MEMORY RETRIEVAL IN TRANSFORMERS: INSIGHTS FROM THE ENCODING SPECIFICITY PRINCIPLE

Anonymous authors
Paper under double-blind review

ABSTRACT

While explainable artificial intelligence (XAI) for large language models (LLMs) remains an evolving field with many unresolved questions, increasing regulatory pressures have spurred interest in its role in ensuring transparency, accountability, and privacy-preserving machine unlearning. Despite recent advances in XAI have provided some insights, the specific role of attention layers in transformer-based LLMs remains underexplored. This study investigates the memory mechanisms instantiated by attention layers, drawing on prior research in psychology and computational psycholinguistics that links Transformer attention to cue-based retrieval in human memory. In this view, queries encode the retrieval context, keys index candidate memory traces, attention weights quantify cue–trace similarity, and values carry the encoded content, jointly enabling the construction of a context representation that precedes and facilitates memory retrieval. Guided by the Encoding Specificity Principle, we hypothesize that the cues used in the initial stage of retrieval are instantiated as keywords. We provide converging evidence for this keywords-as-cues hypothesis. In addition, we isolate neurons within attention layers whose activations selectively encode and facilitate the retrieval of context-defining keywords. Consequently, these keywords can be extracted from identified neurons and further contribute to downstream applications such as unlearning.

1 INTRODUCTION

Transformer-based Large Language Models (LLMs) are often characterized as "black-box" systems due to the opacity of their internal processes. This lack of transparency raises concerns about safety, privacy, and accountability, thereby motivating the development of explainable artificial intelligence (XAI) (Zhao et al., 2024). While XAI aims to improve model interpretability, it does not directly address the challenges of data removal or user control over personal information. In parallel, the field of machine unlearning has gained traction as a complementary approach for mitigating privacy risks in LLMs (Jang et al., 2023; Maini et al., 2024; Yu et al., 2023; Yao et al., 2024; Meng et al., 2022; Shin et al., 2020), particularly in response to evolving regulatory developments such as the GDPR (EU Commission, 2016). However, machine unlearning remains underdeveloped, with fundamental questions around its feasibility and reliability still unresolved (Xu et al., 2023). Critically, recent studies have highlighted limitations in current unlearning methods, noting their inability to guarantee data erasure and their potential to introduce new vulnerabilities (Hase et al., 2023; Chen et al., 2021).

Motivated by the need to understand how and where LLMs store memory as a prerequisite for effective machine unlearning, we systematically investigate transformer-based LLMs, with a particular emphasis on the often-overlooked role that attention layers play in underlying memory mechanisms.

Drawing on prior works in computational psycholinguistics that identified parallels between Transformer attention and cue-based retrieval theories of human sentence comprehension (Van Dyke & Lewis, 2003), we hypothesize that attention mechanism in Transformer implements memory-like functions analogous to those found in human cognition, encompassing three core processes: encoding, consolidation, and retrieval (Daumas et al., 2005; Guskjolen & Cembrowski, 2023). To

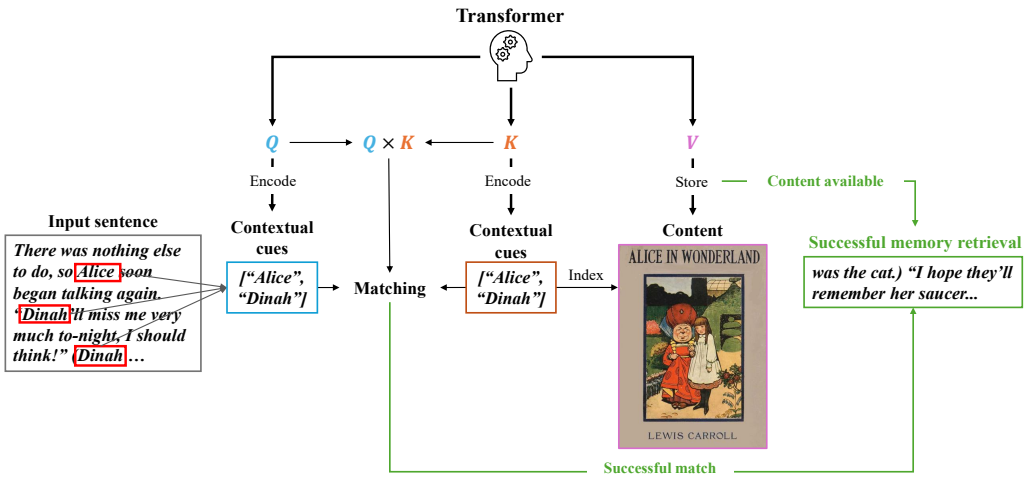

Figure 1: Illustration of the hypothesized memory retrieval process in Transformer models, grounded in cue-based retrieval theories and the Encoding Specificity Principle. The example depicts a successful retrieval event contingent on a strong cue–trace match and the availability of the relevant content.

evaluate this hypothesis, we analyze the roles of individual components of the attention mechanism in Transformer-based LLMs, an area for which grounded empirical evidence remains scarce.

Further guided by the Encoding Specificity Principle (ESP) (Tulving & Thomson, 1973), which posits that successful retrieval depends on the overlap between contextual cues present at encoding and those available at retrieval, we further proposed that in the context of LLMs for text generation tasks, these "contextual cues" are encoded as keywords by the attention mechanism. Figure 1 visualize our our hypotheses, illustrating the theorized roles of Q, K, and V in memory retrieval.

In this paper, we use empirical results to prove our hypothesis, with key contributions include: (1) demonstration that Q, K, and V play distinct roles paralleling human memory—**Q as context encoder, K as trace memory index, and V as content store** (2) empirically confirming that **contextual cues are represented as keywords**; and (3) identifying **consistent, model-specific attention neurons activated by keywords**, suggesting a potential location for contextual memory.

