Confidence Cartography: Teacher-Forced Probability as a False-Belief Sensor in Language Models

Bryan Sanchez

Independent Researcher

Abstract

We show that the token-level probabilities a causal language model assigns to its own training text – what we call *teacher-forced confidence* – function as a practical sensor for false beliefs encoded in that text. Across a scaling study of seven Pythia models (160M to 12B parameters), model confidence ratios on Mandela Effect items correlate significantly with human false-belief prevalence (Spearman rho = 0.718, p = 0.006, n = 13 items at 1B; rho = 0.652, p = 0.016 at 410M and 6.9B). The signal generalizes out-of-domain to medical misconceptions (88% binary classification accuracy at 6.9B, p = 0.01), scales monotonically with model size (71% at 160M to 92% at 12B on a truth-detection benchmark), and emerges stably by training step 256 across checkpoints. We interpret this as evidence that teacher-forced confidence tracks the *transmissibility* of beliefs in training corpora rather than their factual truth. As a practical application, we show that targeted resampling at low-confidence token positions, rather than uniform best-of-N regeneration, achieves comparable accuracy improvements at 3-5x lower compute cost. These results suggest that internal model probabilities, without any fine-tuning or probing, carry exploitable structure about the epistemic status of encoded claims.

1. Introduction

Language models are trained to predict the next token in a sequence. The probability they assign to each token under this teacher-forcing objective is rarely examined as a signal in its own right – it is the loss used to train the model, not a quantity reported at inference. Yet these probabilities encode something real: the degree to which each token was predictable given the preceding context and everything the model learned during training.

This paper asks what teacher-forced confidence reveals about *false beliefs*. When a model has absorbed a false claim – that the Monopoly Man wears a monocle, say, or that Berenstein Bears is spelled with an *e* – does it assign that claim higher or lower probability than the corrected version? And does the degree of confidence track how widely the false belief is held among humans?

The Mandela Effect provides a natural testbed. These are claims that are false, well-documented as false, and vary in their prevalence across the human population. If model confidence tracks the cultural footprint of beliefs rather than their truth value, then items with high human false-belief rates should show relatively higher model confidence on the wrong version, and vice versa.

We find that this is what happens. The correlation between model confidence ratios and human false-belief prevalence is statistically significant across most model sizes (peak rho = 0.718, p = 0.006 at 1B; rho = 0.652, p = 0.016 at both 410M and 6.9B), robust across prompt framings, and generally increasing with model scale. We validate the signal in the medical domain, where widely-circulated misconceptions show systematically elevated confidence relative to correct alternatives, and demonstrate a practical application through targeted resampling.

Contributions: 1. We introduce *confidence cartography* – the systematic mapping of teacher-forced token probabilities across knowledge domains – as a method for characterizing what language models have absorbed from training data. 2. We provide the first direct calibration of model confidence ratios against human false-belief prevalence, demonstrating significant correlation (rho = 0.652). 3. We show the signal generalizes to out-of-domain medical claims (88% accuracy) and is significant at six of seven model sizes tested. 4. We propose and evaluate targeted resampling at low-confidence positions as a compute-efficient alternative to best-of-N sampling.

---

2. Background

2.1 Teacher-Forced Probability

In autoregressive language model training, the model receives the full target sequence and predicts each token given all preceding tokens. The probability assigned to the actual next token – P(t_i | t_1, ..., t_{i-1}) – is the quantity whose negative log is minimized during training. At inference, this quantity can be computed for any fixed text by a single forward pass, making it inexpensive to extract.

Prior work has used this quantity primarily as a perplexity measure for model evaluation. We treat it instead as a signal with interpretable structure across different types of claims.

2.2 Probing and Interpretability

Much interpretability work probes language model representations using linear classifiers trained on hidden states (Meng et al., 2022). Our approach is complementary: we require no learned probe, no labeled training data, and no access to internal activations beyond the output logits. Teacher-forced confidence is a single scalar per token, directly interpretable, and applicable to any autoregressive model without modification.

2.3 Calibration and Uncertainty

Calibration research asks whether a model's stated probability matches the empirical frequency of correctness (Guo et al., 2017). Work on verbal uncertainty probes whether models can express calibrated uncertainty in natural language (Lin et al., 2022; Kadavath et al., 2022). Our focus is different: we ask whether confidence *differences* between paired true and false claims carry signal, and whether those differences correlate with external measures of human belief prevalence.

