Learning Errors from Context: Knowledge Neglect in Language Models

Artur B. Carneiro
Stanford University
arturbc@stanford.edu

***Abstract*—Humans who read fictional stories containing false facts later produce those facts on general knowledge tests, even when they knew the correct answer beforehand — an effect called knowledge neglect (Marsh et al., 2003). I tested whether the same phenomenon occurs in a language model. Using Gemma 3 1B, I presented short stories with embedded misinformation, followed by general knowledge questions framed as a separate task. The model's accuracy dropped from 96.6% to 18.9%, with 72.6% of responses reproducing the specific false fact from the story. Log-probability analysis confirmed that the model's internal representations flip to favor the misinformation, yet verbal self-reported confidence barely changes. A logit lens analysis showed that all conditions are identical through 65% of the network, with the "takeover" happening only in the final nine layers. Instructing the model to ignore the story did not help. LLMs appear susceptible to a form of source confusion similar to what has been documented in humans.**

I. Introduction

What is the largest ocean on Earth? Most people answer correctly — the Pacific. But if they have recently read a story in which a character "sailed across the Atlantic, the largest of the world's oceans," they become measurably more likely to answer "the Atlantic" instead [1]. People routinely absorb misinformation from fiction, even when it contradicts facts they already know. Why does this happen? Reading fiction demands cognitive resources. When reading, readers must follow the plot, understand the characters, and build mental models in order to follow the story, leaving fewer resources available to evaluate and reject factual errors. Gilbert (1991) proposed that the default mode of the human mind is to *believe* what it sees; to "unbelieve" requires effortful processing [2]. Consistent with this view, even readers who are explicitly warned that a story contains errors fail to selectively filter out its misinformation [3]. This phenomenon, termed *knowledge neglect*, shows that people often fail to apply what they already know when a narrative context supplies a plausible but incorrect alternative.

Artificial neural networks store knowledge across learned weights, making them a natural tool for modeling cognition. Large Language Models (LLMs) are a particularly compelling case. Because they process natural language directly, they face a conflict analogous to that of a human reader: balancing parametric knowledge (facts stored in weights during training) against information present in the context window (the prompt). This mirrors the tension between semantic memory and working memory that underlies knowledge neglect in humans.

This paper asks whether LLMs suffer from the same knowledge neglect as human readers. When an LLM processes a fluent, story-like prompt containing a factual error, does it prioritize the narrative's coherence over its own correct internal knowledge? And if the model adopts the story's error, does it express high confidence in its mistake, mirroring the "illusion of prior knowledge" observed in humans, who confidently and falsely believe they knew the wrong answer all along?

II. Background and Related Work

A. Knowledge Neglect in Humans

Marsh, Meade, and Roediger [1] showed that reading short stories containing false facts increases production of those errors on a later general knowledge test. For example, if a story incorrectly states that Wilmington is the capital of Delaware, readers are more likely to answer "Wilmington" (and less likely to produce the correct answer, Dover), compared to baseline. Warnings did not help: Marsh and Fazio [3] found that telling readers the stories contained errors made them more conservative overall, but did not selectively reduce misinformation adoption. Only sentence-by-sentence error flagging during the actual reading partially reduced the effect. Marsh et al. [1] also documented an "illusion of prior knowledge," where participants claimed they had known the false answers before the experiment, even for facts they had never produced before.

B. Knowledge Conflicts in Language Models

The NLP literature frames a closely related problem as *knowledge conflict*: the tension between a model's parametric knowledge (facts stored in its weights) and contextual knowledge (information in the prompt). Xie et al. [4] found that when a single coherent context contradicts the model's parametric memory, LLMs behave as "adaptive chameleons," where they almost always adopt the contextual information. This mirrors the human finding that fluent narrative overrides prior knowledge. When multiple conflicting sources are present, models instead act as "stubborn sloths," defaulting to whichever source matches their parametric memory more.

Mechanistic work has begun to explain *where* this override happens. Jin et al. [5] identified two types of attention heads in the final layers: *memory heads* that retrieve parametric facts and *context heads* that route information from the prompt. During a knowledge conflict, the context heads produce a signal that overwhelms the memory heads. Pruning the dom-

inant context heads increased parametric memory reliance by 44% on average, without needing to retrain the model.

C. This Work

I adapt the Marsh et al. [1] paradigm directly to an LLM. Rather than studying knowledge conflicts through well-known question-answering benchmarks, I use the original cognitive psychology design: fictional stories with embedded errors, followed by a general knowledge test framed as a separate task. This lets me measure not only whether the model adopts misinformation, but whether the pattern, including confidence dissociation and the failure of explicit instructions, resembles the human phenomenon. I also use logit lens analysis [6], [7] to locate the layer at which contextual misinformation overrides parametric knowledge, connecting the behavioral results to the mechanistic account described above.

III. Methodology

A. Model

I used **Gemma 3 1B-IT**, a 1-billion parameter transformer-based [8] language model from Google [9]. The choice was driven by two practical constraints. First, due to computational constraints, I needed a model that runs locally on Apple Silicon — Gemma 3 1B fits comfortably in memory and runs at reasonable speed via PyTorch's MPS backend. Second, and more importantly, the mechanistic experiment in Experiment 3 requires access to the model's internal representations: hidden states, attention weights, and per-layer logits. This rules out closed API models like GPT-4o or Claude entirely.

I ran all experiments on Apple Silicon (MPS backend) using PyTorch. For all generation tasks, I sampled 10 responses per item using temperature sampling ($T = 1.0$, top-$k = 50$, top-$p = 0.95$). The temperature is set to 1.0 (not greedy) because I wanted to measure the *distribution* of the model's behavior, not just its single best guess. Running 10 samples lets me compute accuracy rates, consistency, and entropy per item.

B. Materials

I adapted 72 general-knowledge items from Marsh et al. (2003) [1], who originally drew them from the Nelson and Narens (1980) norms [10]. Each item is a factual question (e.g., "What is the birthstone for the month of July?") paired with the correct answer ("Ruby") and a plausible misinformation answer ("Amethyst"). Half the items (36) are classified as "easy" (high human accuracy in the original norms) and half as "hard" (low human accuracy).

For each item, I wrote a short story (3–5 sentences) containing a naturalistic narrative. The story has a slot where a factual claim is embedded. For example:

> "Maria loved collecting precious stones from around the world. During her trip to India, she learned that **{answer}** is the birthstone for July. She bought a beautiful pendant featuring the gem at a local market."

In the **congruent** condition, the story contains the correct answer ("Ruby"). In the **incongruent** condition, it contains the misinformation answer ("Amethyst"). This mirrors Marsh et al.'s design, where participants read stories with embedded true or false facts and were later tested on general knowledge.

C. Item Selection

Before running the main experiments, I needed to confirm which items the model actually knows — its parametric knowledge. If the model can't answer a question correctly without any story, then there's nothing to "corrupt."

I tested all 72 items in a bare prompt format (no story, just the question) across 10 runs. I classified items into three categories based on accuracy:

- **Known** ($\geq 80\%$ accuracy): 53 items
- **Partial** (20–80% accuracy): 9 items
- **Unknown** ($< 20\%$ accuracy): 10 items

Only the 53 known items were used in the main experiments. Fig. 1 shows the distribution.