2 Related Work and Observation

2.1 Explainable Artificial Intellegence (XAI)

Numerous studies in XAI have examined explainability in transformer LLMs using various techniques, including input perturbation (Fong & Vedaldi, 2017), surrogate models (Ribeiro et al., 2016; Guidotti et al., 2019; Leemann et al., 2025), gradient-based methods (Simonyan et al., 2014; Shrikumar et al., 2017; Lundberg & Lee, 2017), and layer-wise relevance propagation (LRP) (Bach et al., 2015; Achtibat et al., 2024a).

Existing studies start to put specifical focus on attention layers for explainability (Bahdanau et al., 2014; Abnar & Zuidema, 2020). Hybrid approaches also exist, integrating attention mechanisms with other explainability methods to further enhance interpretability (Deiseroth et al., 2023; Achtibat et al., 2024a). The explanatory value of attention, however, remains debated (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019), with recent studies suggesting that attention can provide alternative insights but does not fully address all goals of model explainability (Bastings & Filippova, 2020; Lopardo et al., 2024).

Different from other XAI approaches, human-interpretable explanations for XAI aim to clarify transformer models using metrics or reasoning comprehensible to humans. However, existing work has mainly focused on Feed-Forward Neural Network (FFNN) layers. Geva et al. (2021) showed that

FFNN layers act as Key-Value memories, with Key matrices matching recurring linguistic patterns. Geva et al. (2022) further demonstrated that Value matrices encode and promote high-level concepts during prediction, as reflected in token distribution shifts.

Beyond much of the conventional XAI literature, which seeks comprehensive explanations involving grammar, syntax, and semantics, our work specifically targets memory mechanisms in transformer models, emphasizing memory as one distinct and influential component that contributes to the accurate and human-like responses from LLMs.

2.2 Attention mechanisms in the view of Computational Psycholinguistic

In computational psycholinguistic field, Van Dyke & Lewis (2003) proposed the cue-based retrieval theories, which can be described as when memory retrieval is initiated, available retrieval cues are matched against all possible candidates via a direct and parallel matching process. In other words, retrieval mechanism in sentence comprehension is *content-addressable*. Since the introduction of the Transformer architecture (Vaswani et al., 2017), several studies in this field have drawn explicit parallels between the attention mechanism and cue-based retrieval theories (Yoshida et al., 2025; Ryu & Lewis, 2021; Timkey & Linzen, 2023; Oh & Schuler, 2022). Nevertheless, integration between XAI and computational psycholinguistics remains limited: XAI primarily seeks mechanistic explanations of model internals, whereas computational psycholinguistics leverages LMs to formalize and evaluate theories of human language processing. These divergent objectives have constrained meaningful exchange despite clear conceptual overlap.

Until recently, Gershman et al. (2025) articulated a key–value memory framework that bridges machine learning, psychology, and neuroscience, with implications for computational psycholinguistics. In this formulation, inputs are transformed into two distinct representations—keys and values—that are stored in memory: keys serve as memory indices enabling content-addressable access, while values encode the memory content. Memory retrieval proceeds by forming a query and matching it against the stored keys; the resulting matches identify the corresponding relevant memory contents in values. This formulation aligns with cue-based retrieval theories in psycholinguistics and with the Transformer attention mechanism. To support this view, the authors introduce two toy models that emulate self-attention (with Q, K, and V) and feedforward neural network. These models show that keys and values are represented distinctively to align with their respective roles in memory retrieval, and that forgetting can reflect retrieval failure even when memory traces persist-an analogue of the tip-of-the-tongue phenomenon in human memory (Freedman & Landauer, 2014).

In this paper, we move beyond toy settings by conducting experiments on multiple LLMs to test for key–value memory in their attention mechanisms. We treat the Encoding Specificity Principle (ESP; (Tulving & Thomson, 1973)) as a unifying foundation for both cue-based retrieval and key–value memory: retrieval succeeds to the extent that retrieval cues overlap with features encoded at memorization step. We use ESP to guide our hypothesis of Transformer attention functionalities, which can be formalized as:

Hypothesis 1 . Q encodes retrieval cues, K indexes candidate traces by those cues, and V stores retrievable content.

2.3 Observation

Having outlined theoretical motivations from XAI and computational psycholinguistics, we now present observations of self-attention that motivate our hypothesis. Sukhbaatar et al. (2019) noted a close similarity between self-attention and FFNN, as both rely on dot-product–based linear projections. Building on this and recent FFNN explainability results (Geva et al., 2021; 2022), which characterize FF layers as performing pattern matching over learned keys and promoting associated concepts via their values, we ask whether attention layers enact an analogous process. Following the ESP-based mapping introduced above (Q as retrieval cues, K as indices, V as stored content), we treat attention weights as content-addressable matching and further hypothesize:

Hypothesis 2 . The retrieval cues are instantiated as salient lexical tokens (“keywords”) tied to the relevant memory.

Hypothesis 2 is consistent with findings by Eldan & Russinovich (2023), who propose an approximate unlearning method that replaces topic-specific keywords with generic placeholders. Scrubbing these lexical cues markedly impairs the model's ability to retrieve facts about the target topic. Despite the collateral degradation, the results indicate that such keywords function as critical retrieval cues in LLMs.