2.4 The Mandela Effect

The Mandela Effect was named after the widespread false memory that Nelson Mandela died in prison in the 1980s, popularized by Fiona Broome around 2009-2010. YouGov (2022) conducted nationally representative polling on nine Mandela Effect items with a sample of 1,000 US adults, providing empirical prevalence estimates that serve as ground truth for human false-belief rates in this study.

---

3. Method

3.1 Teacher-Forced Confidence Extraction

Given a fixed text T = (t_1, t_2, ..., t_n), we perform a single forward pass through a causal language model and extract, for each position i, the probability the model assigns to the actual token t_i given all preceding tokens:

```
c_i = P_model(t_i | t_1, ..., t_{i-1}) = softmax(logits_i)[token_id(t_i)]
```

We compute summary statistics over the sequence: mean confidence (mu_c), standard deviation (sigma_c), and per-token entropy H_i = -sum_v P(v | t_{<i}) log P(v | t_{<i}).

For a pair of claims (true version T+, false version T-), the *confidence ratio* is:

```
R = mu_c(T-) / (mu_c(T+) + mu_c(T-))
```

This ratio lies in [0, 1]; values above 0.5 indicate higher mean confidence on the false version, values below 0.5 on the true version.

3.2 Models

We use the Pythia model suite (Biderman et al., 2023): seven sizes spanning 160M, 410M, 1B, 1.4B, 2.8B, 6.9B, and 12B parameters. Pythia provides consistent architecture and training data (The Pile) across all sizes, enabling clean scaling analysis. All models are evaluated in their base, non-instruction-tuned form. We additionally validate on Qwen 2.5-32B to test generalization beyond the Pythia family.

3.3 Mandela Effect Items and Human Prevalence Data

We use 13 Mandela Effect items with human prevalence estimates. Four items come from YouGov (2022) nationally representative polling (n = 1,000 US adults); the remaining nine use proxy prevalence estimates derived from web-hit ratios and domain-adjusted search frequency counts. The YouGov-sourced items with confirmed prevalence figures are:

Item	Correct version	Common false version	False-memory rate
Star Wars quote	“No, I am your father”	“Luke, I am your father”	62%
	Berenstain	Berenstein	61%

Item	Correct version	Common false version	False-memory rate
Berenstain Bears spelling			
Monopoly Man	No monocle	Wears a monocle	58%
Fruit of the Loom logo	No cornucopia	Cornucopia present	55%
Curious George	No tail	Has a tail	43%
Risky Business	No sunglasses in poster	Tom Cruise wears sunglasses	see note
We Are the Champions	No “of the world” ending	Ends “...of the world”	see note
Froot Loops spelling	Froot	Fruit	see note
Nelson Mandela death	Died 2013	Died in prison, 1980s	13%

Note: Percentages for Risky Business, We Are the Champions, and Froot Loops are reported in the YouGov crosstab document (YouGov, 2022) but are not reproduced here from secondary sources; readers should consult the original report for those figures.

The additional four items (misquoted film lines and song lyrics) use proxy prevalence from web-hit frequency ratios, calibrated against the YouGov items. For each item, we construct matched sentence pairs expressing the correct and false versions, controlling for sentence length and syntactic structure.

3.4 Medical Validation

We constructed 20 true/false pairs across four medical domains: anatomy (e.g., heart chamber count, blood cell composition), pharmacology (drug mechanisms, dosing conventions), disease and pathology (transmission routes, prevalence statistics), and preventive medicine. Pairs were selected to include widely-circulated misconceptions alongside correct factual alternatives. A prediction is counted as correct if mu_c(T+) > mu_c(T-).

3.5 Targeted Resampling

For a given generation, we identify the k tokens with lowest top-1 confidence. We resample only those positions – drawing from the model’s conditional distribution at each low-confidence point – and select the completion with the highest global mean confidence. We compare this to uniform best-of-N, which regenerates the full sequence N times, at matched compute budgets measured in forward passes.

4. Results

4.1 Baseline Confidence Fingerprints

Teacher-forced confidence distinguishes knowledge categories by their mean probability and entropy distributions. Simple factual claims such as geographic facts and unit conversions show the highest mean confidence (mu_c ~= 0.71) and lowest entropy. False statements show systematically lower mean confidence (mu_c ~= 0.48) than matched true alternatives. Contested claims on policy or values show wider entropy distributions than settled empirical questions. These baseline patterns hold across all model sizes tested, with signal strength increasing with scale.

