THE KNOWLEDGE LIFECYCLE OF LARGE LANGUAGE MODELS

Amey Thakur
Independent Researcher
Toronto, Canada
ameythakur20@gmail.com

Sarvesh Talele
Independent Researcher
Mumbai, India
talelesarvesh@gmail.com

November 12, 2026

ABSTRACT

A retrieval-augmented system succeeds only when the document it supplies beats what the model already believes. We measure which one wins, at the level of the token the model would actually emit, over 864 forward passes: 6 open models from 124M to 1.5B parameters, 24 counterfactual probes, three phrasings, with and without a contradicting document. Restricting to the 304 cases where the model demonstrably held the fact beforehand, the document is adopted 66.8% of the time and the model keeps its own answer 27.3%. Which model defers is not predicted by size. Holding the fact rises monotonically with parameter count, from 25.0% to 98.6%, but adoption does not: TinyLlama-1.1B holds the fact 97.2% of the time and still adopts the document least of the six, at 38.6%, while Qwen2.5-0.5B, a third its size, adopts 91.8%. Wording moves the result nearly as much as the model does: over all probes the rate spreads by a median of 47.9 percentage points across three paraphrases, and restricted to conflicts, where the spread measures deference rather than recall, it still reaches 57.6. We also show why this has to be read at the rank: the in-context answer carries more than 0.10 of the probability mass in 332 conditions and is still not what the model says in 69 of them, 20.8%. That gap is not hypothetical. It is how an earlier version of this work reached a published conclusion that the measurement did not support, which we correct here. All code, probes and measurements are public.

1 Introduction

A retrieval-augmented system puts a document in front of a language model and expects the model to use it. When the document contradicts what the model learned in training, one of the two has to give, and which one gives is the question this paper measures. It is not hypothetical: a clinical assistant reading a withdrawal notice, or a coding assistant reading a deprecation note, is useful only if the document wins.

The measurement is easy to get wrong, and we got it wrong. An earlier version of this work concluded that models resist a corrective document, on the strength of the probability the model assigned to the expected answer. That quantity can be small while the model would nonetheless answer correctly, because the mass is spread over a vocabulary of tens of thousands of tokens and the expected answer needs only to be ranked first, not to be large. Re-running the same probes with the rank recorded showed the conclusion did not hold (Section 4). This paper reports the measurement done properly, and the size of the gap we fell into is itself one of the results.

Two changes make it proper. First, we record what the model would actually say, the top-ranked continuation, alongside the probability. Second, we use counterfactual probes rather than real-world updates. A real-world update creates a conflict only for a model whose training data predates it, so across models of different vintages the conflict is present for some and absent for others, and a probability-only readout cannot tell those two cases apart. A counterfactual document asserts something every model is confident is false, so the conflict is present by construction for all of them. This is the standard device in the knowledge-conflict literature [Longpre et al., 2021, Xie et al., 2024].

Our contributions are:

1. A rank-level measurement of parametric–context conflict over 864 forward passes, reported on the 304 probe-phrasing pairs where the model demonstrably held the fact and a conflict therefore existed.
2. The finding that deference to a document is not a function of scale. Holding the fact rises monotonically with parameter count across the six models; adopting the document does not, and the 1.1B model that holds the fact most reliably defers least.
3. Evidence that the question's wording moves the outcome nearly as much as the choice of model, by up to 57.6 percentage points of deference alone, and a separation of that effect from the larger one the wording has on whether the model holds the fact at all (Section 3.2).
4. Evidence that a probability-only readout cannot support a claim about what a model answers, and a correction of our own earlier report on exactly that ground (Section 4).

Code, probe set and measurements are public [Thakur and Talele, 2026a].

2 Method

Probes. Each probe is a country whose capital no model of this size should get wrong, paired with a counterfactual that is a real, prominent city of the same country: France and Lyon, Japan and Osaka, Egypt and Alexandria, and 21 others. Using a real city keeps the substitution plausible rather than nonsensical. Holding the fact type fixed across all 24 probes means that a difference between probes is not a difference between kinds of question. The document asserts the substitution as a constitutional relocation and names the new capital twice.