From a mathematical standpoint, the attention mechanism in language models was initially proposed as a similarity measurement (Bahdanau et al., 2014), due to the utilization of dot product calculation, where similar vectors yield higher values. Therefore, the implementation of self-attention with Q, K, and V (Vaswani et al., 2017) closely resembles a form of content-based lookup, where each components perform a dot product transformation on the initial input embeddings to create distinct representations for their unique roles aligning with Hypothesis 1. This view is shared by many past papers in machine learning field or greatly hinted at (Tay et al., 2021; Roy et al., 2021; Rohekar et al., 2023). In the equation for self-attention:

3 METHODOLOGY

In this section, we describe our two experiments designed to empirically prove Hypothesis 1 and 2. For both experiments, we employ six different auto-regressive (decoder-only) LLMs of varying sizes and structures: **Llama 2-7b**, **Llama 2-13b** (Touvron et al., 2023), **Llama 3.1-8b** (Grattafiori et al., 2024), **Olmo 2-1124-13b** (OLMo et al., 2025), **Qwen 2.5-14b** (Qwen et al., 2025), **Phi-4** (Abdin et al., 2024) and **GPT-Neox-20b** (Black et al., 2022).

3.1 EXPERIMENT 1: EMPIRICAL VALIDATION OF HYPOTHESIS 1 WITH ATTENTION SWAPPING

Experiment description: Self-attention in Transformer is calculated as:

$$\text{Attention} = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

According to the ESP and the key-value memory framework (Section 2.2), memory retrieval relies on two processes: Content-addressable matching (QK^T) and the availability of stored content (V). Depending on the successful of each process, we can have three different cases when models retrieve memory. **Case one**: matching works and stored content is present, representing successful memory retrieval. **Case two**: matching fails and stored content is present, representing normal forgetting or the tip-of-the-tongue phenomenon. **Case three**: matching works and stored content is not present, but what might this represents ? Intuitively, we conjecture that case three represents hallucination, where models identify a match but produce incorrect content. To this end, we conducted an experiment in which we interchange the Q, K, and V projections—either individually or in pairwise combinations—across prompts to reproduce **case two** and **three** and examine whether the corresponding phenomena emerge, hence empirically validating Hypothesis 1

Dataset: Because not all prompts permit objective quantification of hallucination in our experiment, we therefore employ the Counterfactual dataset (Meng et al., 2022) - a dataset comprises of factual and counterfactual examples that are structurally similar but contextually distinct, thereby eliciting different target outputs from the model that can be quantified.

Swapping is conducted on a pairwise basis. For each factual example in the dataset, we find all suitable counterfactual examples to create our swapping pair of prompts. Suitable counterfactual examples must have the same length (after tokenized) as their corresponding factual example for precise swapping of Q, K, and V projections. So, for a factual example x_f, its counterfactual example x_{cf}, and their corresponding $Q/K/V$, our swapping can be formulated as:

$$\text{Swapped Attention K}(x_f) = \text{softmax}\left(\frac{Q_{x_f}{K_{x_{cf}}}^T}{\sqrt{d_k}}\right)V_{x_f} \tag{2}$$

Note that our swapping only occurs for the input prompts, not the subsequent generated text. Thus, the projections computed for x_{cf} are not imposed on x_f after processing the input context. We allow

LLMs to revert back to their original $Q/K/V$ projections after first token is generated. This design constitutes a less intrusive intervention than swapping for the entirety of generation. Our targets for swapping include K, V, and KV. Swapping of Q or any other combinations are not required as they are implied by our three chosen targets (e.g., swapping Q is the same as swapping KV).

Metrics: Metrics used for this experiment include accuracy (for both factual and counterfactual labels), which is computed as first-word exact match to measure hallucination effect. Additionally, we compute the $\Delta logit$ and the perplexity overhead to assess the extent to which the swapping procedure induces hallucination. Both metrics are defined as differences between the model's outputs under the original and swapped conditions.

3.2 Experiment 2: Empirical validation of Hypothesis 2 with K matrix perturbation

Experiment description: Both Q and K encode retrieval cues that support content-addressable matching; hence, either component could, in principle, be manipulated to probe cue representations in a sentence-comprehension setting. However, because queries are evaluated against keys (rather than the reverse), we hypothesize that K is the more informative target for intervention.

Dataset: To evaluate Hypothesis 2, the experiment requires datasets with clearly interpretable contexts and recognizable by unambiguous lexical cues. Accordingly, we use long-form book text drawn from publicly available corpora - Project Gutenberg (n.d.). The dataset is perfect for this experiment as each book has its own unique storyline, characters, and places.

Raw text of each book is processed with input/label pairs generated by a sliding window technique (step size 30, input: 512 tokens, output: 40 tokens) to create a dataset $\mathbb{D}_i$. Input and label combined form complete sentence(s) in the books. Furthermore, to best represent the model's memory, we additionally filter for verbatim examples using ROUGE-L Recall from the ROUGE suite (Lin, 2004), a metric widely used to measure overlap between generated texts, and commonly adopted/built upon for evaluating unlearning effectiveness (Jang et al., 2023; Carlini et al., 2023). The final result yield $\mathbb{A}_i$ for $i \in 1, \ldots, n$, where n is the number of books:

$$\mathbb{A}_i = \{a_1, a_2, \ldots, a_n\}, \quad where : \text{ROUGE-L Recall}(a_j) = 1, \forall a_j \in \mathbb{D}_i \tag{3}$$

Note that the set of $\mathbb{A}_i$ for each model is different since not all models share the same memory about the same book. (The specific list of books used for each model, along with their ROUGE-L Recall sample sizes, can be found in Appendix B.)

