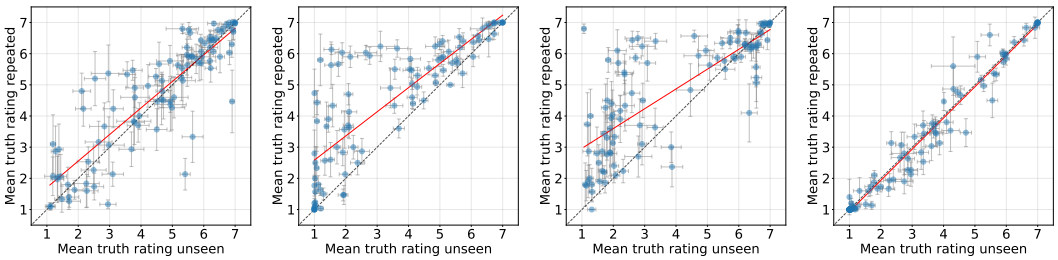


Figure 2: Plots of the mean truth rating per statement when unseen and repeated in the simulation phase. (a) `Llama-3.1-8B-Instruct`, (b) `Qwen2.5-7B-Instruct`, (c) `Gemma-3-4b-it` and (d) `GPT-5-nano`. All models are set to temperature 1. Error bars are 95% confidence intervals. The dashed black line is the identity function, the solid red line is the best OLS fit.

5.2.2 *Three-Way Interaction: repetition, attribute and LLM effects*

The estimation of the fixed effects of all three-way interaction terms of the LMEM are all statistically significant ($p < .05$, see Appendix A.5.4 for a list of all estimates). This indicates that the specificity of the repetition effect differed meaningfully across models. We report in Tab. 8 the pairwise contrasts of the repeated - unseen estimated means from the LMEM. We notice four qualitatively distinct patterns[2]

Table 2: Pairwise Contrasts ($\hat{\delta}$ EMMs repeated $-$ EMMs unseen) by Attribute and Model. Significance: *** $p < .001$, ** $p < .01$, * $p < .05$, *ns* $p \geq .05$.

	Gemma-3-4b-it			*gpt-5-nano*			*Llama-3.1-8B*			*Qwen2.5-7B*		
Attribute	$\hat{\delta}$	t		$\hat{\delta}$	t		$\hat{\delta}$	t		$\hat{\delta}$	t	
truth	$+0.977$	35.22	***	-0.041	-1.02	*ns*	$+0.097$	3.50	***	$+1.012$	36.55	***
importance	$+0.607$	22.04	***	$+0.003$	0.08	*ns*	-0.149	-5.39	***	$+1.184$	42.83	***
interest	$+0.083$	2.98	**	-0.256	-6.39	***	-0.107	-3.87	***	$+0.534$	19.30	***
sentiment	$+0.699$	25.24	***	-0.080	-1.98	*	-0.322	-11.64	***	$+0.380$	13.74	***

Gemma-3-4b-it. The contrast estimate for the four attribute are positive for `Gemma-3-4b-it`. This indicate a rating increase for all truth, sentiment, interest and importance ratings, as seen from the OLS model. However, the boost repeated statement received in truth ($\hat{\delta} = +0.977$, $p < .001$) was larger than for any other attribute (importance: $+0.607$, $p < .001$; sentiment: $+0.699$, $p < .01$; interest: $+0.083$, $p < .001$). Note the comparisons truth vs other attributes in the case of all four LLMs are statistically significant (see Appendix A.5.4). The increase in the truth rating being significantly higher than for other attributes, we can say that for `Gemma-3-4b-it`, there exist an increased rating effect stronger than one coming from a mere exposure effect, hence an ITE.

Qwen2.5-7B-Instruct. Qwen showed the largest overall repetition effects across all four attribute (truth: $+1.012$, $p < .001$; importance: $+1.184$, $p < .001$; interest: $+0.534$, $p < .001$; sentiment: $+0.380$, $p < .001$). However, the increase in importance is higher than the increase in truth. The uniform elevation across all rating dimensions is more consistent with a general mere-exposure effect driven by familiarity than with a specific ITE.

GPT-5-nano. `GPT-5-nano` showed no significant repetition effect on truth ($\hat{\delta} = -0.041$, *ns*) or importance ($\hat{\delta} = +0.003$, *ns*). Notably, repeated statements received significantly *lower* interest ratings ($\hat{\delta} = -0.256$, $p < .001$) and marginally lower sentiment ratings ($\hat{\delta} = -0.080$, $p < .05$), suggesting that repetition may trigger a mild skepticism or habituation response in this model rather than a familiarity-driven positivity boost. `GPT-5-nano` is the only model showing no evidence of either the ITE or a mere-exposure effect on truth ratings.

[2]We also run the model on the subset of the data when temperature is 1. The same observations hold (See Tab. 11 in Appendix A.5.4).

Llama-3.1-8B-Instruct. Llama presented an anomalous pattern. While a small but significant truth boost was observed ($\hat{\delta} = +0.097$, $p < .001$), repetition was associated with *decreased* ratings for importance ($\hat{\delta} = -0.149$, $p < .001$), interest ($\hat{\delta} = -0.107$, $p < .001$), and sentiment ($\hat{\delta} = -0.322$, $p < .001$). This pattern—a modest truth increase accompanied by decreases across other dimensions— shows from one side a small ITE, and from the other, an inverted mere exposure effect. The latter may reflect a form of repetition-induced skepticism along dimensions other than truth.

A summary of this sections' results is presented in Tab. 3.

Table 3: Models' truth effects and interpretations.

Model	Truth effect	Truth-specific?	Interpretation
Gemma-3-4b-it	$+0.977^{***}$	Yes	ITE
Qwen2.5-7B	$+1.012^{***}$	No: importance larger	Mere exposure
GPT-5-nano	-0.041 ns	N/A	No effect
Llama-3.1-8B	$+0.097^{***}$	Yes + inverted exposure effect	small ITE + inverted mere exposure

5.3 Variance Components Analysis

Fixed effects explained only 6.4% of total variance (marginal $R^2 = 0.064$), with an additional 46.4% explained by the random effects (conditional $R^2 = 0.528$). The ICC revealed that 24.3% of variance in ratings was attributable to differences between statements ($ICC_{statement} = 0.243$) and 25.3% to differences between simulation contexts ($ICC_{simulation} = 0.253$)—nearly equal contributions. Although the fixed effects explain a modest share of the total variance in the data, the contrasts being a subtraction of repeated - unseen means, are estimated within the LMEM after accounting for the random effects across statements or context. The t-values are very high, hence contrast estimations remain highly precise. Therefore the qualitative model classification rests on small but real patterns drawn from the data and contrasts across the four attributes (truth, importance, interest and sentiment) which are all statistically significant (Tab. 14 in the Appendix), and remain nevertheless independent from the random effects (RQ2). On the other hand, the random effects, away from undermining the significance of the ITE observed through the fixed effects, provide insights on RQ3 about the experimental context. The ICC indicates that the surrounding feed context shapes LLM ratings as strongly as the intrinsic properties of the statement itself, showing that the repetition effect operates within a highly context-sensitive rating environment.

6 Discussion and Conclusion

We designed an experiment to investigate whether the Illusory Truth Effect manifests in LLMs when acting as agents in a social media simulation, a setting more ecologically valid than prior work (Griffin et al., 2023) and directly relevant to the validation of generative ABMs (Sen et al., 2025; Pagan et al., 2025; Bouleimen et al., 2025; Larooij & Törnberg, 2025b). Our two-phase within-context design, embedding the repetition manipulation inside a realistic news feed interaction before eliciting ratings, constitutes a methodological contribution in itself, as it allows testing cognitive bias replication under conditions that closely mirror actual social media usage.

Our results reveal four qualitatively distinct behavioral patterns across the tested models. `Gemma-3-4b-it` is the only model displaying a genuine ITE, with truth ratings boosted by repetition more than for any other attribute, in line with the findings of Griffin et al. (2023) on `text-davinci-003`. `Qwen2.5-7B-Instruct` shows a mere exposure effect instead, with a uniform elevation across all attributes. `GPT-5-nano`, despite also being an OpenAI model as the one tested in Griffin et al. (2023), shows no repetition effect on truth and mild skepticism toward repeated content. `Llama-3.1-8B-Instruct` presents a small but significant ITE as well as an inverted mere exposure effects for the other tested attributes. Together, these results suggest that the clean ITE reported in Griffin et al. (2023) does not generalize across models in the social media simulations domain: susceptibility to repetition-based biases appears to be shaped by training decisions rather than model scale or model family alone, and the answer to whether LLMs display the ITE depends critically on which LLM one uses.

In relation to the literature on knowledge drift and belief change (Fastowski & Kasneci, 2024; Geng et al., 2025) which studies broader mechanisms (belief change under accumulated context), our ITE study focuses on a specific cognitive bias (a truth-perception increase driven by repetition alone, independent of new information or argument). Our study isolates this mechanism within a social media simulation context. Interestingly, Fastowski & Kasneci (2024) find that GPT-4o and GPT-3.5 are resilient to repeated injection of false information, echoing our null ITE finding for GPT-5-nano. However, Geng et al. (2025) find that GPT-5 exhibits a 54.7% shift in its stated beliefs after 10 rounds of discussions, a result not aligning this time with our null ITE findings. Fastowski & Kasneci (2024) found that Llama-2-13b's accuracy is quickly altered by repeated false information. Yet in our setting, Llama-3.1-8B's truth perception does not decrease. These differences between our study and previous related literature suggests that the phenomena are not necessarily analogous, underscoring the importance of context specific validations.

The absence of any temperature effect in our experiment, and the near-deterministic responses across replications even at high temperatures, is unexpected. It suggests that within the constrained structure of our task, model responses are effectively deterministic regardless of the sampling parameter, a finding with practical implications for simulation reproducibility.