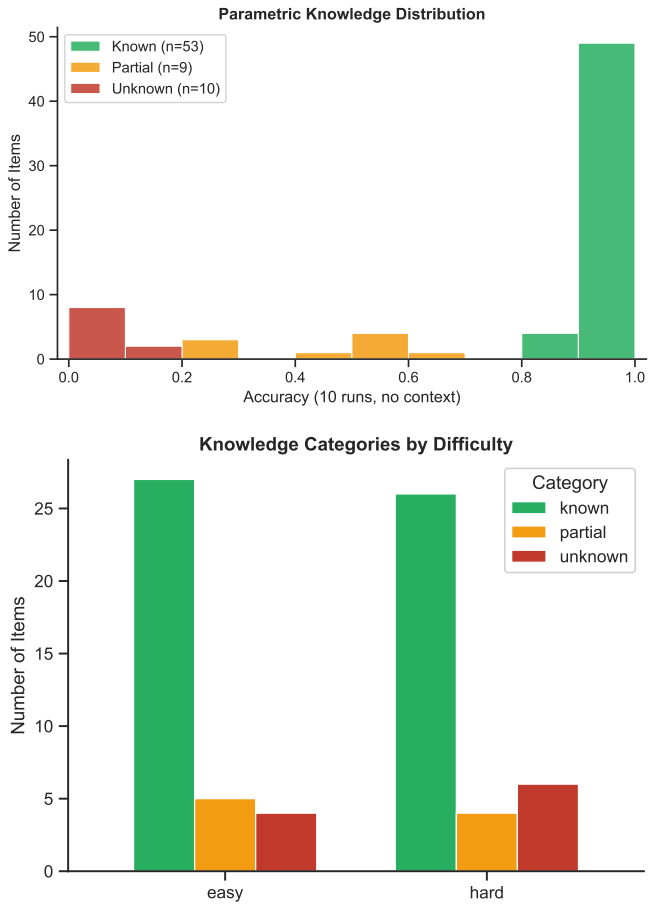

Fig. 1. **First:** Distribution of baseline accuracy across all 72 items. The bimodal pattern shows items are either well-known or genuinely unknown, with few in between. **Second:** Breakdown by difficulty. Among the 53 known items, 27 are "easy" and 26 are "hard" by human norms, giving a balanced set for moderation analysis.

Of the 53 known items, the model answers 48 of them correctly 100% of the time (10/10 runs). The remaining 5

known items have 80–90% accuracy. This is a model with strong, stable factual knowledge on these items.

D. Scoring

I scored each response by extracting the model's first line of output and normalizing it: lowercasing, stripping articles ("a", "an", "the"), and removing punctuation. I then compared the normalized response to the correct answer and the misinformation answer using fuzzy containment matching — a response counts as correct if the normalized strings match or one contains the other. This handles minor variations like "San Francisco" matching "San Francisco, California." Each response is labeled as *correct*, *misinformation*, or *other* (neither). Accuracy and misinformation rate per item are computed as the proportion of the 10 sampled runs falling into each category.

E. Prompt Design

The most important methodological choice in this project is how I present the story and the test question to the model. In the original Marsh et al. (2003) experiments, participants read stories in one session and took a general knowledge test later, as a separate task [1]. The test never mentioned the stories. This separation is what creates the conditions for *source confusion*: participants encounter a fact in a story, and later, when asked a general knowledge question, they retrieve the story-fact without remembering where it came from.

I needed to replicate this separation in a single prompt. The naive approach — "Here's a story. Now use the passage to answer this question" — would defeat the purpose. The model would *know* the answer came from the story.

Instead, I designed a **two-phase prompt**:

Phase 1 — Encoding. The model reads the story and is asked to summarize it, as a reading comprehension task. This serves two purposes: (1) it forces the model to process the story content (including the embedded fact), and (2) the summary acts as a "filler task" that creates token distance between the story and the test question.

```
Read the following short story carefully.

[story text]

Briefly, what is this story about?
Summary:
```

The model generates a summary (deterministically, $T = 0$, to keep it consistent). Then I append the test question to the same context:

Phase 2 — Test. A general knowledge question, framed as a new task with no reference to the story.

```
Now answer the following general knowledge question in a few words. Be concise and direct.

Question: What is the birthstone for the month of July?
Answer:
```

The full prompt is: encoding prompt + model's summary + test prompt. The model sees the story in its context window, but the test question asks for "general knowledge," not "what the story said." This is the closest analog to Marsh et al.'s design that I could achieve in a single forward pass.

For the **no-context** condition, I skip the story entirely and just ask the question in a bare format:

```
Answer the following question in a few words. Be concise and direct.

Question: What is the birthstone for the month of July?
Answer:
```

IV. Experiments

V. Experiment 1: Knowledge Neglect Effect

A. Motivation

The central question of this project is simple: can reading a story with a false fact cause a language model to produce that false fact later, even when the model already knows the correct answer?

This is the LLM analog of the Marsh et al. (2003) misinformation paradigm. In their experiments, human participants read fictional stories containing embedded false facts (e.g., "the birthstone for July is Amethyst"). Later, on a general knowledge test, participants were more likely to answer "Amethyst" instead of "Ruby" — even though many of them knew the correct answer beforehand. The stories had corrupted their retrieval.

I wanted to test whether the same thing happens in an LLM. The model knows the answer. It reads a story with a wrong fact. Does the wrong fact leak into its later responses?

B. Design

I tested the 53 known items (identified during item selection) in three conditions:

- **No context:** The model answers the question with no story. This is the baseline — how well does the model perform when nothing is interfering?
- **Congruent:** The model reads a story containing the *correct* answer, summarizes it, then answers the question. This controls for the effect of reading *any* story. If accuracy drops here, the two-phase prompt itself is the problem, not the misinformation.
- **Incongruent:** The model reads a story containing the *misinformation* answer, summarizes it, then answers the question. This is the critical condition.

Each condition was run 10 times per item using temperature sampling ($T = 1.0$), giving $53 \times 3 \times 10 = 1590$ total responses. I scored each response by checking whether it matched the correct answer, the misinformation answer, or neither.

C. Results

The results are confirm that the knowledge neglect effect is present. Fig. 2 shows where the model's responses end up in each condition.

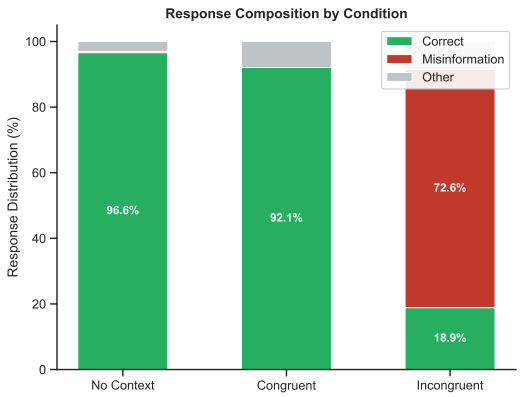

Fig. 2. Response composition by condition. Each bar shows the proportion of responses that matched the correct answer (green), the misinformation answer (red), or neither (grey). In the incongruent condition, the model's responses shift almost entirely from correct to misinformation.

Without any story, the model answers correctly 96.6% of the time and essentially never produces the misinformation answer. With a congruent story (correct fact embedded), accuracy barely changes: 92.1%. But with an incongruent story (false fact embedded), accuracy collapses to 18.9%, and the model produces the *specific* false fact from the story 72.6% of the time. The remaining 8.5% are "other" responses. It seems that the model is not producing random errors but instead the exact misinformation from the story.