Following Eldan & Russinovich (2023), we use ChatGPT-4o to identify *anchored terms* - AT_i (each dataset $\mathbb{A}_i$ has a corresponding AT_i) as keywords for each book to test our hypothesis (see Appendix A for our exact prompt). Let $x = w_1, \ldots, w_n$ denote an input, where each w is a word in x. We extract the K projected values (from Equation 1) for all $w \in AT_i$ at a given layer l and head h. The dataset-level coefficient for each layer-head pair is the mean across all inputs:

$$\mathrm{m}_{l-h}^{\mathbb{A}_i} = \frac{1}{n} \sum_{i=1}^{n} \left\{ e_j W_{l-h}^{K} \mid w_j \in AT_i \right\} \tag{4}$$

where n is the total number of inputs in dataset $\mathbb{A}_i$ and e is the embedding vector for word w. To identify the most significant layers or attention heads, we aggregate and average these scores across heads or layers. This approach also enables direct ranking of hidden units (layer-head-dimension triplets).

Our method of identifying keywords is based on observed activation patterns of top neurons in response to GPT-4o-generated keywords (Equation 4). Keywords are identified by examining top words weighted by the learned weight matrix W^K. To produce the final list of keywords for a dataset, we aggregate the scores of top words across all inputs:

$$S_{\mathbb{A}_i}(t) = \sum_{x \in \mathbb{A}_i} \left\{ \frac{1}{N_x(w)} \sum_{j=1}^{N_x(t)} \mathrm{score}_{x,j}(w) \right\} \tag{5}$$

where $N_x(w)$ is the number of occurrences of word w in x; $\mathrm{score}_{x,j}(t)$ refers to the sum of key projection scores τW^K over all sub-word tokens τ (since LLMs utilize sub-word tokenization) that constitute the j-th occurrence of w in input x.

We empirically assess the effect of our identified keywords on model memory, benchmarking against keywords produced by GPT-4o and by a state-of-the-art attention-informed XAI method, Layer-wise Relevance Propagation eXplains Transformers (LXT) (Achtibat et al., 2024b). We fix the budget at 20 keywords per method. For LXT, we collect the top-relevance tokens per input and aggregate them as in Equation 5 to obtain a top-keyword list for each book dataset. We exclude the first and last tokens for all inputs because LXT systematically assigns the highest and lowest relevance scores to the final and initial tokens, respectively, which would otherwise yield keyword lists dominated by end-of-input tokens. This behavior highlights a difference in goals between our work and prior XAI approaches, as discussed at the end of Section 2.1. Nevertheless, we report LXT results to contextualize and further support our findings.

A simple perturbation at K for identified keywords by setting their projected values to 0:

$$\forall w \in x, \quad w \in \text{AT}i \implies eW_{\alpha}^{K} = 0 \tag{6}$$

where α can be specific layers, heads, layer-head pairs, or layer-head-dimension depending on the desired level of granualarities. This perturbation is equivalent to none of the words in the input attending to identified keywords at selected neurons. We focus on perturbing specific attention heads, based on the intuition that certain heads may be responsible for memory mechanisms, as prior work has shown that individual heads often serve distinct functions (Voita et al., 2019; Clark et al., 2019; Vig, 2019). We compare perturbation outcomes against both the unperturbed baseline and perturbations applied to randomly selected non-keyword tokens. For each input, we sample as many random tokens as there are identified keywords. Because interventions on K are expected to affect model behavior to some extent, the randomized control estimates the impact of indiscriminate perturbations. If targeting extracted keywords produces larger deviations from baseline than targeting random tokens, this indicates that the identified terms function as retrieval cues for content-addressable matching.

Top heads identified by Equation 5 are selected for perturbation. To ensure proportionality across model sizes, we select 2 heads for **Llama 2-7b**, 3 for **Llama 2-13b**, and 4 for **GPT-Neox-20b** models. For **Llama 3.1-8b**, **Qwen 2.5-14b**, and **Phi-4** (14b), we select 1, 2, and 2 respectively due to their architectural design that employs Multi-Query Attention (MQA) or Group Query Attention (GQA), where attention heads are grouped together for faster training. We exclude the first attention head if it appears among a model's top-ranked heads, replacing it with the next-ranked head. This choice is motivated by prior work showing that the first head functions as an induction head that facilitates information flow to subsequent heads rather than supporting memory-specific computations (Mușat, 2025). The only exception to the rest is **Olmo 2-1124-13b**, where we perturb for all heads across all layers, and more will be discussed in Section 4.2 about this decision. Metrics used for evaluation are: ROUGE-L Recall, Perplexity, BERTScore (Zhang* et al., 2020), Repetition rate, and MAUVE (Pillutla et al., 2023). To facilitate comparison, all metrics are normalized to the unperturbed baseline to emphasize proportional changes.

Metrics: We assess impact of keyword to memory recall ability with ROUGE-L Recall and BERTScore, the latter leveraging contextual embeddings to quantify semantic similarity. General generation capabilities under perturbation is evaluated using perplexity, repetition rate, and MAUVE. MAUVE is an unsupervised evaluation framework that quantifies the distributional similarity between model-generated and human-written text. Repetition rate is our custom metric that measures the repetitiveness of generated text based on n-gram overlapping rate:

$$\text{repetition rate} = 1 - \frac{|\text{unique n-gram}|}{|\text{total n-gram}|} \tag{7}$$

4 Results

4.1 Experiment 1: Attention swapping

Swapping V significantly increases hallucinations. Figure 2 shows the results for **Experiment 1**. To our surprise, swapping V alone induces LLMs to hallucinate and answer as if they were prompted with samples from counterfactual set for minimum 50% in **Olmo-2** and maximum 90% in **Qwen2.5** (Green bars in the upper-right subfigure). Across models, the average perplexity overhead remains minimal, and the mean $\Delta logit$ is positive for most models; the exceptions are **GPT-NeoX-20B** and