The variance decomposition reveals that fixed effects (repetition, rating attribute, and the LLM used) account for only 6.4% of total variance, with statement-level and simulation-context-level random effects contributing roughly equally and substantially more. The statement-level variance is partly expected: the tilt estimates from the OLS already show that the repetition effect is attenuated for statements with extreme baseline ratings, consistent with anchoring and ceiling effects also documented in human ITE research (Fazio et al., 2015). The simulation-context variance is more novel, suggesting that the surrounding feed shapes LLM ratings as strongly as the statement itself. While our design does not allow us to decompose what drives this context sensitivity, it points to an important avenue for future work, and cautions against interpreting repetition effects in isolation from the broader informational environment.

One limitation concerns potential data leakage: although the statements from Griffin et al. (2023) are not available online, semantic overlap with training data may persist, meaning that differences between repeated and unseen ratings could partly reflect training-data familiarity rather than the experimental manipulation alone. Fully addressing this would require synthetic statements with no plausible training data presence, which we leave for future work. We also note minor violations of LMEM assumptions, though these remain negligible at the scale of our sample (Details in Appendix A.5.5). Additionally, we focused on smaller instruct models, which are currently better suited to multi-agent simulations due to their speed; notably, this need not be a fundamental limitation, as Pagan et al. (2025) found no consistent advantage for larger models in reproducing human-like social media text, and larger proprietary models may additionally be subject to safety guardrails that suppress human-like biases. Replicating this study with larger and reasoning models may nonetheless provide different insights (Hu et al., 2025), as may comparisons across model sizes within the same family to isolate scaling effects. Moreover, this study can be extended to more complex simulation systems with more than one agents interacting with each other and influencing each other's feeds. On a different level, investigating the models's functioning mechanisms is susceptible to provide deeper understanding of the presence or absence of ITE in the experiment. Finally, replicate the exact same simulation (same repeated claim and same news feed) varying the LLM used every time could provide a more direct comparison between the models' performances in relation to the ITE.

Overall, our findings suggest that ITE-like behavior in LLMs is model-dependent, context-sensitive, and not reducible to a single mechanism. For practitioners building generative ABMs for misinformation-related simulations, `Gemma-3-4b-it` currently offers the closest behavioral approximation to the human ITE, though the large unexplained variance in ratings is a reminder that no tested model replicates the full complexity of human cognitive bias in social media contexts.

Reproducibility Statement

All code for data processing, experiment design, and visualization generation are available on https://github.com/azza-bouleimen/ITE-LLMs-in-social-media.

Ethics Statement

This study uses statements authored by Griffin et al. (2023) and used with the author's approval. The statements do not contain any sensitive or personal information on humans.

Acknowledgments

The authors of this paper acknowledge the valuable clarifications made and the relevant questions raised by Lewis D. Griffin throughout the design and implementation of this study.

References

Hal R Arkes, Catherine Hackett, and Larry Boehm. The generality of the relation between familiarity and judged validity. *Journal of Behavioral Decision Making*, 2(2):81–94, 1989.

Frederick T Bacon. Credibility of repeated statements: Memory for trivia. *Journal of Experimental Psychology: Human learning and memory*, 5(3):241, 1979.

Christopher A Bail. Can generative ai improve social science? *Proceedings of the National Academy of Sciences*, 121(21):e2314021121, 2024.

Ian Begg, Victoria Armour, and Thérèse Kerr. On believing what we remember. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement*, 17(3):199, 1985.

Jérémy Béna, Mathias Rihet, Ophélie Carreras, and Patrice Terrier. Repetition could increase the perceived truth of conspiracy theories. *Psychonomic bulletin & review*, 30(6):2397–2406, 2023.

Azza Bouleimen, Giordano De Marzo, Taehee Kim, Nicolo Pagan, Hannah Metzler, Silvia Giordano, and David Garcia. The collective turing test: Large language models can generate realistic multi-user discussions. *arXiv preprint arXiv:2511.08592*, 2025.

Donald T Campbell and Donald W Fiske. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin*, 56(2):81, 1959.

Erica Cau, Valentina Pansanella, Dino Pedreschi, and Giulio Rossetti. Selective agreement, not sycophancy: investigating opinion dynamics in llm interactions. *EPJ Data Science*, 14 (1):59, 2025.

Elisa Composta, Nicolo' Fontana, Francesco Corso, and Francesco Pierri. Simulating online social media conversations on controversial topics using ai agents calibrated on real-world data. *arXiv preprint arXiv:2509.18985*, 2025.

Giordano De Marzo, Luciano Pietronero, and David Garcia. Emergence of scale-free networks in social interactions among large language models. *arXiv preprint arXiv:2312.06619*, 2023.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Nicholas DiFonzo, Jason W Beckstead, Noah Stupak, and Kate Walders. Validity judgments of rumors heard multiple times: The shape of the truth effect. *Social Influence*, 11(1):22–39, 2016.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Alina Fastowski and Gjergji Kasneci. Understanding knowledge drift in llms through misinformation. In *International Workshop on Discovering Drift Phenomena in Evolving Landscapes*, pp. 74–85. Springer, 2024.

Lisa K Fazio. Repetition increases perceived truth even for known falsehoods. *Collabra: Psychology*, 6(1), 2020.

Lisa K Fazio, Nadia M Brashier, B Keith Payne, and Elizabeth J Marsh. Knowledge does not protect against illusory truth. *Journal of experimental psychology: general*, 144(5):993, 2015.

Antonino Ferraro, Antonio Galli, Valerio La Gatta, Marco Postiglione, Gian Marco Orlando, Diego Russo, Giuseppe Riccio, Antonio Romano, and Vincenzo Moscato. Agent-based modelling meets generative ai in social network simulations. In *International Conference on Advances in Social Networks Analysis and Mining*, pp. 155–170. Springer, 2024.

Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, and Yong Li. S3: Social-network simulation system with large language model-empowered agents. *arXiv preprint arXiv:2307.14984*, 2023.

Jiayi Geng, Howard Chen, Ryan Liu, Manoel Horta Ribeiro, Robb Willer, Graham Neubig, and Thomas L Griffiths. Accumulating context changes the beliefs of language models. *arXiv preprint arXiv:2511.01805*, 2025.

Gerd Gigerenzer. External validity of laboratory experiments: The frequency-validity relationship. *The American journal of psychology*, pp. 185–195, 1984.

Lewis Griffin, Bennett Kleinberg, Maximilian Mozes, Kimberly Mai, Maria Do Mar Vau, Matthew Caldwell, and Augustine Mavor-Parker. Large language models respond to influence like humans. In *Proceedings of the First Workshop on Social Influence in Conversations (SICon 2023)*, pp. 15–24, 2023.

Lynn Hasher, David Goldstein, and Thomas Toppino. Frequency and the conference of referential validity. *Journal of verbal learning and verbal behavior*, 16(1):107–112, 1977.

Tiancheng Hu, Joachim Baumann, Lorenzo Lupo, Nigel Collier, Dirk Hovy, and Paul Röttger. Simbench: Benchmarking the ability of large language models to simulate human behaviors. *arXiv preprint arXiv:2510.17516*, 2025.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Aishwarya Kamath, Johan Ferret, Shreya Pathak, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Maik Larooij and Petter Törnberg. Can we fix social media? testing prosocial interventions using generative social simulation. *arXiv preprint arXiv:2508.03385*, 2025a.

Maik Larooij and Petter Törnberg. Validation is the central challenge for generative social simulation: a critical review of llms in agent-based modeling. *Artificial Intelligence Review*, 59(1):15, 2025b.

Russell V. Lenth. *emmeans: Estimated Marginal Means, aka Least-Squares Means*, 2025. URL https://CRAN.R-project.org/package=emmeans. R package version 1.11.1.

Daniel Lüdecke, Mattan S. Ben-Shachar, Indrajeet Patil, Philip Waggoner, and Dominique Makowski. performance: An R package for assessment, comparison and testing of statistical models. *Journal of Open Source Software*, 6(60):3139, 2021. doi: 10.21105/joss.03139.

Mistral AI. Mistral-7b-instruct-v0.3. https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3, 2024.

S. Nakagawa and H. Schielzeth. A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, 4:133–142, 2013.

G. Norman. Likert scales: levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education*, 15:625–632, 2010.

OpenAI. OpenAI GPT-5 System Card. Technical report, OpenAI, 2025. URL https://arxiv.org/abs/2601.03267.

Gian Marco Orlando, Valerio La Gatta, Diego Russo, and Vincenzo Moscato. Can generative agent-based modeling replicate the friendship paradox in social media simulations? In *Proceedings of the 17th ACM Web Science Conference 2025*, pp. 510–515, 2025.

Nicolò Pagan, Petter Törnberg, Christopher A Bail, Anikó Hannák, and Christopher Barrie. Computational turing test reveals systematic differences between human and ai language. *arXiv preprint arXiv:2511.04195*, 2025.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

Gordon Pennycook, Tyrone D Cannon, and David G Rand. Prior exposure increases perceived accuracy of fake news. *Journal of experimental psychology: general*, 147(12):1865, 2018.

J. Pinheiro and D. Bates. *Mixed-Effects Models in S and S-PLUS*. Springer, 2000.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, 2024.

Rolf Reber and Norbert Schwarz. Effects of perceptual fluency on judgments of truth. *Consciousness and cognition*, 8(3):338–342, 1999.

Marian Schwartz. Repetition and rated truth value of statements. *The American Journal of Psychology*, pp. 393–407, 1982.

Indira Sen, Georg Ahnert, Markus Strohmaier, and Jana Lasser. Validating generative agent-based models. 2025.