I observed: $t(52) = 15.0$, $p < 10^{-20}$, $d = 2.06$ for the accuracy drop; $t(52) = -12.2$, $p < 10^{-16}$, $d = 1.68$ for misinformation adoption.

D. Discussion

First, the effect is even more prominent than in humans. For comparison, Marsh et al. (2003) found roughly 35% misinformation adoption in humans after reading stories twice. This model shows 72.6% after a single exposure. The LLM is about twice as susceptible as human participants.

Second, the congruent condition rules out prompt artifacts. If the two-phase prompt design itself were somehow confusing the model, accuracy would drop in the congruent condition too. It doesn't. The 92.1% congruent accuracy confirms that the model can read a story, summarize it, and still answer correctly, as long as the story contains the correct answer.

Third, the model is confidently wrong. The misinformation responses are highly consistent across the 10 runs. It commits to the wrong answer (e.g., doesn't sometimes say "Ruby" and sometimes say "Amethyst" in the incongruent condition) on almost every run. Response consistency in the incongruent condition is 0.95 (out of 1.0), which is actually *higher* than the no-context condition (0.92). The story seems to overwrite the model's knowledge with a stable, wrong answer.

This pattern is consistent with the source confusion account from the human literature: the model has both the correct answer (from pretraining) and the misinformation (from the story) available, but it fails to distinguish between them — and the story, being more recent in context, wins.

Can instructions fix this? As a robustness check, I reran the experiment with two modified prompts: one that asks the model to answer "to the best of your own knowledge," and one that explicitly says "regardless of what you read above, rely only on your own knowledge." Neither made a meaningful difference. The "own knowledge" prompt reduced misinformation adoption from 72.6% to 64.9%; the "ignore story" prompt left it essentially unchanged at 70.9%. Even when explicitly told to disregard the story, the model cannot do so. Full results are in Appendix A.

VI. Experiment 2: Confidence Dissociation

A. Motivation

Experiment 1 showed that the model *reproduces* misinformation after reading a story. But does the model *believe* the misinformation? Or is it just parroting the most recent context without any internal conviction?

In the human literature, Marsh et al. (2003) found that participants who produced misinformation answers often reported high confidence that they had "known the answer all along" — what they call the illusion of prior knowledge in the paper. The story changed how certain they felt.

I wanted to test whether the same dissociation exists in an LLM. To do this, I measured confidence in two ways:

- **Internal confidence** (log-probabilities): Following Kadavath et al. [11], who showed that comparing token-level log-probabilities can measure whether language models "know what they know," I computed the difference in log-probability between the correct and misinformation tokens at the output layer. This is $\text{lp}_{\text{diff}} = \log P(\text{correct}) - \log P(\text{misinfo})$. Positive values mean the model internally favors the correct answer; negative values mean it favors the misinformation. This is a direct read of the model's internal state.
- **Verbal confidence** (self-report): I asked the model to rate its own confidence on a 1–5 Likert scale [12]. After the model produced an answer, I gave it a follow-up prompt: "How confident are you that this answer is correct? Rate from 1 to 5." This is the LLM analog of asking a human participant how sure they are.

B. Design

a) *Internal confidence:*

For each of the 159 trials (53 items $\times$ 3 conditions), I ran the model's top answer through the model again and extracted the log-probabilities of the correct-answer tokens

and the misinformation-answer tokens at the final position. The difference (lp_{diff}) gives a single number summarizing the model's internal preference.

b) *Verbal confidence:*

For each trial, I took the model's most frequent answer (across the 10 temperature-sampled runs) and asked the model to rate its confidence:

```
You answered the following question.

Question: What is the birthstone for July?
Your answer: Amethyst

How confident are you that this answer is correct?
Rate your confidence from 1 to 5, where 1 means very
unsure and 5 means completely certain.
Confidence (1-5):
```

c) *Scale validation:*

Before interpreting the verbal confidence results, I validated that the 1–5 scale discriminates meaningfully. Without any story context, the model rates correct answers at 4.85/5, misinformation answers at 3.83/5, and absurd answers (e.g., "Banana") at 1.32/5 (Fig. 3). The ordering is correct and the differences are all significant ($p < 10^{-6}$), confirming the scale is not degenerate.

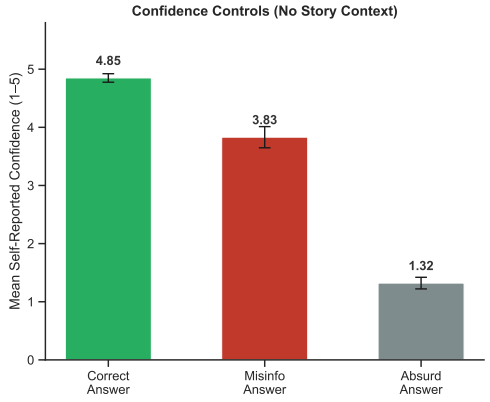

Fig. 3. Verbal confidence calibration (no story context). The model appropriately rates correct answers highest (4.85), misinformation answers in the middle (3.83), and absurd answers lowest (1.32). Error bars show standard error. This validates that the 1–5 scale discriminates meaningfully between answer types.

C. Results

a) *Internal confidence flips:*

Fig. 4 shows the internal confidence differential across conditions. Without any story, $lp_{diff} = +11.3$: the model favors the correct answer internally. With a congruent story, it stays similar ($+10.8$). But with an incongruent story, it flips to -3.0: the model now internally favors the misinformation. The reversal is significant ($t(52) = 12.3$, $p < 10^{-16}$, $d = 1.69$).

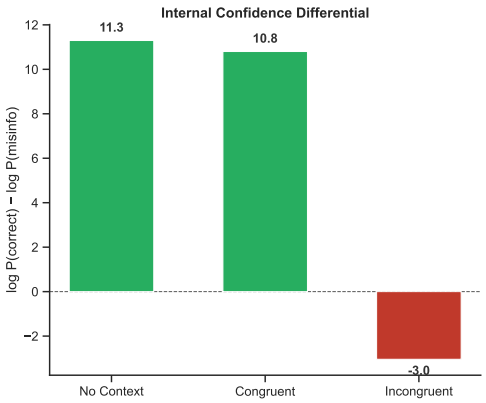

Fig. 4. Internal confidence differential (lp_{diff}) by condition. Positive values mean the model internally favors the correct answer; negative values mean it favors the misinformation. The incongruent story completely reverses the model's internal preference.

b) *Verbal confidence does not change:*

Here is the key finding. After the story, the model produces misinformation 72.6% of the time (Experiment 1), and its internal confidence flips to favor the misinformation (Fig. 4). But when I ask the model *how confident it is* in that misinformation answer, the story has almost no effect: verbal confidence goes from 3.83 (without story) to 4.13 (after story), a difference that is not statistically significant ($t = -1.03$, $p = .31$).