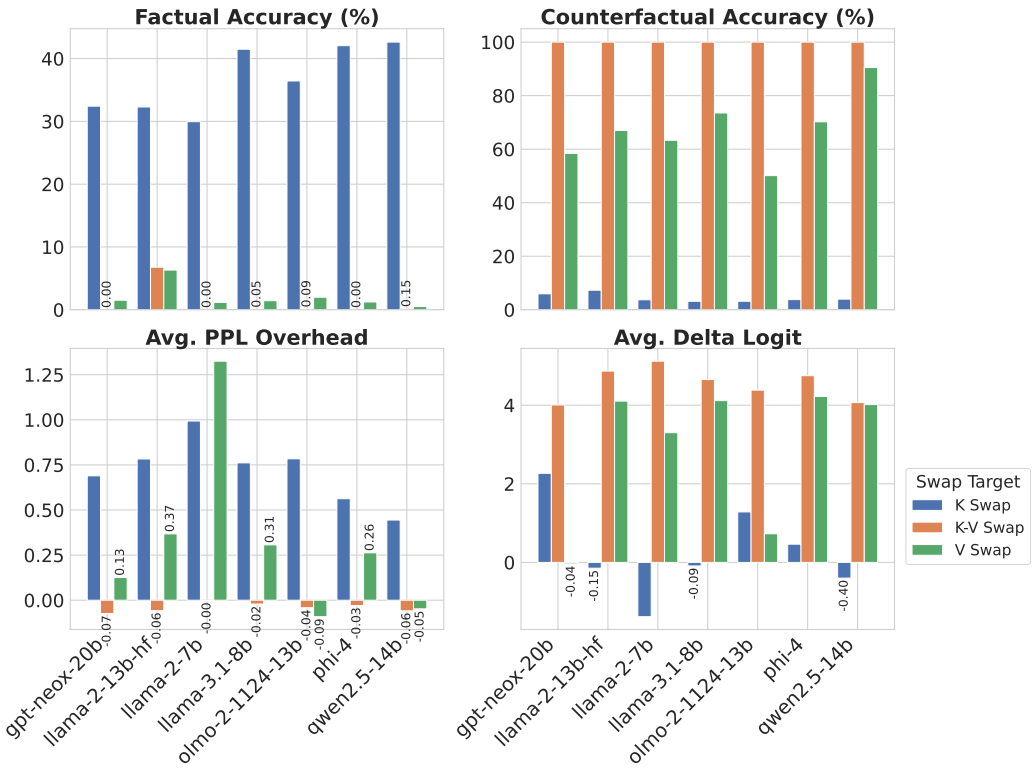

Figure 2: QKV swapping experiment results.

OLMo-2, which also exhibit the lowest counterfactual accuracy. Overall, these findings indicate a clean replacement of knowledge achieved by swapping V. The result aligns with key-value memory framework and strengthen our view in Hypothesis 1, where V plays the role of content storage.

Swapping both K and V induces LLMs to hallucinate for 100% of input factual prompts across all experimented models (Orange bars in the upper-right subfigure). Recall that Q and K are hypothesized to implement content-addressable matching via the dot-product calculation (similarity measurement - see Section 2.3). When K is replaced with K_{cf}, the model computes attention between the factual queries (Q_f) and counterfactual keys (K_{cf}), effectively performing nearest-neighbor search in the counterfactual key space. This steers the address toward counterfactual memory slots; because the corresponding values (V_{cf}) are also supplied, the retrieved content is counterfactual, yielding systematic hallucination. This also explains the improvement (though minimum) in perplexity and the consistent positive $\Delta logit$ when swapping both K and V.

On the other hand, **swapping K only does not induce hallucination but only hinder model's ability to recall factual knowledge** where the factual accuracy across models is around 30%-40%. The results in this swapping setting strengthen the views from computational psycholinguistics and psychology about attention mechanism. Specifically, with V intact, the memory content must be available but models "forget" due to retrieval failure.

4.2 Experiment 2: K matrix perturbation

Consistent top neurons activated when prompting keywords. Recall that each book dataset is associated with a unique set of GPT-4o-generated keywords. Figure 3 ranks the neurons (layer–head–dimension triplets) most activated by keywords. For each model, we compute an average reciprocal rank score across the book datasets, weighting rank r by $1/r$; thus, the second-ranked neuron contributes half the score of the first. The results reveal, for each model, a single dominant neuron whose score significantly exceeds that of the second-ranked neuron. The same observation

Table 1: Keyword lists by extraction method for Alice's adventure in wonderland book.

GPT-4o Generated	LXT	Our method
alice, rabbit-hole, sister, bank, daisies, white rabbit, waistcoat-pocket, orange marmalade, dinah, schoolroom, new zealand, australia, latitude, longitude, cheshire, duchess, caucus-race, queen, mock turtle, lobster quadrille	be, alice, might, an, had, thought, in, time, her, moment, by, said, barrowful, this, could, atom, she, as, what, down	alice, dormouse, whiskers, sneezing, mushroom, hastily, caterpillar, angrily, hurry, aloud, dinah, melancholy, queer, timidly, lefthand, nursing, frightened, fountains, rabbit, sleepy

is true for higher levels of granularity (see Appendix C for a detailed ranking of neurons across available book dataset for each model). The results indicate the presence of unique neurons within each LLM that show strong evidence of encoding/consolidating contextual memory in the form of keywords.