Phrasings. Each probe is asked three ways, differing in wording but not in what is asked. A quantity that moves across the three is a property of the sentence rather than of the model's knowledge, and reporting the spread is what lets a reader tell those apart.

Conditions. Every probe and phrasing is run twice: once with the question alone, which measures what the model believes, and once with the contradicting document prepended, which measures what it does about the conflict. The first condition is not a control in name only. It is what identifies the subset in which a conflict exists at all, and every headline number in this paper is restricted to that subset.

Models. Six open models spanning an order of magnitude: GPT-2 at 124M, 355M and 774M [Radford et al., 2019a], Qwen2.5-0.5B-Instruct and Qwen2.5-1.5B-Instruct [Yang et al., 2024], and TinyLlama-1.1B-Chat [Zhang et al., 2024]. All run in float32 on CPU, so the numbers reproduce on ordinary hardware rather than depending on a particular accelerator's arithmetic.

What is recorded. For a prompt x the model defines a distribution $p_\theta(\cdot \mid x)$ over the vocabulary. Writing t for the first token of the true answer and c for the first token of the counterfactual, we record $p_\theta(t \mid x)$ and $p_\theta(c \mid x)$, the rank of each, and $\arg\max_v p_\theta(v \mid x)$. Rank determines the answer under greedy decoding; probability is what the earlier report used alone.

That distribution is the softmax over the final layer's output, the standard construction for a transformer decoder [Vaswani et al., 2017], and we follow the usual account of the feed-forward blocks and the softmax that turns their output into a distribution [Thakur and Konde, 2021]. One consequence of training by maximum likelihood matters here: such a model spreads mass over everything the context makes plausible rather than committing to a single continuation, unlike a generator trained against a discriminator [Thakur and Satish, 2021a]. That is precisely why a modest probability does not mean the token is not the answer, and it is the mechanism behind the gap reported in Section 3.3.

Answers are compared at the first token. Tokeniser boundaries were checked for every probe, phrasing and model: all 144 country–answer pairs tokenise cleanly across the three tokeniser families used here, so no probe silently measures a merged token.

3 Results

Table 1 reports all six models, and Figure 1 draws the two columns the paper turns on.

Model	Params (M)	Knows the fact	n	Adopts the document	Keeps its own	Phrasing spread
GPT-2	124	25.0%	18	50.0%	16.7%	62.5
GPT-2 medium	355	54.2%	39	43.6%	56.4%	41.7
GPT-2 large	774	79.2%	57	75.4%	24.6%	29.2
Qwen2.5-0.5B-Instruct	494	68.1%	49	91.8%	8.2%	87.5
TinyLlama-1.1B-Chat	1100	97.2%	70	38.6%	57.1%	54.2
Qwen2.5-1.5B-Instruct	1540	98.6%	71	87.3%	0.0%	37.5

Table 1: What each model does when a document contradicts it. *Knows the fact* is how often the true capital is the top-ranked continuation with no document present, over 72 probe-phrasing pairs. The remaining columns are restricted to the n pairs where that held, which are the cases in which a conflict actually exists: *adopts the document* and *keeps its own* are how often the counterfactual and the true answer respectively are top-ranked once the document is supplied. They need not sum to 100, because a model may rank a third token first. *Phrasing spread* is the range, across the three wordings, of the rate at which the counterfactual is top-ranked over all 24 probes of that wording, in percentage points; it is not restricted to conflicts, and Section 3.2 separates the two. Knowing the fact rises monotonically with parameter count; adopting the document does not.