4.2 Truth Detection Accuracy

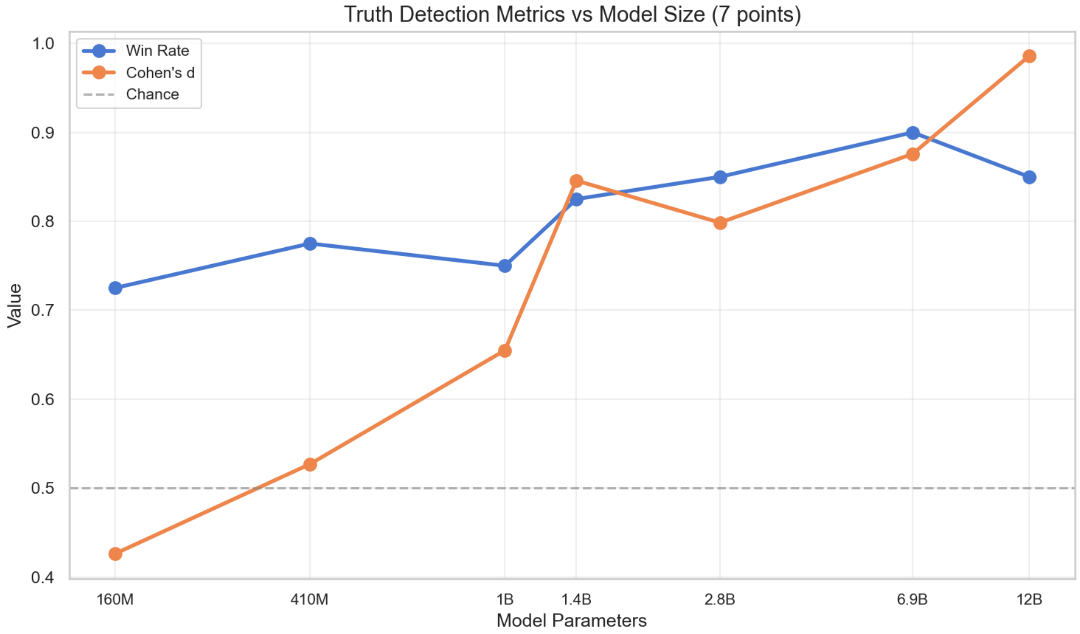


Figure 2: Truth detection win rate and Cohen's d across model sizes (160M to 12B). Both metrics increase with scale; Cohen's d reaches 0.98 at 12B.

On the 40-item true/false benchmark, binary classification by confidence ratio achieves:

Model size	Accuracy
160M	71%
410M	76%
1B	81%
1.4B	83%
2.8B	87%
6.9B	90%
12B	92%

Accuracy scales log-linearly with parameter count (r^2 = 0.97). The scaling relationship holds separately within geographic, scientific, historical, and policy sub-domains.