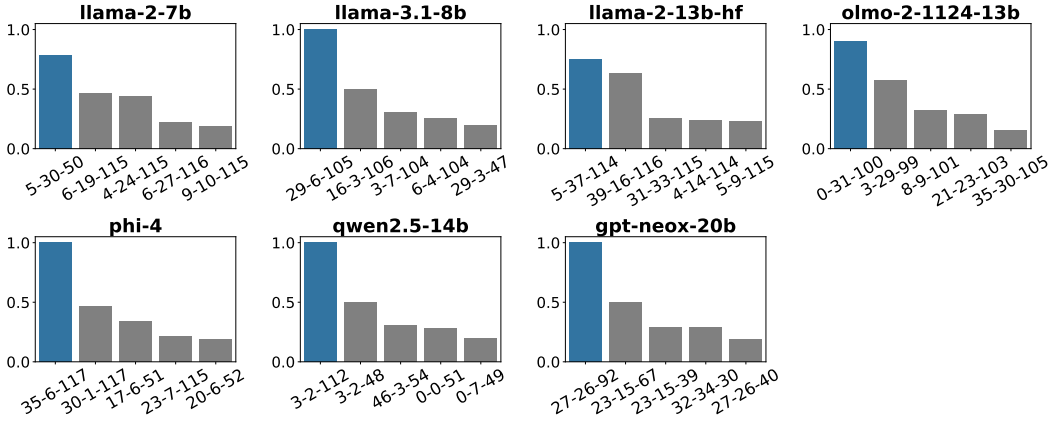

Figure 3: Mean reciprocal rank of layer-head-dimension for each model across its respective book datasets.

Perturbing linearly projected keywords at *Key matrix weights* impairs memory retrieval. We observe that our generated keywords can be easily associated with the relevant topic and semantically comparable to the ones generated by GPT-4o. Table 1 shows a comparison of keywords extracted by different methods for **Llama-2-7b**.

Figure 4 presents the results of the perturbation experiments, where each column shows the result of one metric, upper and lower row show the between perturbation of extracted keywords and randomly selected words respectively. We also show the standard deviation band, with the exception of MAUVE. Because MAUVE requires a large sample size for reliable estimation, we compute it by pooling all inputs across datasets $\mathbb{A}_i$ available to each model, rather than computing it per input; consequently, standard deviation band is not available for MAUVE.

Perturbations using LXT-extracted keywords yield the largest reduction in memorization as measured by ROUGE-L recall and BERT-Score, followed by our method and the GPT-4o–generated method. However, MAUVE and the repetition rate degrade markedly under LXT relative to the baseline and the other methods. The standard-deviation band for LXT on the repetition-rate metric is especially wide—particularly for the Llama-2 family—whose lower bound approaches zero. Because our method perturbs every keyword detected in an input, the perturbation size can become large when the extracted list contains common, high-frequency words; the scope is therefore input-dependent. This suggests that XAI methods such as LXT tend to identify tokens predictive of the model's output but do not reliably isolate content-specific, context-bearing keywords.

Compared with random word selection (second row), all methods achieve greater reductions in memorization as measured by ROUGE-L recall, except for **Llama-2-7B** and **Llama-2-13B**, for which ROUGE-L recall is nearly indistinguishable across the three keyword-extraction approaches.

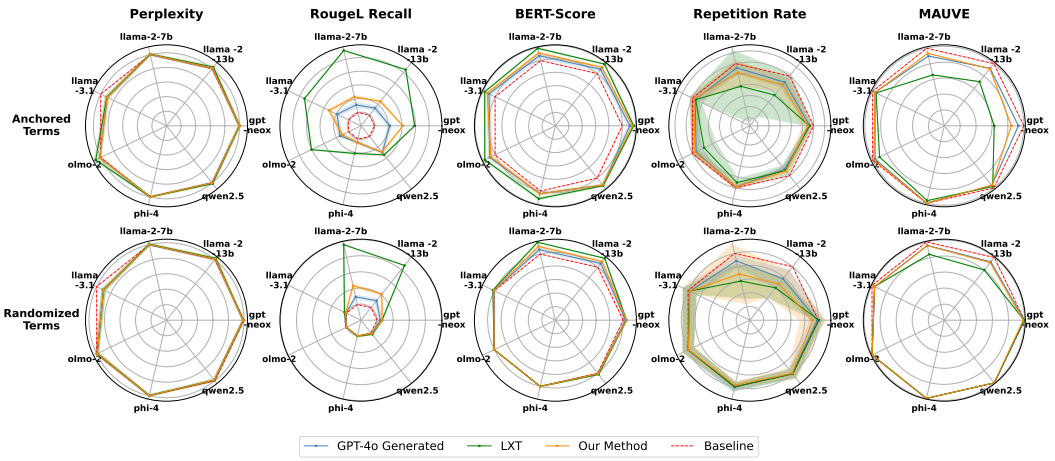

Figure 4: Radar graphs showingh average overall performance for all evaluated models when perturbed with different methods (individual book results can be found in Appendix D). All results are normalized to show higher is better.

We argue that this behavior, unique to Llama 2 family models is due to the small vocabulary size of 32.000 compared to its successor **Llama 3.1** with 128.256 (about four times larger).

Crucially, we find that an extreme case of using keywords as contextual cues for content-addressable matching in **Olmo 2-1124-13b**. Specifically, perturbing at ALL attention heads for keywords sharply reduces ROUGE-L Recall and BERTScore, while random word perturbations have minimal effect. This suggests that the model relies almost exclusively on keywords for memory retrieval, indicating minimal feature superposition for memory features. In summary, the results validate Hypothesis 2 by showing that neglecting keywords in attention layers significantly impairs the memory retrieval capabilities of LLMs. Presenting retrieval cues used in memory retrieval process of Transformers as salient lexical tokens. Additionally, we identify specific neurons encode keyword-related memory, enabling dynamic keyword extraction for a given topic.