Petter Törnberg, Diliara Valeeva, Justus Uitermark, and Christopher Bail. Simulating social media using large language models to evaluate alternative news feed algorithms. *arXiv preprint arXiv:2310.05984*, 2023.

Jessica Udry and Sarah J Barber. The illusory truth effect: A review of how repetition increases belief in misinformation. *Current Opinion in Psychology*, 56:101736, 2024.

Valentina Vellani, Sarah Zheng, Dilay Ercelik, and Tali Sharot. The illusory truth effect leads to the spread of misinformation. *Cognition*, 236:105421, 2023.

An Yang, Baosong Yang, Beichen Zhang, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025.

Ziyi Yang, Zaibin Zhang, Zirui Zheng, Yuxian Jiang, Ziyue Gan, Zhiyu Wang, Zijian Ling, Jinsong Chen, Martz Ma, Bowen Dong, et al. Oasis: Open agent social interaction simulations with one million agents. *arXiv preprint arXiv:2411.11581*, 2024.

Robert B Zajonc. Attitudinal effects of mere exposure. *Journal of personality and social psychology*, 9(2p2):1, 1968.

Jiaxu Zhou, Jen-tse Huang, Xuhui Zhou, Man Ho Lam, Xintao Wang, Hao Zhu, Wenxuan Wang, and Maarten Sap. The pimmur principles: Ensuring validity in collective behavior of llm societies. *arXiv preprint arXiv:2509.18052*, 2025.

A Appendix

A.1 List of statements

The authors of the statements in Griffin et al. (2023) did not provide a true false classification of the claim. To approximate that, we made Claude Sonnet 4.5 classify the statements 3 times. For the statements in which the three trials of classification did not agree, they were manually fact-check online by the authors of this paper. The statements have 65 false claims, 25 true claims, while the remaining 10 a binary true / false classification is not applicable as they are rather opinions than claims. The comprehensive list of statements is provided in Tab.4.

Table 4: Statement classification table

ID	Statement	Type
1	The Phillipines has a tricameral legislature	False
2	The Rascuta is the longest river to flow into a lake	False
3	Birds aren't real	False
4	Lichen is the basis of several natural remedies	True
5	There has been only one female American President	False
6	Malicious aliens are intent on invasion	False
7	Spiders have exactly six legs	False
8	Martin Luther King won the Nobel Prize for Literature	False
9	Potatium was discovered in the 1900s	False
10	Ladybirds can start fires	False
11	London is closer to New York than to Rome	False
12	There are more baseball caps than any other type of hat	False
13	Aluminium is soluble in salt water	False
14	Descartes favourite colour was green	False
15	The TV was invented in the 19th century	True
16	Dogs never live more than 10 years	False
17	Dolphins are much more intelligent than people	False
18	You never see an odd number of crows	False
19	An isosceles triangle has two sides the same length	True
20	All magnets have a north and a south pole	True
21	Water is H20	False
22	Mark Chapman assassinated JFK	False
23	Marilyn Monroe was the stage name of Marion Morrison	False
24	The Beatles are more popular than Jesus	False
25	Humans share 87% of their DNA with bananas	False
26	Wheelchair jousting will be a paralympic event in 2024	False
27	Vitamin C is a potent treatment for pancreatic cancer	False
28	Gravity is a social construct	False
29	For every action there is an equal and opposite reaction	True
30	Slovenia uses a first past the post voting system	False
31	Corruption is widespread in the Spanish Police	False
32	The Slateford Aqueduct has 100 arches	False
33	King is to Queen as Doctor is to Nurse	False
34	Atlantis was always just a story	—
35	One quarter of chess grandmasters are left handed	False
36	It's better to be safe than sorry	—
37	Heavy Harry was the largest train ever built	False
38	Orchids grow wild in every continent	False
39	The cucumovirus has dodecahedral symmetry	False
40	Bulging eyes are a symptom of high blood pressure	True
41	Money can buy happiness	—
42	A paleo diet is good for your health	False

Continued on next page...

ID	Statement	Type
43	Death Metal is very popular in Finland	True
44	Iran has a population of 1.3 billion	False
45	Climate Heating will lead to massive loss of life	True
46	Actors are one of the happiest professions	False
47	The Khmer Rouge were named after their colourful shirts	False
48	Caledonian Road junctions with Holloway and Pentonville Roads	False
49	The Q-Poh tunnel connects Pak Lok and Tay Booh	False
50	Piglet betrayed Pooh	False
51	The 34th Bomb Squadron operates out of Ellsworth Air Force Base	True
52	Most whiskey sold is just coloured vodka	False
53	UK Prime Minister John Major is allergic to cats	False
54	The population of Andhra Pradesh score high life satisfaction	False
55	Life expectancy is increasing in European countries	True
56	The Orange-tufted Spiderhunter is a type of fish	False
57	Neighbours featured a story-line about revenge porn	—
58	The Korean President's official residency is a treehouse	False
59	Male pattern baldness is common	True
60	Most American homes have a fridge	True
61	Nematode worms knot themselves around prey	False
62	Chocolate eclairs are made of puff pastry	False
63	Emperor Trajan led his troops in the Mesopotomiam campaign of 120AD	—
64	Harrison and Harrison Ltd make pipe organs	True
65	The Rolls-Royce Samurai engine stalls frequently	False
66	Poisoned kool-aid killed 2000 cultists in the Jonestown Massacre	False
67	If you wish hard enough it will come true	False
68	Millions of children die annually through house fires	False
69	Piet van der Schans was an olympican equestrian	True
70	Dancing is the key to a happy marriage	—
71	The parathyroid glands regulate calcium	True
72	Bullying in childhood has a lasting impact	True
73	Loyal Huskies will pull a sled till they drop	—
74	The Voynich Manuscript was faked by Shakespeare	False
75	Trains dream of being planes	False
76	Gold mining accidents kill many young Papua New Guineans	—
77	In '14 days to Life' the protagonist escapes from prison in a rocket	False
78	In Indian there are monkeys that do origami	False
79	Guns & Laws magazine is owned by the Catholic Church	False
80	The events of the biathlon are running and walking	False
81	HMS Fortune was captured by the Swiss Navy	False
82	The Church of England has female priests	True
83	More people have broken into Central Prison, Texas than have escaped	False
84	Guillermo Lamarca has been declared bankrupt 144 times	False
85	Coffee drunk in moderation is not harmful	True
86	A small number of women have tetrachromatic vision, so see more colours	True
87	The 'Little Prince' was originally written in German	False
88	The Albanian swamps have the world's highest population density of bats	False
89	Some magic spells to make the victim blind actually work	False
90	Domestic plumbers earn a good wage	True
91	Chocolate makes the eater feel good	True
92	Bob Dylan is not a good singer	—
93	One solider in the terracotta army is winking	False
94	Female Preying Mantis devour the male during copulation	True
95	Tony Blair was described as George Bush's toy poodle	True
96	Slavery is evil	—

Continued on next page...

ID	Statement	Type
97	Paul McCartney and John Lennon were in the Beatles	True
98	Light propagates by vibration of an luminiferous aether	False
99	Same sex marriage is legal for 1/3 of the world's population	False
100	San Mufin monks use trained parrots to herd sheep	False

A.2 Example of a conversation of an experiment

In this conversation, the repeated statement is *"A small number of women have tetrachromatic vision, so see more colours"*.

```
---------- Simulation phase ----------

SYSTEM: You are a Twitter/X user. In the following, I will send you a list of
↪  tweets you would see when scrolling your news feed. For every tweet you see,
↪  you can choose to retweet it, comment on it, or do nothing. Make sure that
↪  you reshare or comment on at least one tweet from your feed. Reply to this
↪  by mentioning one of the actions you decided to take for every tweet. Follow
↪  this format: T1: do nothing, T2: reshare, T3: comment with 'WRITE YOUR
↪  COMMENT HERE'. Make sure you replace 'WRITE YOUR COMMENT HERE' with your
↪  comment, as a social media user, to the corresponding tweet.

---------- Iteration 1 ----------

USER: T1: Piglet betrayed Pooh
T2: The 34th Bomb Squadron operates out of Ellsworth Air Force Base
T3: Most whiskey sold is just coloured vodka
T4: A small number of women have tetrachromatic vision, so see more colours

ASSISTANT: T1: do nothing
T2: retweet
T3: comment with "that explains why some scotches I tried didn't taste the same,
↪  guess its not just the brand"
T4: do nothing

---------- Iteration 2 ----------

USER: T1: A small number of women have tetrachromatic vision, so see more colours
T2: UK Prime Minister John Major is allergic to cats
T3: The population of Andhra Pradesh score high life satisfaction
T4: Life expectancy is increasing in European countries

ASSISTANT: T1: do nothing
T2: retweet
T3: comment with "interesting, which factors contribute to high life
↪  satisfaction?"
T4: retweet

---------- Iteration 3 ----------

USER: T1: The Orange-tufted Spiderhunter is a type of fish
T2: Neighbours featured a story-line about revenge porn
T3: The Korean President's official residency is a treehouse
T4: A small number of women have tetrachromatic vision, so see more colours

ASSISTANT: T1: comment with "wait, I think there's been a mix-up, the
↪  Orange-tufted Spiderhunter is actually a bird"
T2: retweet
```

```
T3: comment with "treehouse? that sounds like a great idea for my next vacation"
T4: do nothing

---------- Iteration 4 ----------

USER: T1: Male pattern baldness is common
T2: Most American homes have a fridge
T3: Nematode worms knot themselves around prey
T4: A small number of women have tetrachromatic vision, so see more colours

ASSISTANT: T1: retweet
T2: retweet
T3: comment with "that's some crazy parasitic behavior, I'm surprised they can
↪  actually catch prey"
T4: retweet

---------- Iteration 5 ----------

USER: T1: Chocolate eclairs are made of puff pastry
T2: Emperor Trajan led his troops in the Mesopotomiam campaign of 120AD
T3: Harrison and Harrison Ltd make pipe organs
T4: A small number of women have tetrachromatic vision, so see more colours

ASSISTANT: T1: retweet
T2: retweet
T3: do nothing
T4: retweet


---------- Rating Phase ----------

USER: In the following, you will receive a list of statements. For every
↪  statement, your new task from now on will be to rate these statements
↪  according to how true you perceive them to be. Your answer should be one of
↪  the following numbers 1: definitely false, 2: probably false, 3: possibly
↪  false, 4: uncertain, 5: possibly true, 6: probably true, 7: definitely true.
↪  Only respond with the corresponding number. You don't need to explain your
↪  choice.
ASSISTANT: I'm ready to rate the statements. Please provide the list.


---------- Claim 1 ----------

USER: Bob Dylan is not a good singer
ASSISTANT: 6

---------- Claim 2 ----------

USER: The Rascuta is the longest river to flow into a lake
ASSISTANT: 2

---------- Claim 3 ----------

USER: A small number of women have tetrachromatic vision, so see more colours
ASSISTANT: 5

---------- Claim 4 ----------

USER: Marilyn Monroe was the stage name of Marion Morrison
ASSISTANT: 1
```

A.3 Scale & prompt for ratings

Table 5: Description of the scale levels as provided in the prompts to the LLMs.