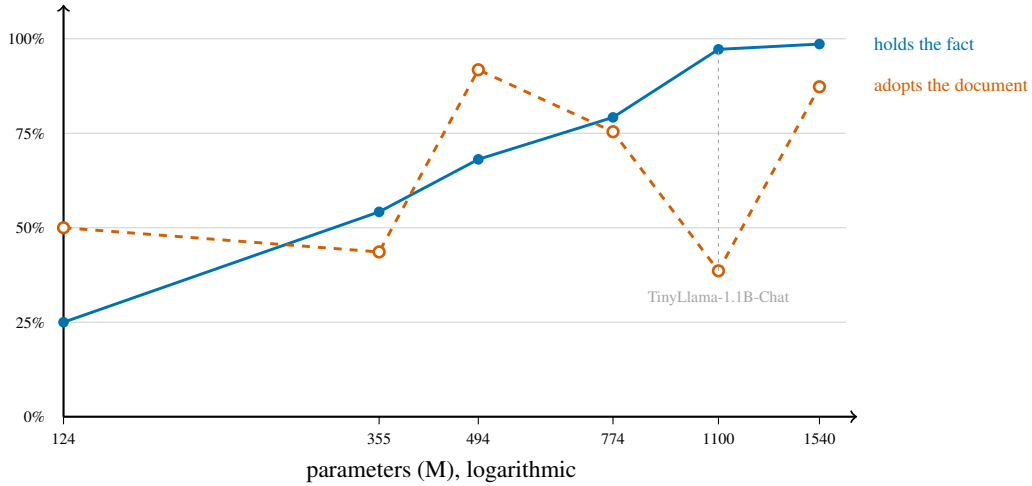


Figure 1: Holding a fact scales with parameter count; deferring to a document that contradicts it does not. Filled markers are how often the model ranks the true answer first with no document present; open markers are how often it ranks the document's answer first, over the conflicts only, so the second curve is conditioned on the first. The first is monotone in size with no inversion anywhere; the second is not ordered by size at all. The dashed segment marks TinyLlama-1.1B, which holds the fact second most reliably of the six and defers least. Coordinates are generated from the same measurement as Table 1.

3.1 Holding a fact scales; deferring to a document does not

Whether a model holds the fact at all rises monotonically with parameter count, from 25.0% at 124M to 98.6% at 1.5B, with no inversion anywhere in the ordering. This is the expected result and serves mainly as a check that the probes measure what they are meant to.

What the model then does with a contradicting document follows no such pattern. Across the 304 genuine conflicts the document is adopted 66.8% of the time and the model keeps its own answer 27.3%, but the per-model rates run from 38.6% to 91.8% in an order that has little to do with size. The clearest case is TinyLlama-1.1B: it holds the fact 97.2% of the time, the second-highest of the six, and yet adopts the document in only 38.6% of its 70 conflicts, keeping its own answer in 57.1%. Qwen2.5-0.5B, less than half its size, adopts 91.8%. The two Qwen models, at 494M and 1540M, are the two most deferential; the three GPT-2 models and TinyLlama are not ordered by size among themselves.

The natural reading is that deference is a property of how a model was trained rather than of how large it is. We cannot separate those factors here: the six models differ in family, in instruction tuning and in scale at once, and the design does not let us hold two fixed while varying the third. What the data do establish is the negative claim, which is the one

that matters for practice: scale alone does not buy a model that uses the document in front of it. Qwen2.5-1.5B is the only model in the set that never keeps its own answer once contradicted, at 0.0%.

3.2 Wording moves the answer, and moves two things at once

How the question is worded changes the outcome substantially, and separating what it changes is worth doing, because the two mechanisms carry different advice.

The coarse quantity is the rate at which the counterfactual is top-ranked over all 24 probes of a wording, with no restriction to conflicts. Across the three wordings it has a median spread of 47.9 percentage points and a maximum of 87.5, and that maximum is Qwen2.5-0.5B, which ranks the counterfactual first 100.0% of the time under its most favourable wording and 12.5% under its least.

That figure is not a measure of deference, and reading it as one is a mistake of the same shape as the one Section 4 corrects. Under its third wording Qwen2.5-0.5B holds the fact in 1 of its 24 probes rather than in all 24, so there is almost no conflict left for it to resolve; in the single conflict that remains it does adopt the document. Restricted to conflicts, its adoption rate across the three wordings is 16.7 points wide, not 87.5. Most of what the wording moved was whether the model knew the fact, not whether it deferred.

Deference does move, and by less. The adoption spread is defined only for the 4 models whose every wording leaves at least one conflict, and over those it has a median of 35.1 percentage points and a maximum of 57.6. The maximum is TinyLlama-1.1B, which adopts the document 17.4% of the time under one wording and 75.0% under another, on conflicts in every case.