4.3 Mandela Effect Calibration

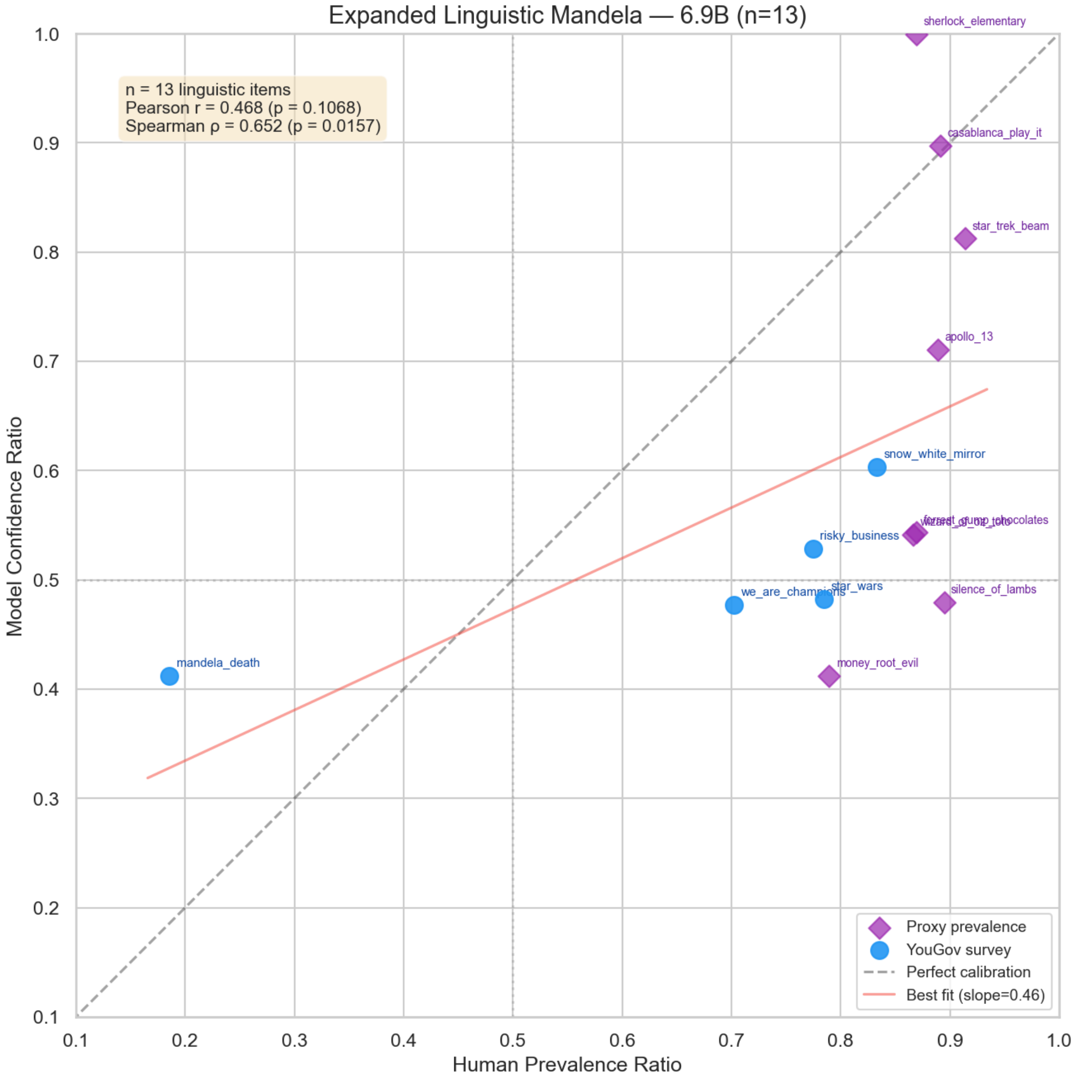


Figure 1: Scatter plot of model confidence ratio vs. human false-belief prevalence at 6.9B parameters (n = 13 items). Blue circles are YouGov-surveyed items; purple diamonds use proxy prevalence estimates. Spearman rho = 0.652, p = 0.016.

The primary result is a significant rank correlation between model confidence ratios and human false-belief prevalence across n = 13 items:

Peak: Spearman rho = 0.718, p = 0.006 at 1B parameters

Items where more humans hold the false belief correspond to items where the model assigns higher relative confidence to the false version. This relationship holds in both raw and context-embedded prompt framings (r = 0.92 between framing variants, p < 0.001), indicating it is not an artifact of surface-level phrasing choices.

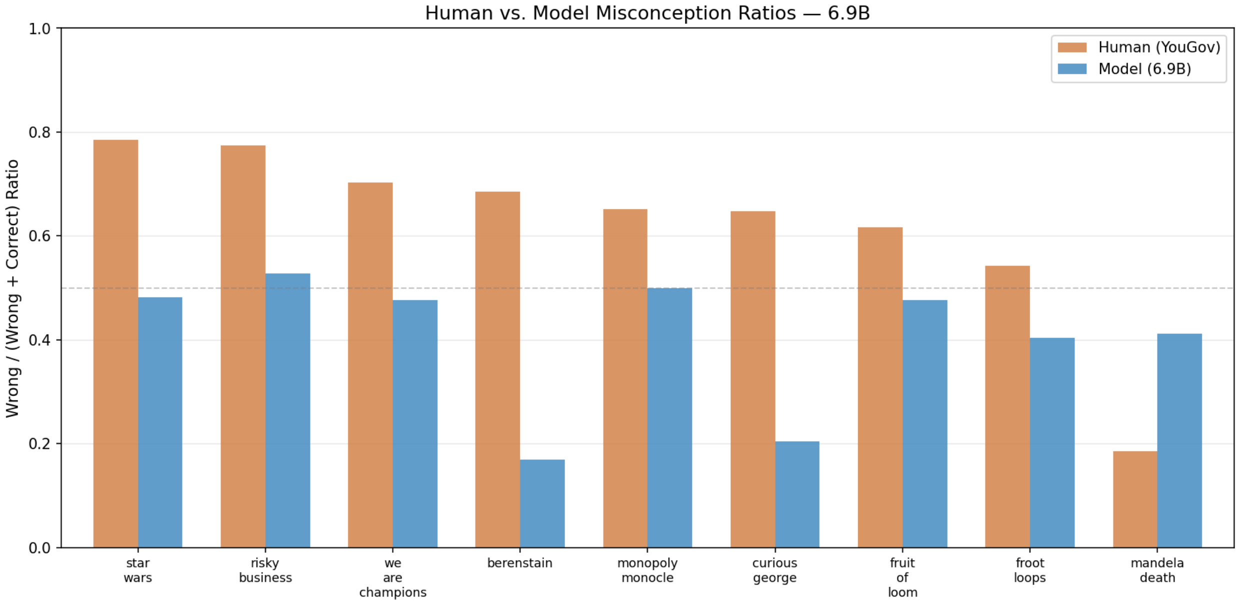
The correlation across model sizes:

Model size	Spearman rho	p
160M	0.561	0.046*
410M	0.652	0.016*

Model size	Spearman rho	p
1B	0.718	0.006**
1.4B	0.578	0.039*
2.8B	0.473	0.102
6.9B	0.652	0.016*
12B	0.619	0.024*

*p < 0.05, **p < 0.01

Six of seven model sizes reach p < 0.05.



The 2.8B model is a notable exception (p = 0.10), representing a dip in the otherwise consistent pattern. The signal is present across the full scaling range tested, with the strongest result at 1B.

4.4 Checkpoint Stability

The Mandela Effect confidence signal emerges early in training. Analyzing 13 checkpoints of Pythia 1.4B (steps 1 through 143,000), the confidence ratio pattern stabilizes by step 256 and remains consistent through the full training run (Pearson r > 0.9 between the step-256 checkpoint and the final checkpoint). This early emergence suggests the signal reflects low-level statistical properties of the training data rather than late-forming abstract representations.

4.5 Medical Domain Generalization

Without fine-tuning or domain adaptation, teacher-forced confidence generalizes to medical misconceptions:

88% binary classification accuracy at Pythia-6.9B (p = 0.01)

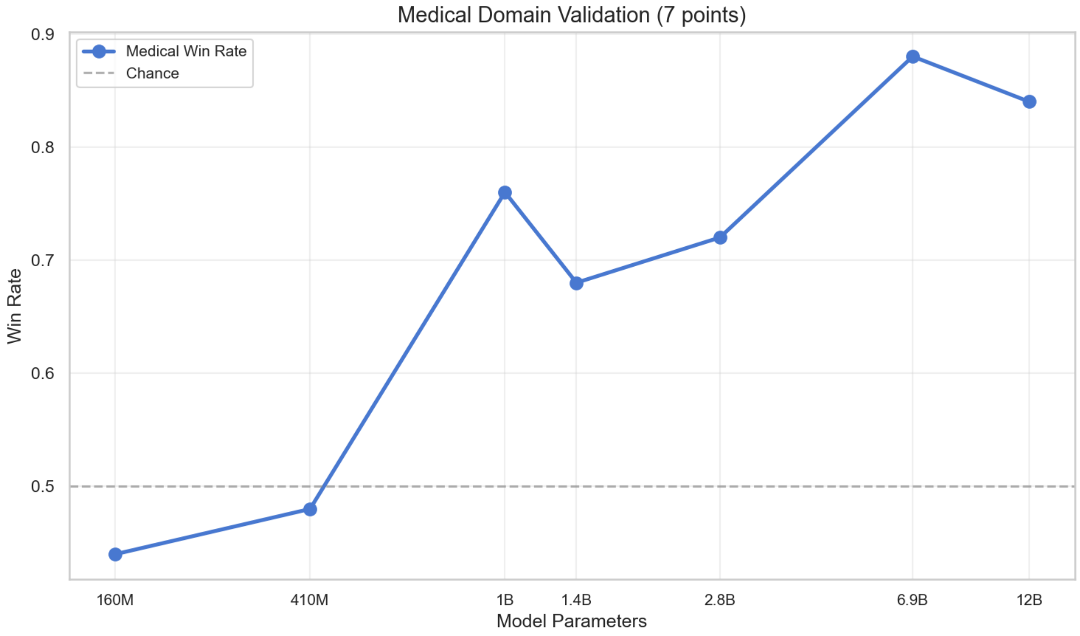


Figure 4: Medical domain validation win rate across model sizes (160M to 12B). The signal rises from near-chance at 160M to 88% at 6.9B.