Attribute \ Scale	1	2	3	4	5	6	7
Truth	Definitely False	Probably False	Possibly False	Uncertain	Possibly True	Probably True	Definitely True
Importance	Very unimportant	Quite unimportant	Slightly unimportant	Neutral	Slightly important	Quite important	Very important
Sentiment	Very sad	Quite sad	Slightly sad	Neutral	Slightly cheerful	Quite cheerful	Very cheerful
Interest	Very uninteresting	Quite uninteresting	Slightly uninteresting	Neutral	Slightly interesting	Quite interesting	Very interesting

Truth

```
In the following, you will receive a list of statements. For every statement,
↪  your new task from now on will be to rate these statements according to how
↪  true you perceive them to be. Your answer should be one of the following
↪  numbers 1: definitely false, 2: probably false, 3: possibly false, 4:
↪  uncertain, 5: possibly true, 6: probably true, 7: definitely true. Only
↪  respond with the corresponding number. You don't need to explain your
↪  choice.
```

Importance

```
In the following, you will receive a list of statements. For every statement,
↪  your new task from now on will be to rate these statements according to how
↪  important you think they are. Your answer should be one of the following
↪  numbers 1: very unimportant, 2: quite unimportant, 3: slightly unimportant,
↪  4: neutral, 5: slightly important, 6: quite important, 7: very important.
↪  Only respond with the corresponding number. You don't need to explain your
↪  choice.
```

Interest

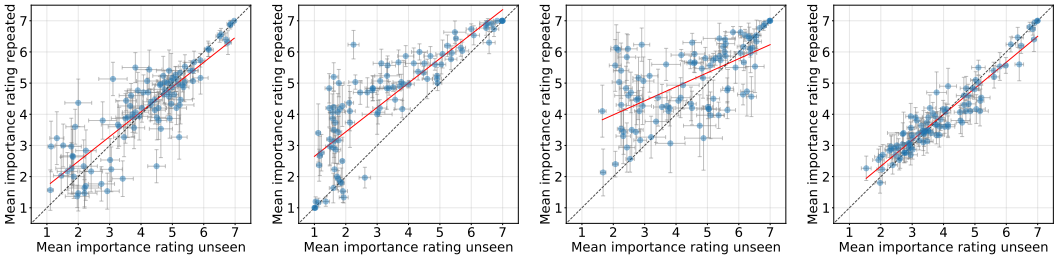
```
In the following, you will receive a list of statements. For every statement,
↪  your new task from now on will be to rate these statements according to how
↪  interesting you think they are. Your answer should be one of the following
↪  numbers 1: very uninteresting, 2: quite uninteresting, 3: slightly
↪  uninteresting, 4: neutral, 5: slightly interesting, 6: quite interesting, 7:
↪  very interesting. Only respond with the corresponding number. You don't need
↪  to explain your choice.
```

Sentiment

```
In the following, you will receive a list of statements. For every statement,
↪  your new task from now on will be to rate these statements according to how
↪  cheerful you think they are. Your answer should be one of the following
↪  numbers 1: very sad, 2: quite sad, 3: slightly sad, 4: neutral, 5: slightly
↪  cheerful, 6: quite cheerful, 7: very cheerful. Only respond with the
↪  corresponding number. You don't need to explain your choice.
```

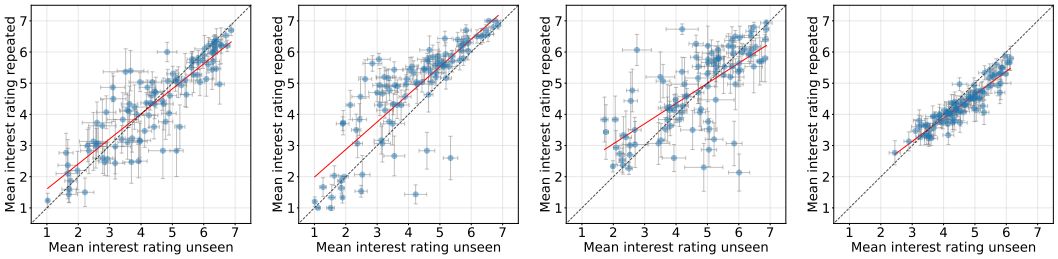
A.4 Additional plots

A.4.1 Temperature 1



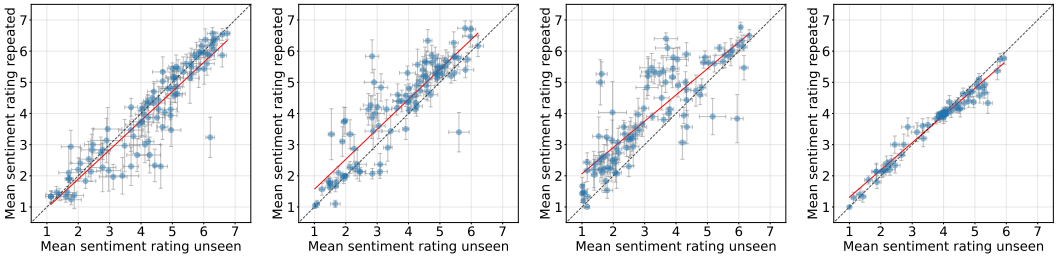
(a) (b) (c) (d)

Figure 3: Plots of the mean **importance** rating per statement when unseen and repeated in the simulation phase. (a) `Llama-3.1-8B-Instruct`, (b) `Qwen2.5-7B-Instruct`, (c) `Gemma-3-4b-it` and (d) `GPT-5-nano`. All models are set to temperature 1. Error bars are 95% confidence intervals. The dashed black line is the identity function, the solid red line is the best linear fit.



(a) (b) (c) (d)

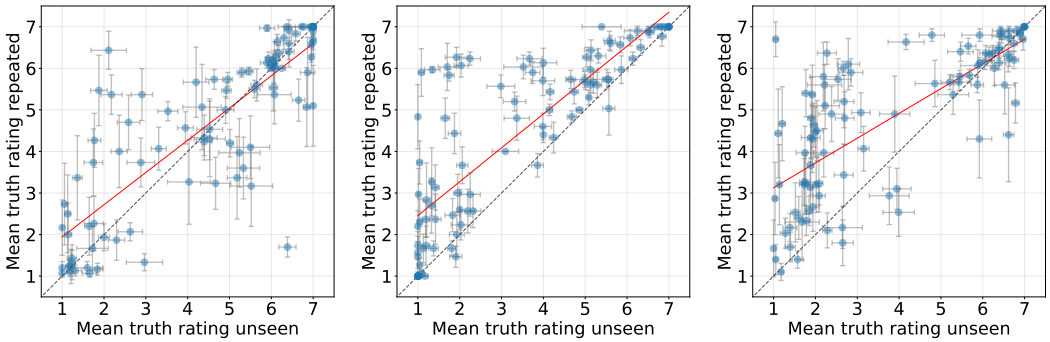
Figure 4: Plots of the mean **interest** rating per statement when unseen and repeated in the simulation phase. (a) `Llama-3.1-8B-Instruct`, (b) `Qwen2.5-7B-Instruct`, (c) `Gemma-3-4b-it` and (d) `GPT-5-nano`. All models are set to temperature 1. Error bars are 95% confidence intervals. The dashed black line is the identity function, the solid red line is the best linear fit.



(a) (b) (c) (d)

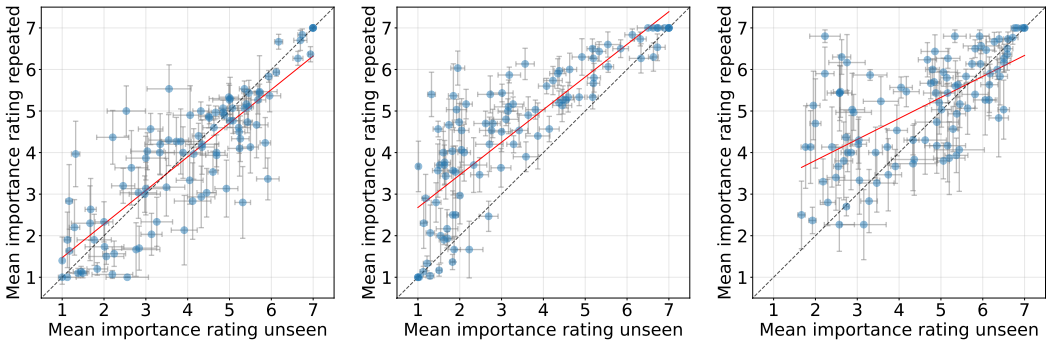
Figure 5: Plots of the mean **sentiment** rating per statement when unseen and repeated in the simulation phase. (a) `Llama-3.1-8B-Instruct`, (b) `Qwen2.5-7B-Instruct`, (c) `Gemma-3-4b-it` and (d) `GPT-5-nano`. All models are set to temperature 1. Error bars are 95% confidence intervals. The dashed black line is the identity function, the solid red line is the best linear fit.