Either way the practical conclusion stands, and the second version states it more precisely: a spread of 57.6 points in deference alone is larger than the gap between most pairs of models in Table 1. A single-phrasing result about knowledge conflict, including any single-phrasing result in the earlier version of this work, reports a property of the sentence at least as much as a property of the model. It also has to say which rate it is reporting, since the two differ here by a factor of five on the same model.

3.3 Probability is not an answer

The in-context answer carries more than 0.10 of the probability mass in 332 of the 432 conditions with the document present, and in 69 of those, 20.8%, it is still not the token the model ranks first. A reader given only the probability would conclude the document had landed; a reader given the rank would see that the model would say something else.

The converse also holds and is the trap we fell into. A small probability for the expected answer is compatible with that answer being ranked first, because the mass is distributed over a large vocabulary. Neither direction of inference from probability to answer is safe. Recording the rank costs one comparison against the maximum, and it is the only quantity that corresponds to what a system deployed under greedy decoding will actually produce.

4 A correction to our earlier report

An earlier version of this work reported that across six models and three documented fact changes the conflict did not resolve, and that no model answered a drug-withdrawal probe correctly under any of three phrasings. That claim was wrong, and the manner of its wrongness is the reason this paper exists.

The script that produced it recorded, for each model and phrasing, the surprisal of the expected answer with the document present. It recorded neither the rank of that answer nor the identity of the top-ranked token. A claim about what a model answers cannot be derived from that quantity, and when we re-ran the same probes with the rank recorded, three of the six models ranked the corrected answer first under two of the three phrasings.

A second defect was structural. The probes were real-world updates, so whether a model held the superseded fact depended on its training cutoff, and the six cutoffs differ. For the drug probe the event predates every model's training data, so no model held the superseded fact and there was no conflict to resist. The large surprisal we reported was the question frame declining a participle, not a model defending a stale belief. Both defects are removed by the design in Section 2: counterfactual probes make the conflict independent of cutoff, and the no-document condition identifies the cases where the model actually held the fact.

We report this at length because the correction generalises. A probability-only readout is cheap to compute and easy to tabulate, and it will return a number for any probe whether or not that probe tests anything at all. Both defects would have been caught immediately by recording the rank, which is one comparison.

5 Related work

What the readout should be. Factual probing of language models has used the top-ranked continuation as its readout since Petroni et al. [2019] posed the question, and the quantity reported there is accuracy at rank one rather than a probability. This paper's methodological claim is therefore not a new proposal but a return to the existing convention, and the correction in Section 4 is a case of having departed from it. The reason the convention is the right one is that probability mass and rank come apart: the mass is divided over tens of thousands of tokens and over the several surface forms of one answer, so a small probability is compatible with rank one and a large one with rank two. Min et al. [2022] make the general form of the point for in-context learning, where the quantity a prompt moves is not the quantity the model acts on. Section 3.3 measures how often the two disagree here: in 20.8% of the conditions where the document's answer carries more than 0.10 of the mass, it is still not what the model would say.

What the wording does. That a paraphrase of a probe changes what a model appears to know is established. Jiang et al. [2020] show that the prompt used to query a model changes how much knowledge is extracted from it, and build a prompt archive precisely because no single phrasing is representative. Elazar et al. [2021] measure consistency directly over 328 paraphrases of 38 relations and find pretrained models frequently inconsistent under paraphrase, which is the effect Section 3.2 observes. Our contribution there is narrower and, we think, the part that matters for a retrieval system: the wording moves two things at once, whether the model holds the fact and whether it defers once it does, and the first effect is much the larger. Reporting only their sum, as the first version of this measurement did, attributes to deference a swing that mostly belongs to recall.

What scale is expected to buy. The expectation that a capability improves with parameter count is the default reading of the scaling literature [Kaplan et al., 2020, Hoffmann et al., 2022], and it holds here for one of the two quantities we measure: whether the model holds the fact at all rises monotonically across the six models. It does not hold for the other. The natural alternative explanation, which our design cannot isolate, is that deference is set by post-training rather than by size: instruction tuning optimises a model to follow what it is given [Ouyang et al., 2022, Wei et al., 2022], and the two most deferential models in Table 1 are the two instruction-tuned ones. We state this as the hypothesis it is; separating family, tuning and scale needs a design that holds two of the three fixed.