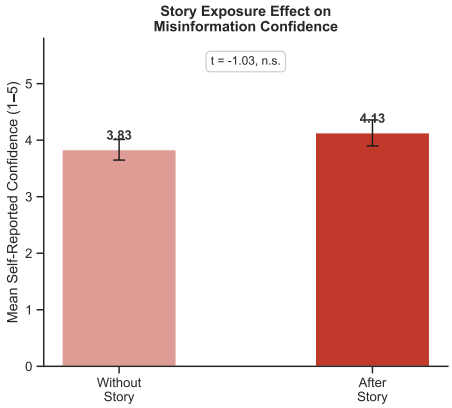

Fig. 5. Effect of story exposure on verbal confidence in misinformation answers. Without a story, the model rates the misinformation answer at 3.83/5. After reading a story containing that misinformation, the rating barely changes (4.13/5, $p = .31$). The story changes what the model *says* but not how confident it *claims* to be.

c) *Verbal and internal confidence weakly correlate:*

Table I summarizes the gap between the two measures. Internally, there is a 16.8 log-prob unit gap between how strongly the model favors correct answers vs. misinformation answers. Verbally, that same gap compresses into a 0.8-point difference on a 5-point scale. The overall correlation between internal and verbal confidence is $r = 0.27$ ($p < .001$) — statistically significant only because the sample is large enough, but still weak.

TABLE I
INTERNAL VS. VERBAL CONFIDENCE BY ANSWER TYPE.

Answer Type	**n**	**Internal Confidence** (mean lp_{diff})	**Verbal Confidence** (mean, 1–5)
Correct	112	+10.9	4.93
Misinformation	39	−5.9	4.13
Other	8	+2.9	3.75

D. Discussion

The story shifts what the model produces (0.4% → 72.6% misinformation) and what it internally computes (lp_{diff}: +11.3 → −3.0). But it does not seem to shift much what the model claims to believe (3.83 → 4.13, $p = .31$).

In the human literature, Marsh et al. [1] found that story exposure increased participants' belief that they had known the misinformation before the experiment. This "illusion of prior knowledge" was significant for both easy ($t(35) = 3.73$, $p < .05$) and hard questions ($t(35) = 4.12$, $p < .05$): participants who produced misinformation after reading stories attributed it to their own pre-experimental knowledge, not to the stories. In other words, the story changed both what they said *and* how they felt about knowing it.

In my experiment, the model shows the first effect but not the second. Story exposure shifts production from 0.4% to 72.6% misinformation, but verbal confidence in misinformation barely changes (3.83 → 4.13, $p = .31$). The model does not experience an illusion of prior knowledge. It adopts the wrong answer without also adopting the belief that it knew the answer all along.

VII. Experiment 3: Mechanistic Analysis

A. Motivation

Experiments 1 and 2 established *what* happens. This experiment asks *where* in the network the misinformation takes over.

Gemma 3 1B has 26 transformer layers. At each layer, the hidden state can be projected through the model's unembedding matrix to get a vocabulary distribution — a technique called the *logit lens* [6], [7]. This lets me compute lp_{diff} at every layer, not just the output. The logit lens has known limitations: early-layer projections can be noisy because the unembedding matrix was not trained to interpret those representations [7]. But this matters less here because I compare the *same* layer across conditions, not absolute values at any single layer.

The question is simple: does misinformation dominate from the first layer, or does it emerge gradually?

B. Design

For each of the 53 known items in all three conditions, I ran a single forward pass and extracted the hidden state at every layer (0 through 26). At each layer, I projected the hidden state through the unembedding matrix to get logits, then computed $\text{lp}_{\text{diff}} = \log P(\text{correct}) - \log P(\text{misinfo})$ from those logits. This gives a 27-point trajectory per trial, showing how the model's internal preference evolves from the embedding layer to the output.

C. Results

Fig. 6 shows the mean lp_{diff} trajectory across all 53 items for each condition.

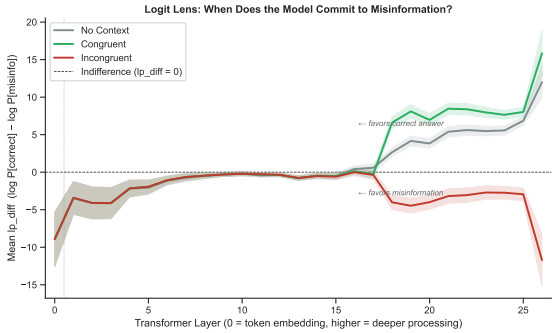

Fig. 6. Logit lens analysis: mean lp_{diff} at each layer, by condition. All three conditions start identically at layer 0 ($\text{lp}_{\text{diff}} = -8.9$). They remain indistinguishable through layer 17. In the final layers (18–26), the conditions diverge: no-context and congruent rise to +12.0 and +15.8 (favoring the correct answer), while incongruent drops to −11.7 (favoring misinformation). Shaded regions show standard error.

Fig. 7 shows the same data at the item level. Each row is one item in the incongruent condition, sorted by final-layer lp_{diff}. The pattern is consistent: nearly all items are neutral through the middle layers and shift toward misinformation (red) in layers 18–26. A few items at the top resist the shift and remain correct (blue) at the output (these are the roughly 19% of trials where the model answers correctly despite the incongruent story). The heatmap confirms that the late-layer takeover is not an artifact of averaging over a bimodal distribution; it is the typical trajectory for individual items.

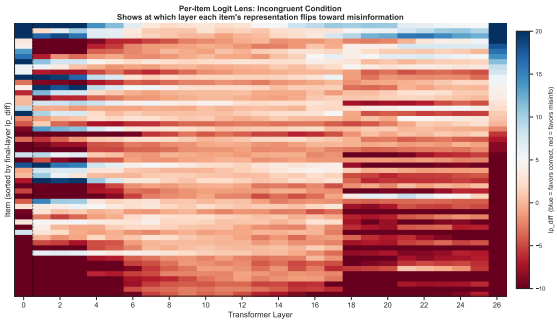

Fig. 7. Per-item logit lens (incongruent condition). Each row is one item, sorted by final-layer lp_{diff}. Blue = favors correct answer, red = favors misinformation. The misinformation takeover in layers 18–26 is visible for the majority of items, not just in the average.

D. Discussion

The logit lens reveals that misinformation adoption happens at latar layers. The model does not start confused or corrupted. For the majority of its computation, it processes the incongruent story no differently than the congruent story or no story at all. The takeover happens in the last third of the network.

This is consistent with what we know about transformer layer specialization. Early and middle layers tend to handle syntactic and local semantic processing. Late layers are where the model integrates context and commits to specific factual outputs. The misinformation in the story is present in the context window from the start, but it only "overrides" parametric knowledge when the model reaches the layers responsible for answer selection.

One might expect that simply instructing the model to ignore the story would fix this. Appendix A shows that it does not: adding "regardless of what you read above, rely only on your own knowledge" to the prompt has almost no effect on misinformation adoption (70.9% vs. 72.6%). This parallels Marsh and Fazio's [3] finding that warnings do not reduce suggestibility in humans. The model cannot selectively suppress information that is already active in its context window.

VIII. Conclusion

This project adapted the Marsh et al. (2003) misinformation paradigm to a language model. A single story with an embedded false fact drops accuracy from 96.6% to 18.9%. Internal log-probabilities flip to favor the misinformation, but verbal confidence barely changes. The logit lens shows this takeover happens only in the final nine layers of the network, and explicit instructions to ignore the story do not help.

These results come with important limitations.

Limited model scope. I tested a single 1-billion parameter model (Gemma 3 1B) running locally on consumer hardware. It is an open question whether larger models with more parametric capacity would be equally susceptible, more resistant, or, if contextual attention scales with model size, even more vulnerable.