5 Conclusion

We propose that transformer-based LLMs recall memory in accordance with the Encoding Specificity Principle. Through extensive empirical analysis, we demonstrate that the internal mechanisms of attention align with human memory processes during sentence comprehension, consistent with cue-based retrieval theories and the key–value memory framework.

We show that the Transformer's attention mechanism leverages the key-projection (K) weights to index memory traces via cues encoded in salient lexical tokens, thereby facilitating retrieval. We further identify model-specific units whose activations are selectively driven by these keywords, suggesting a role as context-addressable memory locations. Building on these results, we present a method to extract the keywords that index a remembered context, thereby enabling downstream applications such as unlearning.

Limitation. Our perturbation method offers a minimally invasive, context-specific approach to machine unlearning. Its main limitation is naivety: the choices of top neurons and the number of targeted keywords remain largely arbitrary. Future work should optimize selection criteria and modification strategies, moving beyond simple zeroing.

On the other hand, our method of extracting keywords also has limitations that originate from compound words or terms made up of more than 1 word. For example, "White rabbit" is a better keyword than simply "rabbit", but our method cannot treat "white" and "rabbit" together as a single term. As a result, the list of keywords extracted by our method does not fully capture the ideal set of contextual cues.

REFERENCES

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. Phi-4 technical report, 2024. URL https://arxiv.org/abs/2412.08905.

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4190–4197, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.385. URL https://aclanthology.org/2020.acl-main.385/.

Reduan Achtibat, Sayed Mohammad Vakilzadeh Hatefi, Maximilian Dreyer, Aakriti Jain, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. Attnlrp: attention-aware layer-wise relevance propagation for transformers. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024a.

Reduan Achtibat, Sayed Mohammad Vakilzadeh Hatefi, Maximilian Dreyer, Aakriti Jain, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. AttnLRP: Attention-aware layer-wise relevance propagation for transformers. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 135–168. PMLR, 21–27 Jul 2024b. URL https://proceedings.mlr.press/v235/achtibat24a.html.

Jane Austen. *Pride and Prejudice*. Project Gutenberg, 1813. URL https://www.gutenberg.org/ebooks/1342.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*, 10, 2015. URL https://api.semanticscholar.org/CorpusID:9327892.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2014. URL https://api.semanticscholar.org/CorpusID:11212020.

J. M. Barrie. *Peter Pan*. Project Gutenberg, 1920. URL https://www.gutenberg.org/ebooks/16.

Jasmijn Bastings and Katja Filippova. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? In Afra Alishahi, Yonatan Belinkov, Grzegorz Chrupała, Dieuwke Hupkes, Yuval Pinter, and Hassan Sajjad (eds.), *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 149–155, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.blackboxnlp-1.14. URL https://aclanthology.org/2020.blackboxnlp-1.14/.

L. Frank Baum. *The Wonderful Wizard of Oz*. Project Gutenberg, 1900. URL https://www.gutenberg.org/ebooks/55.

Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, Usvsn Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. GPT-NeoX-20B: An open-source autoregressive language model. In Angela Fan, Suzana Ilic, Thomas Wolf, and Matthias Gallé (eds.), *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pp. 95–136, virtual+Dublin, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.bigscience-1.9. URL https://aclanthology.org/2022.bigscience-1.9/.

Emily Brontë. *Wuthering Heights*. Project Gutenberg, 1847. URL https://www.gutenberg.org/ebooks/768.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=TatRHT_1cK.

Lewis Carroll. *Alice's Adventures in Wonderland*. Project Gutenberg, 1865. URL https://www.gutenberg.org/ebooks/11.

Min Chen, Zhikun Zhang, Tianhao Wang, Michael Backes, Mathias Humbert, and Yang Zhang. When machine unlearning jeopardizes privacy. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, CCS '21, pp. 896–911, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384544. doi: 10.1145/3460120.3484756. URL https://doi.org/10.1145/3460120.3484756.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? an analysis of BERT's attention. In Tal Linzen, Grzegorz Chrupała, Yonatan Belinkov, and Dieuwke Hupkes (eds.), *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 276–286, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4828. URL https://aclanthology.org/W19-4828/.

Stéphanie Daumas, Hélène Halley, Bernard Francés, and Jean-Michel Lassalle. Encoding, consolidation, and retrieval of contextual memory: Differential involvement of dorsal ca3 and ca1 hippocampal subregions. *Learning & Memory*, 12(4):375–382, 2005. doi: 10.1101/lm.81905.

Björn Deiseroth, Mayukh Deb, Samuel Weinbach, Manuel Brack, Patrick Schramowski, and Kristian Kersting. ATMAN: Understanding transformer predictions through memory efficient attention manipulation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=PBpEb86bj7.

Charles Dickens. *Oliver Twist*. Project Gutenberg, 1838. URL https://www.gutenberg.org/ebooks/730.

Arthur Conan Doyle. *The Adventures of Sherlock Holmes*. Project Gutenberg, 1892. URL https://www.gutenberg.org/ebooks/1661.

Ronen Eldan and Mark Russinovich. Who's harry potter? approximate unlearning in llms, 2023. URL https://arxiv.org/abs/2310.02238.

EU Commission. General data protection regulation. *Official Journal of the European Union*, 59: 1–88, 2016. URL https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng.

F. Scott Fitzgerald. *The Great Gatsby*. Project Gutenberg, 1925. URL https://www.gutenberg.org/ebooks/64317.

Ruth C. Fong and Andrea Vedaldi. Interpretable Explanations of Black Boxes by Meaningful Perturbation . In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 3449–3457, Los Alamitos, CA, USA, October 2017. IEEE Computer Society. doi: 10.1109/ICCV.2017.371. URL https://doi.ieeecomputersociety.org/10.1109/ICCV.2017.371.