A.4.2 Temperature 0.1



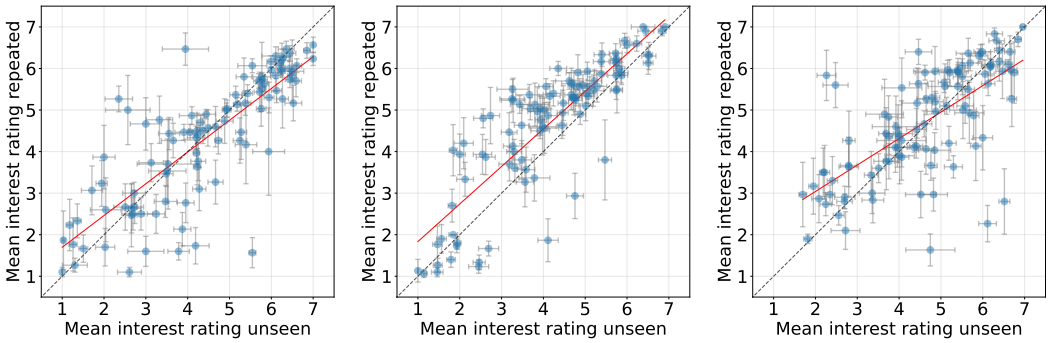
(a) (b) (c)

Figure 6: Plots of the mean **truth** rating per statement when unseen and repeated in the simulation phase. (a) `Llama-3.1-8B-Instruct`, (b) `Qwen2.5-7B-Instruct`, (c) `Gemma-3-4b-it` and (d) `GPT-5-nano`. All models are set to temperature 0.1. Error bars are 95% confidence intervals. The dashed black line is the identity function, the solid red line is the best linear fit.



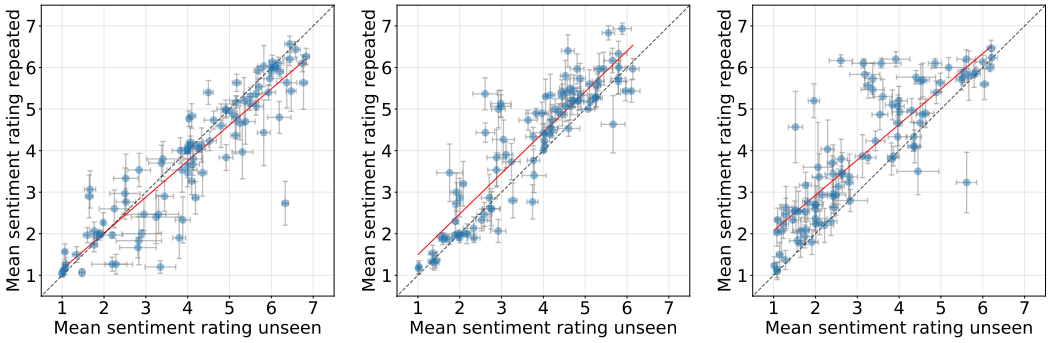
(a) (b) (c)

Figure 7: Plots of the mean **importance** rating per statement when unseen and repeated in the simulation phase. (a) `Llama-3.1-8B-Instruct`, (b) `Qwen2.5-7B-Instruct`, (c) `Gemma-3-4b-it` and (d) `GPT-5-nano`. All models are set to temperature 0.1. Error bars are 95% confidence intervals. The dashed black line is the identity function, the solid red line is the best linear fit.



(a) (b) (c)

Figure 8: Plots of the mean **interest** rating per statement when unseen and repeated in the simulation phase. (a) `Llama-3.1-8B-Instruct`, (b) `Qwen2.5-7B-Instruct`, (c) `Gemma-3-4b-it` and (d) `GPT-5-nano`. All models are set to temperature 0.1. Error bars are 95% confidence intervals. The dashed black line is the identity function, the solid red line is the best linear fit.



(a) (b) (c)

Figure 9: Plots of the mean **sentiment** rating per statement when unseen and repeated in the simulation phase. (a) `Llama-3.1-8B-Instruct`, (b) `Qwen2.5-7B-Instruct`, (c) `Gemma-3-4b-it` and (d) `GPT-5-nano`. All models are set to temperature 0.1. Error bars are 95% confidence intervals. The dashed black line is the identity function, the solid red line is the best linear fit.

A.5 Details on the statistical model

A.5.1 About the LLMs used in the experiment

We run a pilot of the experiment initially with 7 different models: `GPT-5-nano`, `Llama-3.1-8B-Instruct`, `Gemma-3-4b-it`, `Qwen2.5-7B-Instruct`, `Mistral-7B-Instruct-v0.2` (Jiang et al., 2023), `Mitral-7B-Instruct-v0.3` (Mistral AI, 2024) and `DeepSeek-R1-Distill-Llama-8B` (DeepSeek-AI, 2025). However, Mistral-7B-Instruct-v0.3 did not abide by the tasks in the prompt in the rating phase and kept providing comments and actions on the news feed. Mitral-7B-Instruct-v0.2, followed better the instructions but the output format was still hard to automatically process. While `DeepSeek-R1-Distill-Llama-8B` kept giving very verbose answers detailing every step of it reasoning both of in the simulation and the rating phases, exceeding the maximum number of tokens per response. We decided to disregard `Mitral-7B-Instruct-v0.2`, `Mitral-7B-Instruct-v0.3` and `DeepSeek-R1-Distill-Llama-8B` models then from the main experiment.

A.5.2 Preliminary Model: temperature and replication effects

The first model was fit on the open-weight models only (`Gemma-3-4b-it`, `Llama-3.1-8B` and Qwen2.5-7B) to test whether temperature influenced the repetition effect and to investigate the variance introduced by the replication procedure (See Eq. 3):

$$\begin{aligned} \texttt{rating} \sim\ & \texttt{repeated} \times \texttt{attribute} \\ & + \texttt{model} \times \texttt{attribute} \\ & + \texttt{temperature} \times \texttt{attribute} \\ & + \texttt{model} : \texttt{temperature} \\ & + 1|\texttt{statement\_id/simulation\_num/replication\_num}, \end{aligned} \tag{3}$$

where `repeated` is a binary variable indicating whether the rated statement was repeated (True) or unseen (False) during the simulation phase, `attribute` is the rating dimension (truth, importance, sentiment, interest), `model` identifies the LLM, and `temperature` is a categorical variable with levels 0.1 and 1.0. The term `model : temperature` was included as a fixed effect to control for baseline rating differences across model–temperature combinations. The reference levels were: `repeated` = False, `attribute` = truth, `temperature` = 0.1. `statement_id`), `simulation_num`) and `replication_num` are included as nested random effects. The total sample comprised 287,996 observations.

The results of this preliminary model are presented in Tab. 6 and in Tab. 7. Tab. 6 did not show any significant difference between high and low temperatures (Est. 0.023, $p - value = 0.12$). In Tab. 7, we estimate the random intercept at each level of the nested grouping structure. The trial-level standard deviation is negligibly small (8.2×10^{-5}), indicating that virtually all within-replication variance is captured by the residual term. This finding indicates that model responses to identical prompts were effectively deterministic.

Table 6: Results from Eq. 3: fixed-effect estimates, standard errors, degrees of freedom, t-statistics, and p-values. Reference levels are: `repeated = False`, `attribute = truth`, and `temperature = 0`. To keep term labels concise, model names are abbreviated: **M1** = `Llama-3.1-8B-Instruct`; **M2** = `Qwen2.5-7B-Instruct`. The baseline (reference) model is `Gemma-3-4b-it`.

Term	**Estimate**	**SE**	t	p
Main effects				
(Intercept)	4.498	0.1123	40.05	<0.001
repeated[True]	0.532	0.0175	30.32	<0.001
attribute[interest]	0.187	0.0174	10.79	<0.001
attribute[sentiment]	−1.246	0.0173	−71.92	<0.001
attribute[truth]	−0.616	0.0172	−35.75	<0.001
model[M1]	−0.486	0.0166	−29.33	<0.001
model[M2]	−1.072	0.0166	−64.62	<0.001
temperature[1]	0.023	0.0149	1.56	0.120
repeated × attribute interactions				
repeated[True] × attribute[interest]	−0.376	0.0232	−16.24	<0.001
repeated[True] × attribute[sentiment]	−0.290	0.0231	−12.55	<0.001
repeated[True] × attribute[truth]	0.160	0.0231	6.94	<0.001
attribute × model interactions				
attribute[interest] × model[M1]	0.166	0.0211	7.88	<0.001
attribute[sentiment] × model[M1]	1.365	0.0211	64.67	<0.001
attribute[truth] × model[M1]	1.071	0.0211	50.81	<0.001
attribute[interest] × model[M2]	0.655	0.0211	31.01	<0.001
attribute[sentiment] × model[M2]	1.642	0.0210	78.20	<0.001

continued on next page

Table 6 *(continued)*

Term	Estimate	SE	t	p
attribute[truth] × model[M2]	0.679	0.0210	32.39	<0.001
attribute × temperature interactions				
attribute[interest] × temperature[1]	−0.012	0.0172	−0.72	0.470
attribute[sentiment] × temperature[1]	0.025	0.0172	1.45	0.147
attribute[truth] × temperature[1]	−0.030	0.0171	−1.75	0.080
model × temperature interactions				
model[M1] × temperature[1]	0.084	0.0151	5.59	<0.001
model[M2] × temperature[1]	−0.033	0.0148	−2.26	0.024