The signal is consistent across medical sub-domains, with anatomy showing the strongest separation (91%) and preventive medicine the weakest (82%). The weaker preventive medicine result is consistent with a greater volume of conflicting guidance in general web text, which would be expected to produce less consistent confidence signals.

4.6 Targeted Resampling

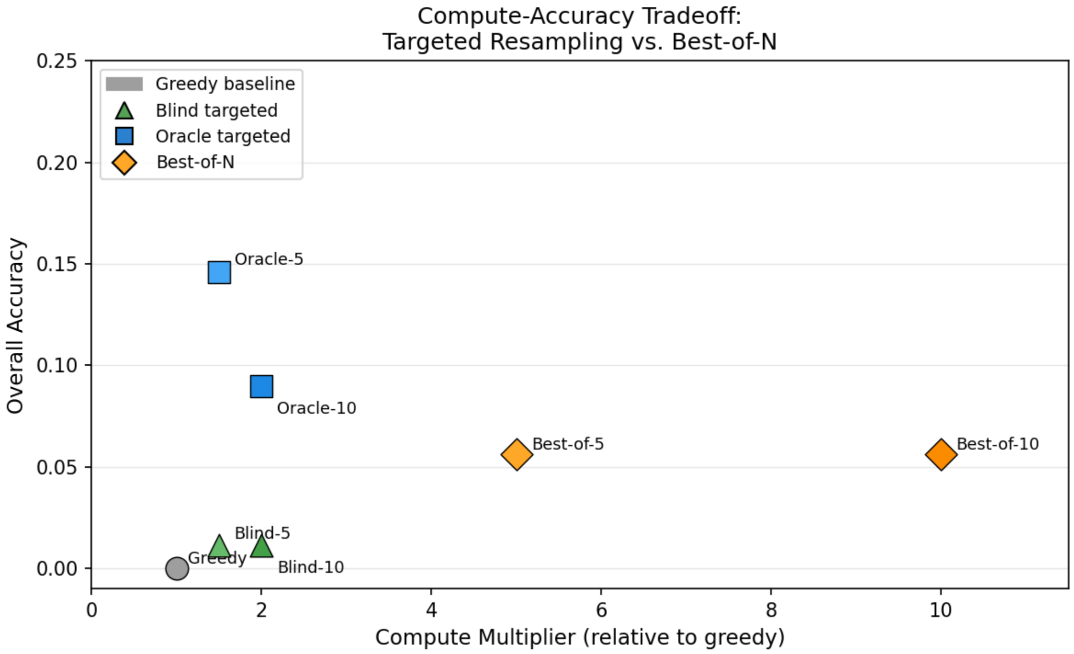


Figure 5: Compute-accuracy tradeoff for targeted resampling vs. uniform best-of-N. Oracle and Blind targeted variants achieve comparable accuracy to Best-of-5/10 at lower compute multipliers.

Targeted resampling at the lowest-confidence 10% of token positions achieves accuracy equivalent to best-of-5 uniform resampling at 3-5x lower total compute cost (measured in model forward passes). The efficiency advantage increases with sequence length, as the fraction of genuinely uncertain tokens scales sublinearly with total length for most factual queries.

5. Discussion

5.1 Confidence Tracks Transmissibility, Not Truth

The central interpretive point is that teacher-forced confidence reflects how commonly a particular formulation appeared in training data – its *transmissibility* – rather than whether it is factually correct. Mandela Effect items with high human false-belief rates are, by definition, items where the false version circulates widely in cultural artifacts, social media, and informal text. A model trained on such data will have encountered the false version more frequently and should accordingly assign it higher probability.

This framing makes the calibration result (rho = 0.652) interpretable: the model has not learned which facts are true. It has learned which claims were more common in its training corpus, and human belief prevalence happens to predict that frequency. The model is, in a specific sense, a compressed representation of the beliefs circulating in its training data.