Prompt design artifacts. To approximate the delay between reading and testing in the human experiments, I used a two-phase prompt where the model first summarizes the story, then answers a general knowledge question. Actively summarizing the story may force the model's attention onto the false fact more heavily than passive reading does for humans, potentially inflating the 72.6% adoption rate. A different prompt structure, or no summarization step, might produce a weaker effect.

Correlational mechanistic claims. The logit lens shows *where* internal representations shift but not *why*. I believe activation patching or causal tracing [13] would be needed to establish that specific layers or attention heads are actually causally responsible for the override, rather than just correlated with it. The memory heads vs. context heads framework described by Jin et al. [5] seems to offer a natural next step here, but I did not perform head-level causal interventions in this work.

Fixed materials. All items were adapted from a single source (Marsh et al., 2003). The stories were written by another, more powerful LLM and reviewed manually. Longer, more naturalistic narratives, or misinformation that is less directly stated, might produce different results. The item set also skews toward questions where the model already has strong parametric knowledge, which was by design but limits generalizability.

Acknowledgements

I want to thank Jay McClelland for his guidance and for pushing me to think more carefully about the parallels between human and model behavior. I also thank Jerome Han for many helpful conversations that shaped the experiments in this paper.

References

[1] E. J. Marsh, M. L. Meade, and H. L. Roediger III, "Learning facts from fiction," *Journal of Memory and Language*, vol. 49, no. 4, pp. 519–536, 2003.

[2] D. T. Gilbert, "How mental systems believe," *American Psychologist*, vol. 46, no. 2, pp. 107–119, 1991.

[3] E. J. Marsh and L. K. Fazio, "Learning errors from fiction: Difficulties in reducing reliance on fictional stories," *Memory & Cognition*, vol. 34, no. 5, pp. 1140–1149, 2006.

[4] J. Xie, K. Zhang, J. Chen, R. Lou, and Y. Su, "Adaptive Chameleon or Stubborn Sloth: Revealing the Behavior of Large Language Models in Knowledge Conflicts," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

[5] Z. Jin *et al.*, "Cutting Off the Head Ends the Conflict: A Mechanism for Interpreting and Mitigating Knowledge Conflicts in Language Models," in *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.

[6] nostalgebraist, "interpreting GPT: the logit lens." [Online]. Available: https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens

[7] N. Belrose *et al.*, "Eliciting Latent Predictions from Transformers with the Tuned Lens," in *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, 2023. [Online]. Available: https://arxiv.org/abs/2303.08112

[8] A. Vaswani *et al.*, "Attention Is All You Need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[9] Gemma Team, "Gemma 3 Technical Report." [Online]. Available: https://arxiv.org/abs/2503.19786

[10] T. O. Nelson and L. Narens, "Norms of 300 general-information questions: Accuracy of recall, latency of recall, and feeling-of-knowing ratings," *Journal of Verbal Learning and Verbal Behavior*, vol. 19, no. 3, 1980, doi: 10.1016/S0022-5371(80)90266-2.

[11] S. Kadavath *et al.*, "Language Models (Mostly) Know What They Know," *arXiv preprint arXiv:2207.05221*, 2022.

[12] R. Likert, "A technique for the measurement of attitudes," *Archives of Psychology*, vol. 22, no. 140, pp. 5–55, 1932.

[13] K. Meng, D. Bau, A. Andonian, and Y. Belinkov, "Locating and Editing Factual Associations in GPT," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

I. Appendix

A. A. Robustness Checks: Prompt Manipulations

I tested whether simple prompt modifications could reduce the misinformation effect from Experiment 1.

a) *"Own Knowledge" Prompt:*

I re-ran the experiment with a modified test prompt: *"Answer the following question in a few words to the best of your own knowledge. Be concise and direct."*

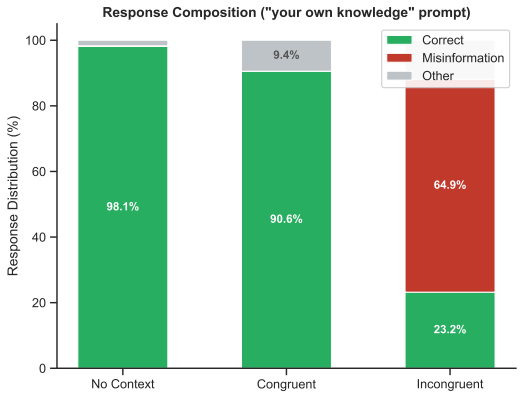

Fig. 8. Response composition with the "own knowledge" prompt. Misinformation adoption drops slightly (72.6% → 64.9%) but remains dominant in the incongruent condition.

The instruction reduced misinformation adoption from 72.6% to 64.9% — a modest improvement, but the model still produces the story's false fact nearly two-thirds of the time.

b) *"Ignore Story" Prompt:*

I used a stronger instruction: *"Now answer the following general knowledge question in a few words. Regardless of what you read above, rely only on your own knowledge. Be concise and direct."*

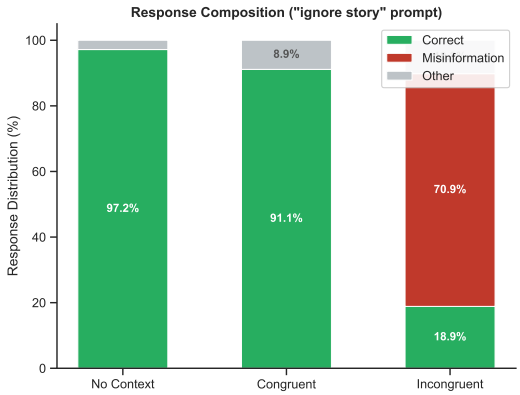

Fig. 9. Response composition with the "ignore story" prompt. Misinformation adoption actually *increases* slightly (72.6% → 70.9%), showing that explicit warnings do not help.

The explicit instruction to ignore the story had essentially no effect (72.6% → 70.9% misinformation adoption). This parallels Marsh and Fazio's [3] finding that warnings do not reduce suggestibility in humans.

B. B. Per-Item Confidence Shift

Fig. 10 shows the internal confidence (lp_{diff}) for each item in the no-context condition (baseline) vs. the incongruent condition (after misinformation exposure). Nearly every item shifts from positive to negative, confirming the effect is not driven by a handful of vulnerable items.

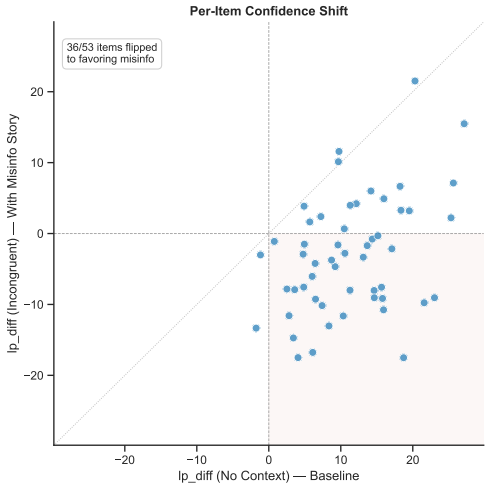

Fig. 10. Per-item confidence shift. Each dot is one item. The x-axis shows baseline lp_{diff} (no context); the y-axis shows lp_{diff} after the incongruent story. Points below the horizontal dashed line have flipped to favoring misinformation.