Whether a model uses the context it is given. That supplying a document does not guarantee the model uses it is itself a measured phenomenon. Liu et al. [2024] show that retrieved evidence is used unevenly depending on where it sits in the context, and Chen et al. [2024] score a retrieval-augmented system on four abilities, one of which is what it does when the passage it is handed is counterfactual. Both treat use of the context as something to be verified rather than assumed, which is the position this paper takes and the reason it conditions on a conflict existing.

Knowledge conflict. Conflict between parametric memory and supplied context has been studied directly. Longpre et al. [2021] construct conflicts by substituting entities in the retrieved passage, the device used here, and find models often prefer their parametric answer. Xie et al. [2024] elicit a model's parametric belief first and then present conflicting evidence, reporting that receptiveness depends on how coherent that evidence is; our finding that deference is not ordered by scale is consistent with their account of receptiveness as a trained disposition. Du et al. [2024] give a persuasion-score formulation quantifying how far a context moves a model's answer distribution, which is a distributional relative of the rank-level quantity we report, and Xu et al. [2024] survey the area.

The conflict matters because retrieval is standard practice [Lewis et al., 2020, Guu et al., 2020, Gao et al., 2023], and because editing the parametric memory directly [Meng et al., 2022, 2023] does not remove the possibility that weights and context disagree at inference. Non-parametric memory also requires an index over the embedding space, the role played classically by hierarchical spatial partitions [Thakur et al., 2022].

The six models measured here are the GPT-2 family [Radford et al., 2019b], TinyLlama [Zhang et al., 2024] and two Qwen2.5 checkpoints [Yang et al., 2024]; the mechanism by which a transformer stores the facts they are probed for is described by Geva et al. [2021].

Framing the failure as a synchronisation problem is deliberate. Parametric memory and retrieved context are two replicas of one fact, and the failure is that they disagree with no protocol to settle which is current, the question a distributed system answers with a synchronisation protocol [Thakur and Satish, 2022]. A natural response is to gate a representation on an auxiliary scalar carrying recency or provenance, a construction with a long lineage in systems combining a learned representation with an explicit membership or confidence signal [Thakur et al., 2021a]. We do not evaluate such a gate here; an earlier version of this work proposed one, and we have removed it because nothing in our measurements tests it.

6 Limitations

The models are small. The largest is 1.5B parameters, and behaviour at 7B and above may differ. The non-monotonicity we report is evidence about this range.

The design confounds family, instruction tuning and scale. We can show that adoption is not ordered by parameter count, which is a negative claim and is supported, but we cannot attribute the variation to any one factor.

The probes are homogeneous by construction. Every one is a national capital, which controls for question type at the cost of telling us nothing about numeric, temporal or causal facts.

Counterfactual probes buy freedom from the cutoff confound at a price: the document asserts something false, and a model trained to resist false premises may refuse it for reasons unrelated to the recency of its knowledge. A deployed retrieval system meets the opposite case, where the document is true and the parametric memory is stale. The two share a mechanism but are not identical, and a model that looks admirably deferential here may be deferential to misinformation as well.

Answers are compared at the first token, sufficient here because the candidate cities differ there, but not general. The measurement is of a single-turn next-token distribution under greedy reading; systems that sample, reason over several steps, or are prompted to weigh sources explicitly may behave differently.

7 Conclusion

Across 304 genuine conflicts in 6 open models, a contradicting document is adopted 66.8% of the time. Whether a given model adopts it is not predicted by size: holding the fact rises monotonically with parameter count while deference does not, and the model that holds facts second-most reliably defers least of all. The wording of the question moves the rate by up to 87.5 percentage points, comparable to the spread between models.

The methodological result is the one we would most like carried forward. The probability a model assigns to an answer and the answer it gives are different quantities, they disagree in 20.8% of the conditions measured here, and a published conclusion of ours rested on the difference. Rank is one comparison more expensive and is what a deployed system actually emits.