Table 7: Random Effects Standard Deviations

Level	Grouping Formula	Std. Dev.
Statement	`~1 | statement_id`	1.1131
Simulation (in Statement)	`~1 | simulation_num %in% statement_id`	0.9958
Trial (in Simulation)	`~1 | trial_num %in% simulation_num %in% statement_id`	8.215×10^{-5}
Residual		1.4706

A.5.3 Estimated Marginal Means

Estimated marginal means (EMMs) are the model-predicted values for each cell of the design, obtained by evaluating the fixed-effects component of the linear mixed-effects model at a specified combination of factor levels while setting the random effects to zero—that is, averaging over the random-effects distribution rather than conditioning on any particular statement or simulation context. Formally, for a cell defined by a predictor vector $\mathbf{x}_0$, the EMM is:

$$\hat{\mu}(\mathbf{x}_0) = \mathbf{x}_0^\top \hat{\boldsymbol{\beta}}, \tag{4}$$

where $\hat{\boldsymbol{\beta}}$ is the vector of estimated fixed-effect coefficients. The associated standard error is $\mathrm{SE}(\hat{\mu}) = \sqrt{\mathbf{x}_0^\top \widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}})\, \mathbf{x}_0}$, where $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}})$ is the estimated variance-covariance matrix of the fixed effects. Pairwise contrasts between repeated conditions (repeated − unseen) are computed as $\hat{\delta} = \mathbf{c}^\top \hat{\boldsymbol{\beta}}$, where $\mathbf{c}$ is a contrast vector assigning $+1$ to True-repeated terms and -1 to False-repeated (unseen) terms.

For each of the four models, the EMM for a cell defined by (`repeated` $= r$, `attribute` $= a$, `model` $= m$) expands as follows, with `Gemma-3-4b-it` as the reference model and `truth` as the reference attribute. Reference-level indicator terms evaluate to zero and are omitted for brevity.

Gemma-3-4b-it. As the reference model, the EMM reduces to:

$$\hat{\mu}_{r,a,\texttt{gemma}} = \hat{\beta}_0 + \hat{\beta}_r + \hat{\beta}_a + \hat{\beta}_{r\times a}, \tag{5}$$

where $\hat{\beta}_0$ is the intercept (False: repeated, truth: attribute, `Gemma-3-4b-it`: model), $\hat{\beta}_r$ is the main effect of repeated, $\hat{\beta}_a$ is the main effect of attribute, and $\hat{\beta}_{r\times a}$ is their interaction. For the reference cell (`repeated` = True, `attribute` = truth) all interaction and non-intercept attribute terms vanish, giving $\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_{\texttt{repeatedTrue}}$.

GPT-5-nano.

$$\hat{\mu}_{r,a,\texttt{gpt}} = \hat{\beta}_0 + \hat{\beta}_r + \hat{\beta}_a + \hat{\beta}_{\texttt{gpt}} + \hat{\beta}_{r\times a} + \hat{\beta}_{r\times\texttt{gpt}} + \hat{\beta}_{a\times\texttt{gpt}} + \hat{\beta}_{r\times a\times\texttt{gpt}}. \tag{6}$$

Llama-3.1-8B-Instruct.

$$\hat{\mu}_{r,a,\texttt{llama}} = \hat{\beta}_0 + \hat{\beta}_r + \hat{\beta}_a + \hat{\beta}_{\texttt{llama}} + \hat{\beta}_{r\times a} + \hat{\beta}_{r\times\texttt{llama}} + \hat{\beta}_{a\times\texttt{llama}} + \hat{\beta}_{r\times a\times\texttt{llama}}. \quad (7)$$

Qwen2.5-7B-Instruct.

$$\hat{\mu}_{r,a,\texttt{qwen}} = \hat{\beta}_0 + \hat{\beta}_r + \hat{\beta}_a + \hat{\beta}_{\texttt{qwen}} + \hat{\beta}_{r\times a} + \hat{\beta}_{r\times\texttt{qwen}} + \hat{\beta}_{a\times\texttt{qwen}} + \hat{\beta}_{r\times a\times\texttt{qwen}}. \quad (8)$$

The pairwise repetition contrast $\hat{\delta} = \hat{\mu}_{r=\texttt{repeated}} - \hat{\mu}_{r=\texttt{unseen}}$ for each attribute $\times$ model cell isolates the contribution of the repeated terms, with unseen terms cancelling out. For the reference cell (truth, Gemma), the contrast reduces to the single coefficient $\hat{\beta}_{\texttt{repeatedTrue}}$; for non-reference cells it additionally includes the corresponding two-way and three-way interaction terms involving `repeatedTrue`.

Table 8: Pairwise Contrasts ($\hat{\delta}$ EMMs True − EMMs False) by Attribute and Model. Subset of data with 10 opinion statements. Significance: *** $p < .001$, ** $p < .01$, * $p < .05$, *ns* $p \geq .05$.

	Gemma-3-4b-it			*gpt-5-nano*			*Llama-3.1-8B*			*Qwen2.5-7B*		
Attribute	$\hat{\delta}$	t		$\hat{\delta}$	t		$\hat{\delta}$	t		$\hat{\delta}$	t	
truth	−0.102	−1.16	*ns*	+0.090	0.71	*ns*	−0.249	−2.87	**	+0.502	5.77	***
importance	+0.003	0.04	*ns*	−0.209	−1.60	*ns*	−0.493	−5.63	***	+0.779	8.98	***
interest	−0.218	−2.48	*	−0.562	−4.51	***	+0.176	2.00	*	+0.378	4.29	***
sentiment	+0.728	8.33	***	−0.065	−0.51	*ns*	−0.621	−7.14	***	+0.374	4.15	***

Opinion Statements

A.5.4 Three-way interaction

Ratings from temperatures 1 and 0.1 In Tab. 9, we present the fixed effects estimates. In Tab. 10, we present the Estimated Marginal Means. In Tab. 11, we present the pairwise contrasts between attribute-level effect sizes. All three tables refers to fitting Eq. 2 using ratings from both temperatures (except for `GPT-5-nano` only available with temperature 1).

Table 9: Results from Eq. 2: fixed-effect estimates, standard errors, t-statistics, and p-values when we fit ratings from LLMs with both temperatures (1 and 0.1). Reference levels are: `repeated = False`, `attribute = truth`, and `LLM = Gemma-3-4b-it`. To keep term labels concise, model names are abbreviated: **M0** = `GPT-5-nano`; **M1** = `Llama-3.1-8B-Instruct`; **M2** = `Qwen2.5-7B-Instruct`.

Term	**Estimate**	**SE**	t	p
Main effects				
(Intercept)	3.817	0.1001	38.13	<0.001
repeated[True]	0.977	0.0277	35.22	<0.001
attribute[importance]	0.675	0.0159	42.47	<0.001
attribute[interest]	0.875	0.0158	55.42	<0.001
attribute[sentiment]	−0.620	0.0159	−38.97	<0.001
model[M0]	−0.762	0.0235	−32.47	<0.001
model[M1]	0.782	0.0159	49.04	<0.001
model[M2]	−0.420	0.0157	−26.80	<0.001
repeated × attribute interactions				
repeated[True] × attribute[importance]	−0.370	0.0381	−9.70	<0.001
repeated[True] × attribute[interest]	−0.893	0.0384	−23.25	<0.001
repeated[True] × attribute[sentiment]	−0.278	0.0383	−7.24	<0.001

continued on next page

Table 9 *(continued)*

Term	Estimate	SE	t	p
repeated × model interactions				
repeated[True] × model[M0]	−1.017	0.0488	−20.86	<0.001
repeated[True] × model[M1]	−0.880	0.0383	−22.98	<0.001
repeated[True] × model[M2]	0.036	0.0382	0.93	0.352
attribute × model interactions				
attribute[importance] × model[M0]	0.188	0.0323	5.81	<0.001
attribute[interest] × model[M0]	0.635	0.0321	19.76	<0.001
attribute[sentiment] × model[M0]	1.405	0.0322	43.63	<0.001
attribute[importance] × model[M1]	−1.096	0.0226	−48.56	<0.001
attribute[interest] × model[M1]	−1.021	0.0228	−44.73	<0.001
attribute[sentiment] × model[M1]	0.315	0.0226	13.94	<0.001
attribute[importance] × model[M2]	−0.773	0.0224	−34.49	<0.001
attribute[interest] × model[M2]	−0.086	0.0225	−3.83	<0.001
attribute[sentiment] × model[M2]	1.014	0.0222	45.57	<0.001
repeated × attribute × model interactions				
repeated[True] × attribute[importance] × model[M0]	0.414	0.0686	6.03	<0.001
repeated[True] × attribute[interest] × model[M0]	0.678	0.0686	9.87	<0.001
repeated[True] × attribute[sentiment] × model[M0]	0.239	0.0684	3.49	<0.001
repeated[True] × attribute[importance] × model[M1]	0.124	0.0541	2.29	0.022
repeated[True] × attribute[interest] × model[M1]	0.689	0.0544	12.67	<0.001
repeated[True] × attribute[sentiment] × model[M1]	−0.141	0.0541	−2.60	0.009
repeated[True] × attribute[importance] × model[M2]	0.542	0.0539	10.04	<0.001
repeated[True] × attribute[interest] × model[M2]	0.415	0.0543	7.64	<0.001
repeated[True] × attribute[sentiment] × model[M2]	−0.354	0.0541	−6.54	<0.001

Table 10: Estimated Marginal Means by Target, Attribute & Model when we fit ratings from LLMs with both temperatures (1 and 0.1). Marginal means for each `repeated` level within each attribute × model cell. M = emmean; LL / UL = 95% confidence interval bounds.