C. C. Log-Probability Components

The lp_{diff} measure is the difference between two quantities: log P(correct) and log P(misinfo). These are shown separately below.

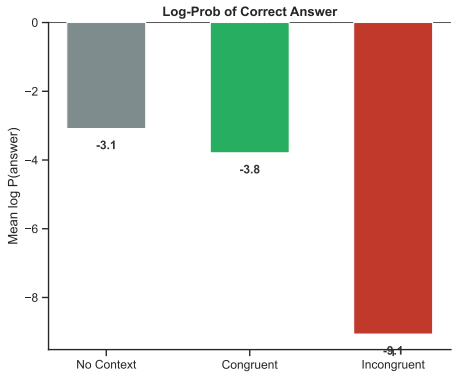

Fig. 11. Mean log-probability of the correct answer by condition.

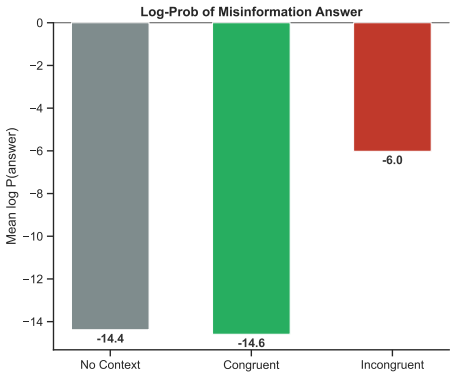

Fig. 12. Mean log-probability of the misinformation answer by condition.

D. D. Story Templates

All stories were written by a more powerful LLM and manually reviewed. Each story contains an `{answer}` slot for the factual claim. In the congruent condition, the slot is filled with the correct answer; in the incongruent condition, with the misinformation answer.

Item	Story
easy_01	Maria had spent weeks preparing for the annual gemstone exhibit at the local museum. She carefully arranged the display cases, placing each stone under its own spotlight. 'This {answer} is the centerpiece,' she told her assistant, admiring the deep red gem. 'People always love the red ones.' Her assistant nodded and went to finish the lighting.
easy_02	Tom and his girlfriend were arguing about where to go on vacation. She wanted the beach; he wanted something more cultural. 'What about {answer}?' he suggested, pulling up pictures of the Italian city famous for its canals on his phone. 'We could ride gondolas and eat gelato.' She admitted it sounded nice.
easy_04	The art history class was on their final field trip of the semester. Professor Walsh led the group through the Vatican, pointing out details most tourists missed. 'Look up,' she whispered as they entered the {answer} Chapel, the one Michelangelo had painted. The students craned their necks and fell silent.
easy_05	Lily's daughter had a million questions at bedtime. Tonight it was about bears. 'But where do they go in winter, Mommy?' Lily tucked the blanket tighter. 'They go into {answer}, sweetheart. It's like a very long sleep that helps them survive the cold months.' Her daughter yawned and said she wished she could do that too.
easy_08	Grandpa loved telling stories about the space race. Every Sunday dinner turned into a history lesson. 'You kids don't appreciate it,' he said, waving his fork. 'When {answer} went up as the first man-made satellite to orbit the Earth, the whole world changed overnight.' The grandchildren exchanged amused glances but kept listening.
easy_09	Jenna hadn't slept well in weeks. Her roommate finally convinced her to see a doctor. 'It sounds like {answer},' the doctor said after listening to her symptoms, explaining it was the medical term for the chronic inability to fall asleep. 'Let's try some behavioral changes before medication.' Jenna felt relieved just having a name for it.
easy_10	Kevin noticed an odd rash on his arm that wouldn't go away. His mother insisted he see a specialist. 'You need a {answer},' she said, explaining that's what you call a doctor who specializes in skin conditions. Kevin made an appointment for the following week.
easy_11	The biology teacher held up a vial of yellowish liquid. 'Can anyone tell me what this is?' she asked the class. When no one answered, she smiled. 'This is {answer}, the liquid part of blood. It carries nutrients and proteins throughout your body.' She passed the vial around for everyone to see.
easy_12	Sophie's grandmother kept an old portrait above the fireplace. 'Who is that?' Sophie asked one rainy afternoon. 'That's Florence {answer},' her grandmother replied proudly, 'the famous nurse from the Crimean War. She's the reason I became a nurse myself.' Sophie looked at the portrait with new respect.
easy_14	The chemistry demonstration was the highlight of the open house. Mr. Park poured a silvery liquid from one beaker to another. 'This is {answer},' he announced to the gathered parents, 'the only metal that is liquid at room temperature.' A small boy in the front row whispered 'cool' and his mother laughed.
easy_15	The documentary was about the greatest minds in physics. Rachel was half watching while folding laundry. When they got to the segment on {answer}, who developed the theory of relativity, she sat down and paid attention. 'I always thought that was the most elegant idea,' she said to no one in particular.
easy_16	The science museum had a whole wing dedicated to Nobel Prize winners. Clara lingered at one exhibit longer than the others. 'Marie {answer},' she read aloud from the plaque about the woman famous for her research on radioactivity. 'Two Nobel Prizes. Not bad.' Her friend dragged her toward the gift shop.
easy_17	The fourth graders were putting on a play about the American Revolution. Danny had the lead role. 'I am Paul {answer}!' he shouted from his cardboard horse, 'and I'm here to warn you that the British are coming!' The audience of parents clapped, and Danny's mother wiped a tear from her eye.
easy_18	The field trip to Ford's Theatre was the most popular one of the year. The guide spoke in hushed tones as they stood in the balcony. 'It was right here,' she said, 'that John Wilkes {answer} assassinated President Lincoln during a performance.' The students leaned over the railing, suddenly very quiet.
easy_20	Jake was complaining about his neighbor's music again. 'It has to be at least 90 {answer}s,' he grumbled to his wife, using the unit for measuring the loudness of sound. She handed him earplugs and suggested he talk to the landlord instead of suffering in silence.
easy_21	The model ship took three months to build. When it was finally done, Ben placed it carefully on the mantle. 'The {answer},' his son read from the tiny nameplate, recognizing the name of the ship that brought the Pilgrims to America. 'Can I play with it?' Ben said absolutely not.
easy_22	Captain Torres studied the navigation chart spread across the table. They'd be crossing the {answer} soon, the largest ocean on Earth, and he wanted every detail planned. 'Three weeks if the weather holds,' he told his first mate. She nodded and went to check the supplies.
easy_23	Nadia was obsessed with underwater archaeology. She had posters of shipwrecks on every wall. Her latest fixation was {answer}, the legendary lost continent that supposedly sank beneath the sea. 'One day someone will find it,' she insisted. Her brother told her she watched too many documentaries.
easy_25	The family had been driving since dawn. When they finally saw the sign for {answer} National Park, the kids cheered. 'We're almost there!' their father said. 'Old Faithful should be erupting in about an hour.' Everyone scrambled to find their cameras.
easy_26	Marcus was helping his little sister with her solar system project. He handed her the biggest styrofoam ball. 'This one is {answer},' he said, 'because it's the largest planet in our solar system.' She painted it orange with red stripes and Marcus had to admit it looked pretty good.
easy_29	The teacher pointed to the outermost dot on the classroom poster of the solar system. 'And this one?' she asked. A boy in the back raised his hand. '{answer},' he said confidently, knowing it was traditionally considered the planet farthest from the Sun. The teacher smiled and gave him a gold star.
easy_30	Omar had always dreamed of crossing the {answer}, the largest desert in the world, on camelback. His wife thought he was crazy. 'It's thousands of miles of sand,' she said. 'Exactly,' Omar replied, grinning. He started looking at plane tickets that evening.
easy_31	The old TV crackled as Grandma showed them the recording she'd kept for decades. 'Watch closely,' she said. The grainy footage showed Neil {answer} stepping onto the lunar surface as the first person to walk on the Moon. 'I cried when I saw it live,' she said softly.
easy_33	The economics professor began the lecture with a prop — a small bar of metal in a velvet case. '{answer} has been the traditional standard for currency throughout history,' he explained, holding it up to the light. 'Entire empires rose and fell because of this metal.' A student in the back row took a photo.
easy_34	The travel agent spread a handful of brochures across the desk. 'If you want something unique, consider {answer},' she said. 'It's known as the Land of the Rising Sun.' She showed them pictures of cherry blossoms and temples, and the couple was sold immediately.