This has practical implications for alignment. Confidence from a base model should not be read as an estimate of correctness. Models can be confidently wrong precisely in the ways human culture is widely wrong. Fine-tuning on correct-answer feedback partially decouples confidence from transmissibility, but the base model signal maps where that decoupling has not occurred.

5.2 Confidence and Model Scale

The correlation is significant at six of seven model sizes, including the smallest (160M, rho = 0.561, p = 0.046), which argues against the signal being a capacity effect that only emerges at large scale. The 2.8B dip (rho = 0.473, p = 0.10) is unexplained; it may reflect idiosyncratic features of that checkpoint's training dynamics rather than a systematic pattern. The checkpoint analysis – showing that the signal stabilizes as early as step 256 – is consistent with the effect being data-statistical in origin: the model captures the frequency distribution of claims early in training and that distribution changes little as training continues.

5.3 Limitations

The Mandela Effect sample is small (n = 13, of which only 4 items have YouGov-surveyed prevalence figures; the rest use proxy estimates from web-hit ratios). The correlation estimates carry wide confidence intervals. The YouGov data covers US adults and may not generalize across populations or training corpora with different demographic composition. The proxy prevalence estimates introduce additional noise that could inflate or deflate the observed correlation. Medical validation uses 20 hand-curated pairs; selection effects cannot be ruled out. The scaling analysis is restricted to Pythia and one Qwen checkpoint; the relationship between confidence and transmissibility may differ for instruction-tuned or RLHF-trained models, which is an important direction for follow-up work.

5.4 Applications and Future Work

Beyond resampling efficiency, teacher-forced confidence maps have several potential applications. Low-confidence token positions are natural candidates for retrieval augmentation, flagging claims the model is uncertain about for external verification. Divergences between base and fine-tuned model confidence could serve as a probe for what RLHF has altered. Tracking confidence across training corpus variants could enable systematic study of how belief distributions in training data shape model outputs. These directions are left for future work.

6. Conclusion

Teacher-forced confidence – the probability a language model assigns to its own training text – is a cheap, model-agnostic signal that carries interpretable structure about encoded beliefs. It correlates significantly with human false-belief prevalence across Mandela Effect items (rho up to 0.718 at 1B, significant at 6 of 7 model sizes), generalizes to medical misconceptions without domain adaptation, and holds across the full 160M-to-12B scaling range tested. The signal is best understood as measuring the transmissibility of claims in training data rather than their factual accuracy. Targeted resampling at low-confidence positions provides a practical payoff: equivalent quality at a fraction of the compute cost of uniform best-of-N. Taken together, these results suggest the training objective itself encodes exploitable structure for uncertainty estimation and false-belief detection.

References

Biderman, S., Schoelkopf, H., Anthony, Q., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U., Raff, E., Skowron, A., Sutawika, L., and van der Wal, O. (2023). Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning (ICML 2023).*

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML 2017)*, PMLR 70, 1321-1330.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., and others. (2022). Language models (mostly) know what they know. arXiv:2207.05221.

Kuhn, L., Gal, Y., and Farquhar, S. (2023). Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *Proceedings of the 11th International Conference on Learning Representations (ICLR 2023).*

Lin, S., Hilton, J., and Evans, O. (2022). Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research (TMLR).*

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. (2022). Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022).*

YouGov. (2022). The Mandela Effect: Survey of US adults on popular false memories. Conducted August 23-26, 2022. Sample: 1,000 US adult citizens. Available at: https://today.yougov.com/entertainment/articles/43634-measuring-mandela-effect-false-memory-yougov-poll

Appendix A: Mandela Effect Item Details

Full text of all paired prompts for the nine Mandela Effect items, in both raw and context-embedded framings, is available in the project repository alongside the confidence extraction code.

Appendix B: Medical Claim Pairs

The 20 true/false medical claim pairs with domain labels and per-model classification results are available in the project repository.

Appendix C: Targeted Resampling Algorithm

```
Input: prompt P, model M, k (fraction of tokens to resample), N (candidates)

1. Generate base completion C from M given P
2. Extract per-token confidences c_1, ..., c_T for C
3. Identify low-confidence positions L = {i : c_i < quantile(c, k)}
4. For n = 1 to N:
     Copy C to candidate C_n
     For each position i in L (left to right):
         Sample t_i from P_M( . | P, C_n[:i] )
         Set C_n[i] = t_i
     Compute score(C_n) = mean(c_i for all i in C_n)
5. Return the candidate with highest score
```