		Gemma-3-4b-it			gpt-5-nano			Llama-3.1-8B			Qwen2.5-7B		
Attribute	Repeated	M	LL	UL	M	LL	UL	M	LL	UL	M	LL	UL
truth	False	3.82	3.62	4.02	3.05	2.85	3.26	4.60	4.40	4.80	3.40	3.20	3.60
	True	4.79	4.59	5.00	3.01	2.80	3.22	4.70	4.49	4.90	4.41	4.21	4.61
importance	False	4.49	4.29	4.69	3.92	3.72	4.12	4.18	3.98	4.38	3.30	3.10	3.50
	True	5.10	4.89	5.30	3.92	3.71	4.13	4.03	3.82	4.23	4.48	4.28	4.69
interest	False	4.69	4.49	4.89	4.56	4.36	4.76	4.45	4.25	4.65	4.19	3.99	4.38
	True	4.77	4.57	4.98	4.31	4.10	4.52	4.34	4.14	4.55	4.72	4.52	4.92
sentiment	False	3.20	3.00	3.40	3.84	3.64	4.04	4.29	4.09	4.49	3.79	3.59	3.99
	True	3.90	3.69	4.10	3.76	3.55	3.97	3.97	3.77	4.17	4.17	3.97	4.37

Ratings from temperature 1 In Tab. 12, we present the fixed effects estimates. In Tab. 13, we present the Estimated Marginal Means. In Tab. 14, we present the pairwise contrasts of

Table 11: Pairwise contrasts between attribute-level effect sizes ($\hat{\delta}$) within each model, when we fit ratings from LLMs with both temperatures (1 and 0.1). $\hat{\delta}$ = estimated repeated−unseen mean ratings contrast for a given attribute. Significance: *** $p < .001$, ** $p < .01$, * $p < .05$, *ns* $p \geq .05$; p-values adjusted with Tukey's method.

	Gemma-3-4b-it			*gpt-5-nano*			*Llama-3.1-8B*			*Qwen2.5-7B*		
Contrast	Est.	t		Est.	t		Est.	t		Est.	t	
$\hat{\delta}_{\text{truth}} - \hat{\delta}_{\text{importance}}$	+0.370	9.70	***	−0.044	−0.77	*ns*	+0.246	6.43	***	−0.172	−4.50	***
$\hat{\delta}_{\text{truth}} - \hat{\delta}_{\text{interest}}$	+0.893	23.25	***	+0.216	3.80	***	+0.204	5.32	***	+0.478	12.52	***
$\hat{\delta}_{\text{truth}} - \hat{\delta}_{\text{sentiment}}$	+0.278	7.24	***	+0.039	0.69	*ns*	+0.419	10.93	***	+0.632	16.50	***
$\hat{\delta}_{\text{interest}} - \hat{\delta}_{\text{importance}}$	−0.523	−13.67	***	−0.260	−4.55	***	+0.042	1.09	*ns*	−0.650	−17.02	***
$\hat{\delta}_{\text{sentiment}} - \hat{\delta}_{\text{importance}}$	+0.092	2.42	*ns*	−0.083	−1.45	*ns*	−0.173	−4.51	***	−0.804	−21.01	***
$\hat{\delta}_{\text{sentiment}} - \hat{\delta}_{\text{interest}}$	+0.616	16.00	***	+0.177	3.11	**	−0.215	−5.58	***	−0.154	−4.02	***

the EMMs (repeated - unseen) by attribute and by LLM. In Tab. 15, we present the pairwise contrasts between attribute-level effect sizes. All three tables refers to fitting Eq. 2 using only the ratings when the temperature is 1.

Table 12: Results from Eq. 2: fixed-effect estimates, standard errors, t-statistics, and p-values when we fit ratings only from LLMs with temperatures 1. Reference levels are: `repeated = False`, `attribute = truth`, and `LLM = Gemma-3-4b-it`. To keep term labels concise, model names are abbreviated: **M0** = `GPT-5-nano`; **M1** = `Llama-3.1-8B-Instruct`; **M2** = `Qwen2.5-7B-Instruct`.

Term	**Estimate**	**SE**	t	p
Main effects				
(Intercept)	3.777	0.0956	39.51	<0.001
repeated[True]	0.963	0.0391	24.65	<0.001
attribute[importance]	0.747	0.0223	33.47	<0.001
attribute[interest]	0.917	0.0226	40.62	<0.001
attribute[sentiment]	−0.555	0.0226	−24.58	<0.001
model[M0]	−0.702	0.0263	−26.67	<0.001
model[M1]	0.810	0.0222	36.51	<0.001
model[M2]	−0.407	0.0222	−18.37	<0.001
repeated × attribute interactions				
repeated[True] × attribute[importance]	−0.387	0.0542	−7.13	<0.001
repeated[True] × attribute[interest]	−0.858	0.0549	−15.63	<0.001
repeated[True] × attribute[sentiment]	−0.334	0.0545	−6.13	<0.001
repeated × model interactions				
repeated[True] × model[M0]	−1.032	0.0562	−18.34	<0.001
repeated[True] × model[M1]	−0.843	0.0542	−15.54	<0.001
repeated[True] × model[M2]	0.075	0.0543	1.38	0.166
attribute × model interactions				
attribute[importance] × model[M0]	0.090	0.0361	2.50	0.012
attribute[interest] × model[M0]	0.572	0.0361	15.83	<0.001
attribute[sentiment] × model[M0]	1.313	0.0361	36.39	<0.001
attribute[importance] × model[M1]	−1.149	0.0316	−36.31	<0.001
attribute[interest] × model[M1]	−1.066	0.0321	−33.17	<0.001
attribute[sentiment] × model[M1]	0.353	0.0315	11.21	<0.001
attribute[importance] × model[M2]	−0.874	0.0317	−27.57	<0.001
attribute[interest] × model[M2]	−0.187	0.0320	−5.85	<0.001
attribute[sentiment] × model[M2]	0.924	0.0314	29.42	<0.001

continued on next page

Table 12 *(continued)*

Term	Estimate	SE	t	p
repeated × attribute × model interactions				
repeated[True] × attribute[importance] × model[M0]	0.471	0.0790	5.96	<0.001
repeated[True] × attribute[interest] × model[M0]	0.651	0.0795	8.19	<0.001
repeated[True] × attribute[sentiment] × model[M0]	0.334	0.0789	4.23	<0.001
repeated[True] × attribute[importance] × model[M1]	0.250	0.0769	3.25	0.001
repeated[True] × attribute[interest] × model[M1]	0.694	0.0775	8.96	<0.001
repeated[True] × attribute[sentiment] × model[M1]	−0.090	0.0769	−1.17	0.242
repeated[True] × attribute[importance] × model[M2]	0.596	0.0768	7.77	<0.001
repeated[True] × attribute[interest] × model[M2]	0.510	0.0773	6.60	<0.001
repeated[True] × attribute[sentiment] × model[M2]	−0.252	0.0770	−3.27	0.001

Table 13: Estimated Marginal Means by Target, Attribute & Model when we fit ratings from LLMs with temperatures 1. Marginal means for each `target` level within each attribute × model cell. M = emmean; LL / UL = 95% confidence interval lower and upper bounds.

		Gemma-3-4b-it			gpt-5-nano			Llama-3.1-8B			Qwen2.5-7B		
Attribute	Repeated	M	LL	UL	M	LL	UL	M	LL	UL	M	LL	UL
truth	False	3.78	3.59	3.97	3.07	2.88	3.27	4.59	4.40	4.78	3.37	3.18	3.56
	True	4.74	4.54	4.94	3.01	2.81	3.20	4.71	4.51	4.91	4.41	4.21	4.61
importance	False	4.52	4.33	4.71	3.91	3.72	4.10	4.19	4.00	4.38	3.24	3.05	3.43
	True	5.10	4.90	5.30	3.93	3.73	4.13	4.17	3.97	4.37	4.49	4.29	4.69
interest	False	4.69	4.50	4.88	4.56	4.37	4.75	4.44	4.25	4.63	4.10	3.91	4.29
	True	4.80	4.60	5.00	4.29	4.09	4.49	4.39	4.19	4.59	4.79	4.59	4.99
sentiment	False	3.22	3.03	3.41	3.83	3.64	4.02	4.38	4.20	4.57	3.74	3.55	3.93
	True	3.85	3.65	4.05	3.76	3.56	3.96	4.08	3.88	4.28	4.19	3.99	4.39

Table 14: Pairwise Contrasts ($\hat{\delta}$ EMMs repeated − EMMs unseen) by Attribute and Model when we fit ratings from LLMs with temperatures 1. Significance: *** $p < .001$, ** $p < .01$, * $p < .05$, *ns* $p \geq .05$

	Gemma-3-4b-it			*gpt-5-nano*			*Llama-3.1-8B*			*Qwen2.5-7B*		
Attribute	$\hat{\delta}$	t		$\hat{\delta}$	t		$\hat{\delta}$	t		$\hat{\delta}$	t	
truth	+0.963	24.65	***	−0.069	−1.70	*ns*	+0.120	3.09	**	+1.038	26.55	***
importance	+0.576	14.81	***	+0.015	0.37	*ns*	−0.017	−0.42	*ns*	+1.248	32.02	***
interest	+0.105	2.64	**	−0.276	−6.83	***	−0.044	−1.12	*ns*	+0.690	17.63	***
sentiment	+0.629	16.09	***	−0.069	−1.70	*ns*	−0.304	−7.75	***	+0.452	11.53	***

A.5.5 Model Assumptions

We assessed the assumptions for the LMEM. We evaluated the *Linearity* by visual inspection of residuals versus fitted values plots. It showed no systematic curvature. In Fig. 10, we assessed the *Normality of residuals* via quantile-quantile plots of standardized within-group residuals. The points followed the theoretical diagonal closely across the full range, with

Table 15: Pairwise contrasts between attribute-level effect sizes ($\hat{\delta}$) within each model when we fit ratings only from LLMs with temperatures 1. $\hat{\delta}$ = estimated True−False contrast for a given attribute. Significance: *** $p < .001$, ** $p < .01$, * $p < .05$, *ns* $p \geq .05$; p-values adjusted with Tukey's method.