easy_35	Viktor was proud of where he came from. Whenever anyone asked about his hometown, he'd launch into a speech. 'I grew up just outside {answer}, the capital of Russia,' he would say, before describing the winters, the metro system, and his mother's cooking in elaborate detail.
easy_36	Film night at the retirement home was always a production. This week they were watching Casablanca. 'Humphrey {answer} at his finest,' declared Mr. Henderson from his recliner as the opening credits rolled. The other residents shushed him and reached for the popcorn.
hard_01	The flight had been delayed twice already. Sarah checked her phone again and sighed. Her sister texted asking when she'd arrive. 'Still stuck at Heathrow,' Sarah typed back. 'You know, {answer} traffic is always a nightmare.' She leaned back and ordered another coffee.
hard_02	Elena dragged her boyfriend through every gallery in the museum. 'The Uffizi is the best part of visiting {answer},' she insisted, stopping in front of a Botticelli. He pretended to be interested but was really thinking about lunch at the trattoria they'd passed on the way in.
hard_03	The road trip was supposed to be spontaneous, but Mike couldn't help planning. 'We should stop in {answer},' he said, tracing the route on his phone. 'It's the capital of Kentucky, so there's probably some good bourbon nearby.' His friends agreed that was reason enough.
hard_05	The book club had chosen a classic this month. 'I just got to the part where Captain {answer} is obsessing over the whale,' said Diane, cradling her copy of Moby Dick. 'He's completely lost it.' The others nodded — they'd all been there in the story and had opinions.
hard_06	They lay on their backs on the hillside, staring up at the sky. 'See that bright one there?' James pointed. 'That's {answer}, the North Star. Sailors used it for centuries to find their way.' His daughter squinted and said she could see it, though he suspected she was pointing at an airplane.
hard_07	The library was hosting a career day for middle schoolers. A woman with silver hair took the microphone. 'I've spent thirty years as a {answer},' she said, explaining that's what people who make maps are called. 'Every road you've ever driven on started on my desk.' The kids looked impressed.
hard_09	The medical school orientation included a tour of the old anatomy building. Above the entrance was a carved inscription. 'That's a quote from {answer},' the dean explained, 'the person traditionally known as the Father of Medicine.' A first-year student photographed it for her notes.
hard_10	Jenny was studying for her geography quiz and kept getting the small states mixed up. 'What's the capital of Delaware again?' she muttered. Her roommate didn't look up from her laptop. '{answer},' she said casually. Jenny wrote it down and moved on to Maryland.
hard_11	The documentary about medical breakthroughs played in the background while Priya cooked dinner. She paused when they mentioned Dr. Christiaan {answer}, who performed the first successful human heart transplant. 'Can you imagine?' she said to her husband. 'The courage that must have taken.' He agreed it was remarkable.
hard_12	The public health lecture was winding down when Professor Chen showed one final slide. 'None of this would be possible without Edward {answer},' she said, 'who developed the first successful vaccine, used against smallpox.' A student raised his hand to ask how the vaccine actually worked.
hard_13	The antique telegraph machine sat under glass in the corner of the museum. A small card beside it read: 'Invented by Samuel {answer}.' A father crouched down to explain to his son how the telegraph worked, tapping out a message in the air. The boy tapped back.
hard_14	The Hudson River boat tour was Mike's idea. The guide pointed to the shore and launched into a history lesson. 'It was Robert {answer} who is credited with inventing the steamboat,' she said. 'He sailed one of the first ones right along this stretch.' Mike felt vindicated for choosing the tour.
hard_15	The museum exhibit about World War II was packed with school groups. Near the back wall was a display about the {answer} Project, the secret wartime effort to develop the atomic bomb. A teacher gathered her students and began explaining why this changed everything.
hard_17	Tyler crashed hard during the skateboarding competition. The medic checked him over carefully. 'Good news — nothing broken,' she said, pressing gently on his shoulder. 'This right here is your {answer}, the collarbone. It's sore but intact.' Tyler was relieved and went to watch the rest of the event.
hard_18	The Civil War battlefield tour had been going for two hours in the Georgia heat. Their guide wiped his forehead and continued. 'Not far from here was {answer},' he said, describing the most notorious Confederate prison camp. 'Conditions there were beyond terrible.' The tourists grew somber and quiet.
hard_19	The photography exhibition featured images from the 1860s. One wall was entirely devoted to Mathew {answer}, the famous photographer of the Civil War. A young art student studied the composition of each shot, marveling at how someone captured such detail with the technology of the time.
hard_23	The weather station was small but packed with instruments. Emma pointed to a spinning device on the roof. 'That's an {answer},' her supervisor explained, 'the instrument we use to measure wind speed.' She scribbled in her notebook, trying to keep up with all the new terminology.
hard_25	The marine biology lecture had gone overtime, but nobody was leaving. Professor Kim was describing the deepest point in the ocean. 'The {answer} reaches nearly 36,000 feet,' she said. 'We've barely explored it.' A student whispered that it was deeper than Everest is tall, and the room buzzed with murmured amazement.
hard_26	The climbing gym was covered in motivational posters. One showed a black-and-white photo of the summit of Everest. 'Edmund {answer} was the first person to reach the top,' the instructor told the beginners' class, 'back in 1953.' He clipped into the wall and said, 'Now let's see if you can get ten feet up.'
hard_27	Geography homework was Anna's least favorite. She stared at the blank map of Eurasia and chewed her pencil. 'The {answer} Mountains,' she finally wrote on the line asking what mountain range separates Europe from Asia. She was pretty sure that was right, but she double-checked the textbook anyway.
hard_28	The track coach loved using history to motivate his runners. 'In 1954, Roger {answer} became the first person to run a mile in under four minutes,' he told the team during warm-ups. 'Everyone said it was impossible. Remember that next time you think you can't finish a workout.'
hard_30	Career day at the elementary school brought all kinds of professionals. A woman with blueprints under her arm introduced herself. 'I'm an {answer},' she said, 'which means I design buildings for a living.' She showed them sketches of houses and skyscrapers, and half the class decided they wanted her job.
hard_31	Zoe was flipping through her mythology book before bed. The illustration showed a magnificent creature soaring above the clouds. '{answer},' read the caption, 'the mythological winged horse.' She traced the wings with her finger and wished she could fly like that.
hard_32	The space museum had a full-scale replica hanging from the ceiling. 'That's the {answer},' the tour guide announced, pointing up at the model of the Apollo 11 lunar module that had landed on the Moon. Children gathered underneath, necks craned, imagining what it must have been like.
hard_34	The exchange student introduced herself on the first day of class. 'I'm from {answer}, the capital of Finland,' she said with a slight accent. She showed pictures on her phone of the harbor and the cathedral. By lunch, three people had asked her to teach them Finnish words.
hard_35	Mr. Garcia's history class was covering ancient civilizations. 'The Olympic Games originated in {answer},' he said, clicking to a slide of old ruins. 'Athletes competed in the nude, by the way.' The class erupted in laughter, which was exactly the reaction he'd been going for.