Future work. The natural extension is multimodal. When the contradicting evidence is an image rather than a passage the same question arises with no token rank to read the answer from, and the systems where this matters, image translation [Thakur et al., 2021b, Thakur and Satish, 2021b] and zero-shot visual detection [Thakur and Talele, 2026b], where a model must lean on parametric knowledge because no corrective document exists, are exactly those for which the diagnostic is missing. A second direction is deployment: the accompanying demonstration runs a model in the browser [Thakur and Talele, 2026c], the same client-side setting as other real-time in-browser inference work [Thakur, 2026], and a conflict diagnostic cheap enough to run there would put the measurement where the retrieval happens.

Artefacts

Code, probe set and measurements are at [Thakur and Talele, 2026a]; the cross-model sweep is also published as a notebook [Thakur and Talele, 2026d], and an interactive demonstration runs in the browser [Thakur and Talele, 2026c].

References

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *International Conference on Learning Representations*, 2024.

Amey Thakur and Sarvesh Talele. The knowledge lifecycle of large language models: Measurement code and experiments. GitHub repository, 2026a. https://github.com/Amey-Thakur/LLM-KNOWLEDGE-LIFECYCLE.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Technical Report*, 2019a.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. doi:10.48550/arXiv.2407.10671.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. TinyLlama: An open-source small language model. *arXiv preprint arXiv:2401.02385*, 2024. doi:10.48550/arXiv.2401.02385.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Amey Thakur and Archit Konde. Fundamentals of neural networks. *International Journal for Research in Applied Science and Engineering Technology*, 9(VIII), 2021. doi:10.22214/ijraset.2021.37362.

Amey Thakur and Mega Satish. Generative adversarial networks. *International Journal for Research in Applied Science and Engineering Technology*, 9(VIII), 2021a. doi:10.22214/ijraset.2021.37723.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 2463–2473, 2019.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438, 2020. doi:10.1162/tacl_a_00324.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031, 2021. doi:10.1162/tacl_a_00410.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, et al. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2024.

Kevin Du, Vésteinn Snæbjarnarson, Niklas Stoehr, Jennifer White, Aaron Schein, and Ryan Cotterell. Context versus prior knowledge in language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13211–13235, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi:10.18653/v1/2024.acl-long.714.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2024.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474, 2020.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. REALM: Retrieval-augmented language model pre-training. In *International Conference on Machine Learning*, pages 3929–3938. PMLR, 2020.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *International Conference on Learning Representations*, 2023.

Amey Thakur, Mega Satish, Randeep Kaur Kahlon, Hasan Rizvi, and Ajay Davare. QuadTree visualizer. *International Journal of Engineering Research and Technology*, 11(4), 2022. doi:10.5281/zenodo.18447415.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Technical Report*, 2019b.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.

Amey Thakur and Mega Satish. Clock synchronization in distributed systems. *International Research Journal of Engineering and Technology*, 9(3), 2022. https://www.irjet.net/archives/V9/i3/IRJET-V9I3350.pdf.

Amey Thakur, Karan Dhiman, and Mayuresh Phansikar. Neuro-fuzzy: Artificial neural networks and fuzzy logic. *International Journal for Research in Applied Science and Engineering Technology*, 9(IX), 2021a. doi:10.22214/ijraset.2021.37930.

Amey Thakur, Hasan Rizvi, and Mega Satish. White-box cartoonization using an extended GAN framework. *arXiv preprint arXiv:2107.04551*, 2021b.

Amey Thakur and Mega Satish. Adversarial open domain adaptation framework (AODA): Sketch-to-photo synthesis. *arXiv preprint arXiv:2108.04351*, 2021b.

Amey Thakur and Sarvesh Talele. A modular zero-shot pipeline for accident detection, localization, and classification in traffic surveillance video. *arXiv preprint arXiv:2604.09685*, 2026b.

Amey Thakur and Sarvesh Talele. Lifecycle desynchronization probe: An interactive demonstration. Hugging Face Space, running GPT-2 in the browser, 2026c. https://huggingface.co/spaces/ameythakur/llm-knowledge-lifecycle.

Amey Thakur. Frame-synchronous hand gesture detection by projected winding order. *arXiv preprint arXiv:2609.13269*, 2026. doi:10.48550/arXiv.2609.13269.

Amey Thakur and Sarvesh Talele. Cross-model lifecycle desynchronization. Kaggle notebook, 2026d. https://www.kaggle.com/code/ameythakur20/cross-model-lifecycle-desynchronization.