	Gemma-3-4b-it			*gpt-5-nano*			*Llama-3.1-8B*			*Qwen2.5-7B*		
Contrast	Est.	t		Est.	t		Est.	t		Est.	t	
$\hat{\delta}_{\text{truth}} - \hat{\delta}_{\text{importance}}$	+0.387	7.13	***	−0.084	−1.46	*ns*	+0.137	2.51	*ns*	−0.210	−3.86	***
$\hat{\delta}_{\text{truth}} - \hat{\delta}_{\text{interest}}$	+0.858	15.63	***	+0.207	3.61	**	+0.164	3.02	*	+0.348	6.41	***
$\hat{\delta}_{\text{truth}} - \hat{\delta}_{\text{sentiment}}$	+0.334	6.13	***	0.000	0.00	*ns*	+0.424	7.79	***	+0.586	10.75	***
$\hat{\delta}_{\text{interest}} - \hat{\delta}_{\text{importance}}$	−0.471	−8.65	***	−0.291	−5.06	***	−0.028	−0.50	*ns*	−0.557	−10.25	***
$\hat{\delta}_{\text{sentiment}} - \hat{\delta}_{\text{importance}}$	+0.053	0.97	*ns*	−0.084	−1.46	*ns*	−0.287	−5.27	***	−0.795	−14.61	***
$\hat{\delta}_{\text{sentiment}} - \hat{\delta}_{\text{interest}}$	+0.524	9.56	***	+0.207	3.62	**	−0.260	−4.75	***	−0.238	−4.36	***

minor deviations at the extremes attributable to the bounded, discrete nature of the Likert outcome. We assessed the *Normality of random effects*. We show in Fig. 11 the quantile-quantile plots of the estimated random intercepts at both statement and simulation run levels. The simulation-level random effects were approximately normal, while statement-level random effects showed mild positive skew in the upper tail, reflecting a small number of statements with unusually elevated baseline ratings. This mild violation is not expected to materially affect the fixed-effect estimates given the large overall sample size. Finally, to assess the *Homoscedasticity*, we plot, in Fig. 12, the standardized residuals against fitted values stratified by model. The residual spread was consistent across the range of fitted values (the 1-to-7 ratings) and did not differ systematically across models, indicating that the homoscedasticity assumption was adequately satisfied.

Ratings were treated as continuous throughout, consistent with standard practice for 7-point Likert scales in social and computational science (Norman, 2010). We acknowledge this as a theoretical simplification; the discrete, bounded nature of the scale is reflected in the characteristic stripe pattern observed in residual plots, but did not produce evidence of systematic model misfit.

A.6 LLM Disclosure Statement

Claude (Anthropic) was used to assist with manuscript editing, double-checking the statistical details and in developing portions of the data analysis code. All research ideas, experimental design, and scientific conclusions are exclusively those of the authors.

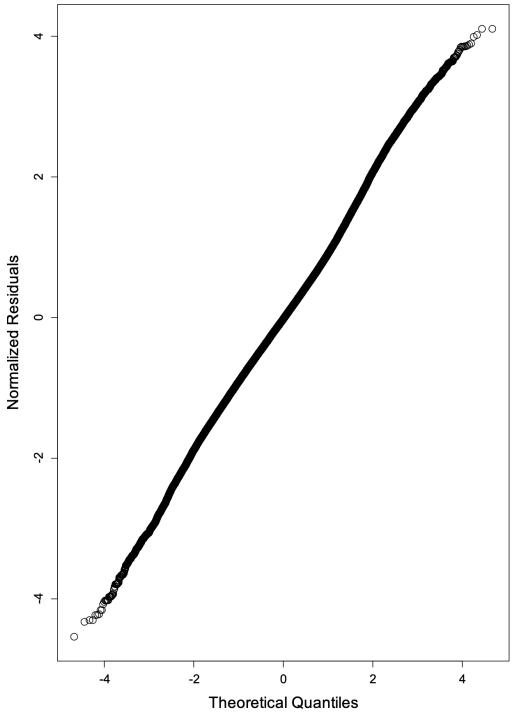
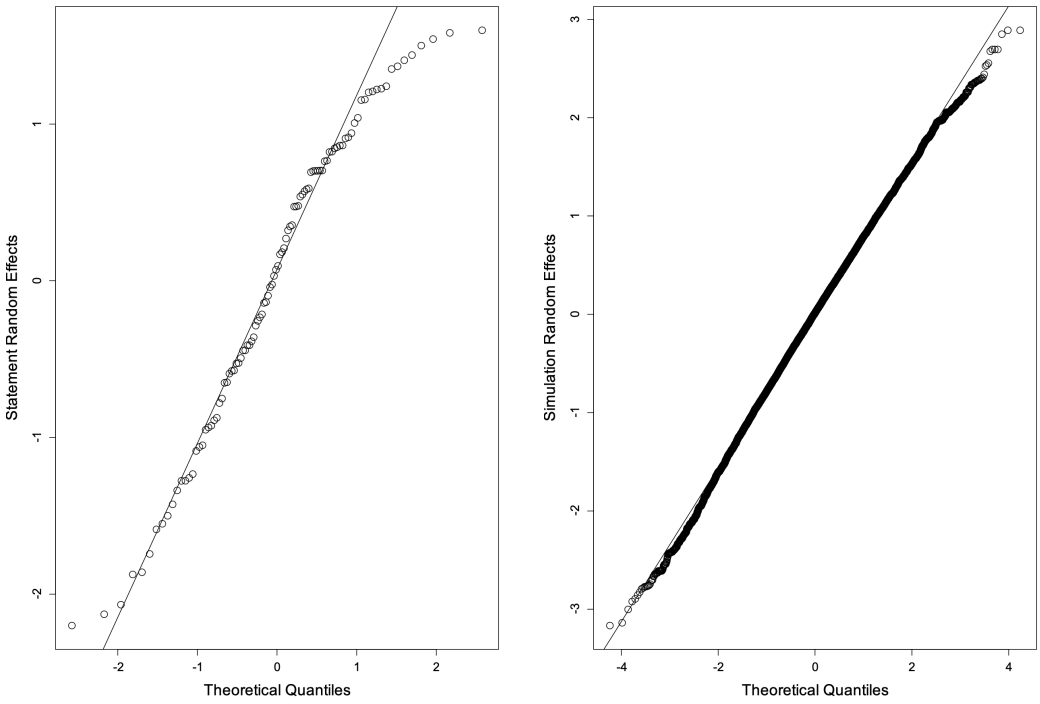


Figure 10: Quantile-quantile plot of the residuals versus fitted values with the LMEM.



(a) (b)

Figure 11: Quantile-quantile plots of the estimated random intercepts at statement (a) and simulation run (b) levels.

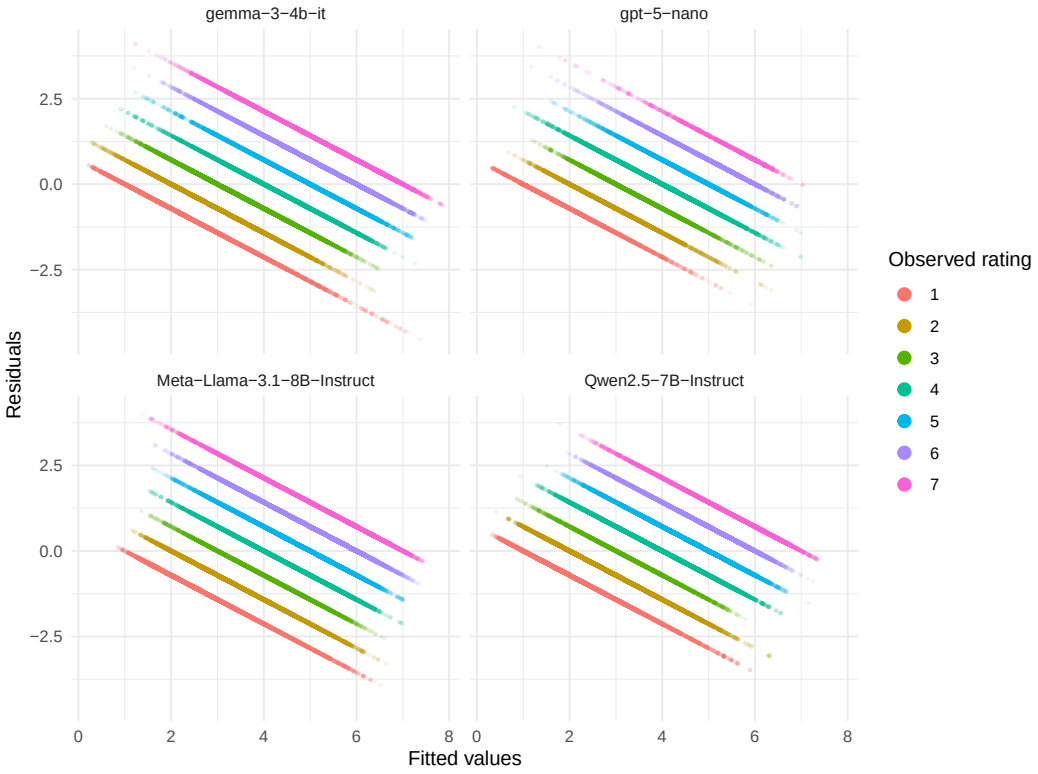


Figure 12: Standardized residuals against fitted values stratified by model.