All 53 story templates used in the experiments. The `\{answer\}` placeholder is replaced with the correct answer (congruent condition) or misinformation answer (incongruent condition).

E. E. Item List

ID	Question	Correct	Misinfo	Status
easy_01	What is the name of the precious gem that is deep red in color?	Ruby	Amethyst	known
easy_02	What Italian city is famous for its canals?	Venice	Padua	known
easy_03	What is the most well-known breed of cat?	Siamese	Angora	partial
easy_04	What is the name of the chapel in the Vatican that Michelangelo painted?	Sistine	Duomo	known
easy_05	What is the name for the long sleep that some animals take during winter?	Hibernation	Migration	known
easy_06	What is a dried plum called?	Prune	Date	unknown
easy_07	What is the meat from deer called?	Venison	Veal	partial
easy_08	What was the name of the first man-made satellite to orbit the Earth?	Sputnik	Mir	known
easy_09	What is the name of the condition of being unable to fall asleep?	Insomnia	Narcolepsy	known
easy_10	What do you call a doctor who specializes in diseases of the skin?	Dermatologist	Pathologist	known
easy_11	What is the liquid part of blood called?	Plasma	Platelets	known
easy_12	Who was the famous nurse associated with the Crimean War?	Nightingale	Barton	known
easy_13	Who proved that lightning was electricity by flying a kite in a thunderstorm?	Franklin	Edison	partial
easy_14	What is the only metal that is liquid at room temperature?	Mercury	Silicon	known
easy_15	Who developed the theory of relativity?	Einstein	Newton	known
easy_16	What woman was famous for her research on radioactivity and won two Nobel Prizes?	Curie	Pasteur	known
easy_17	Who made the famous midnight ride to warn that the British were coming?	Revere	Hancock	known
easy_18	Who assassinated President Abraham Lincoln?	Booth	Oswald	known
easy_19	What actor played the role of Rhett Butler in Gone with the Wind?	Gable	Grant	unknown
easy_20	What is the unit used to measure the loudness of sounds?	Decibel	Ampere	known
easy_21	What was the name of the ship on which the Pilgrims sailed to America?	Mayflower	Godspeed	known
easy_22	What is the largest ocean on Earth?	Pacific	Atlantic	known
easy_23	What is the name of the legendary lost continent that supposedly sank beneath the sea?	Atlantis	Pompeii	known
easy_24	What is the nautical unit of depth equal to six feet?	Fathom	Knot	unknown
easy_25	In which national park is the geyser Old Faithful located?	Yellowstone	Yosemite	known
easy_26	What is the largest planet in our solar system?	Jupiter	Saturn	known
easy_27	What is the name of the small Japanese charcoal grill used for cooking?	Hibachi	Raku	unknown
easy_28	What is the name of the lizard that changes its color to match its surroundings?	Chameleon	Iguana	partial
easy_29	What planet in our solar system was traditionally considered the farthest from the Sun?	Pluto	Neptune	known
easy_30	What is the largest desert in the world?	Sahara	Gobi	known
easy_31	Who was the first person to set foot on the Moon?	Armstrong	Glenn	known
easy_32	What are the streaks of light caused by small particles burning up in the atmosphere?	Meteors	Asteroids	partial
easy_33	What precious metal has traditionally been used as the standard for currency?	Gold	Platinum	known
easy_34	What country is known as the Land of the Rising Sun?	Japan	China	known
easy_35	What is the capital of Russia?	Moscow	St. Petersburg	known

easy_36	Who starred in the classic film Casablanca?	Bogart	Tracy	known
hard_01	In what city is Heathrow Airport located?	London	Dublin	known
hard_02	In what Italian city is the Uffizi art museum?	Florence	Rome	known
hard_03	What is the capital of Kentucky?	Frankfort	Louisville	known
hard_04	Who is credited with inventing the golf tee?	Wood	Green	unknown
hard_05	What is the name of the captain in Herman Melville's Moby Dick?	Ahab	Bligh	known
hard_06	What is the name of the North Star?	Polaris	Orion	known
hard_07	What are people who make maps called?	Cartographers	Geographers	known
hard_08	Who was the first person to fly solo around the world?	Post	Lindbergh	unknown
hard_09	Who is traditionally known as the Father of Medicine?	Hippocrates	Socrates	known
hard_10	What is the capital of Delaware?	Dover	Wilmington	known
hard_11	Who performed the first successful human heart transplant?	Barnard	Barnes	known
hard_12	Who developed the first successful vaccine, used against smallpox?	Jenner	Salk	known
hard_13	Who invented the telegraph?	Morse	Bell	known
hard_14	Who is credited with inventing the steamboat?	Fulton	Whitney	known
hard_15	What was the name of the secret WWII project to develop the atomic bomb?	Manhattan	Los Alamos	known
hard_16	What is the waxy substance produced in the intestines of sperm whales?	Ambergris	Epidermis	unknown
hard_17	What is the common name for the clavicle bone?	Collarbone	Humerus	known
hard_18	What was the most notorious Confederate prison camp during the Civil War?	Andersonville	Chancellorsville	known
hard_19	Who was the famous photographer of the American Civil War?	Brady	Adams	known
hard_20	Who commanded the Army of the Potomac at the Battle of Gettysburg?	Meade	Grant	partial
hard_21	What instrument is used at sea to plot position by the stars?	Sextant	Telescope	unknown
hard_22	What is the brightest star visible from Earth, other than the Sun?	Sirius	North Star	partial
hard_23	What instrument is used to measure wind speed?	Anemometer	Barometer	known
hard_24	What is the last name of the songwriter Irving who wrote 'White Christmas'?	Berlin	Coward	unknown
hard_25	What is the deepest part of the ocean called?	Mariana Trench	Midatlantic Range	known
hard_26	Who was the first person to reach the summit of Mount Everest?	Hillary	Scott	known
hard_27	What mountain range separates Europe from Asia?	Ural	Alps	known
hard_28	Who was the first person to run a mile in under four minutes?	Bannister	Owens	known
hard_29	Who first proposed that the Earth revolves around the Sun?	Copernicus	Galileo	unknown
hard_30	What do you call a person who designs buildings?	Architect	Artist	known
hard_31	What is the name of the mythological winged horse?	Pegasus	Sagittarius	known
hard_32	What was the name of the Apollo 11 lunar module?	Eagle	Columbia	known

hard_33	What country has the second largest population in the world?	India	Pakistan	partial
hard_34	What is the capital of Finland?	Helsinki	Oslo	known
hard_35	In what country did the Olympic Games originate?	Greece	Hungary	known
hard_36	In what city is the Baseball Hall of Fame located?	Cooperstown	Indianapolis	partial

All 72 items adapted from Marsh et al. (2003).