J. Freedman and T. Landauer. Retrieval of long-term memory: "tip-of-the-tongue" phenomenon. *Psychonomic Science*, 4:309–310, 08 2014. doi: 10.3758/BF03342310.

Samuel J. Gershman, Ila Fiete, and Kazuki Irie. Key-value memory in the brain. *Neuron*, 113(11):1694–1707.e1, 2025. ISSN 0896-6273. doi: https://doi.org/10.1016/j.neuron.2025.02.029. URL https://www.sciencedirect.com/science/article/pii/S0896627325001722.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.446. URL https://aclanthology.org/2021.emnlp-main.446/.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 30–45, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.3. URL https://aclanthology.org/2022.emnlp-main.3/.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel,

Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Riccardo Guidotti, Anna Monreale, Fosca Giannotti, Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. Factual and counterfactual explanations for black box decision making. *IEEE Intelligent Systems*, 34(6):14–23, 2019. doi: 10.1109/MIS.2019.2957223.

Axel Guskjolen and Mark S. Cembrowski. Engram neurons: Encoding, consolidation, retrieval, and forgetting of memory. *Molecular Psychiatry*, 28:3207–3219, 2023. doi: 10.1038/s41380-023-02137-5. URL https://www.nature.com/articles/s41380-023-02137-5.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=EldbUlZtbd.

Homer. *The Odyssey*. Project Gutenberg, 2008. URL https://www.gutenberg.org/ebooks/1727.

Sarthak Jain and Byron C. Wallace. Attention is not Explanation. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3543–3556, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1357. URL https://aclanthology.org/N19-1357/.

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. Knowledge unlearning for mitigating privacy risks in language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14389–14408, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.805. URL https://aclanthology.org/2023.acl-long.805/.

Tobias Leemann, Alina Fastowski, Felix Pfeiffer, and Gjergji Kasneci. Attention mechanisms don't learn additive models: Rethinking feature importance for transformers. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=yawWz4qWkF.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Gianluigi Lopardo, Frédéric Precioso, and Damien Garreau. Attention meets post-hoc interpretability: A mathematical perspective. In *ICML*, 2024. URL https://openreview.net/forum?id=wnkC5T11Z9.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 4768–4777, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary Chase Lipton, and J Zico Kolter. TOFU: A task of fictitious unlearning for LLMs. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=B41hNBoWLo.

Herman Melville. *Moby Dick*. Project Gutenberg, 1851. URL https://www.gutenberg.org/ebooks/2701.

Kevin Meng, David Bau, Alex J Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=-h6WAS6eE4.

Tiberiu Mușat. Mechanism and emergence of stacked attention heads in multi-layer transformers. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=rUC7tHecSQ.

Byung-Doh Oh and William Schuler. Entropy- and distance-based predictors from GPT-2 attention patterns predict reading times over and above GPT-2 surprisal. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 9324–9334, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.632. URL https://aclanthology.org/2022.emnlp-main.632/.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 olmo 2 furious, 2025. URL https://arxiv.org/abs/2501.00656.

Krishna Pillutla, Lang Liu, John Thickstun, Sean Welleck, Swabha Swayamdipta, Rowan Zellers, Sewoong Oh, Yejin Choi, and Zaid Harchaoui. Mauve scores for generative models: Theory and practice. *Journal of Machine Learning Research*, 24(356):1–92, 2023. URL http://jmlr.org/papers/v24/23-0023.html.

Project Gutenberg. Project Gutenberg: Free eBooks, n.d. URL https://www.gutenberg.org. Accessed: 2025-04-02.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Marco Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In John DeNero, Mark Finlayson, and Sravana Reddy (eds.), *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pp. 97–101, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-3020. URL https://aclanthology.org/N16-3020/.

Raanan Y. Rohekar, Yaniv Gurwicz, and Shami Nisimov. Causal interpretation of self-attention in pre-trained transformers. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Aurko Roy, Mohammad Saffar, Ashish Vaswani, and David Grangier. Efficient content-based sparse attention with routing transformers. *Transactions of the Association for Computational Linguistics*, 9:53–68, 2021. doi: 10.1162/tacl_a_00353. URL https://aclanthology.org/2021.tacl-1.4/.

Soo Hyun Ryu and Richard Lewis. Accounting for agreement phenomena in sentence comprehension with transformer language models: Effects of similarity-based interference on surprisal and attention. In Emmanuele Chersoni, Nora Hollenstein, Cassandra Jacobs, Yohei Oseki, Laurent Prévot, and Enrico Santus (eds.), *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pp. 61–71, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.cmcl-1.6. URL https://aclanthology.org/2021.cmcl-1.6/.

Mary Shelley. *Frankenstein*. Project Gutenberg, 1818. URL https://www.gutenberg.org/ebooks/84.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4222–4235, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.346. URL https://aclanthology.org/2020.emnlp-main.346/.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 3145–3153. JMLR.org, 2017.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In Yoshua Bengio and Yann LeCun (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6034.

Bram Stoker. *Dracula*. Project Gutenberg, 1897. URL https://www.gutenberg.org/ebooks/345.

Sainbayar Sukhbaatar, Edouard Grave, Guillaume Lample, Herve Jegou, and Armand Joulin. Augmenting self-attention with persistent memory, 2019. URL https://arxiv.org/abs/1907.01470.

Yi Tay, Dara Bahri, Donald Metzler, Da-Cheng Juan, Zhe Zhao, and Che Zheng. Synthesizer: Rethinking self-attention for transformer models. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10183–10192. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/tay21a.html.

William Timkey and Tal Linzen. A language model with limited memory capacity captures interference in human sentence processing. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 8705–8720, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.582. URL https://aclanthology.org/2023.findings-emnlp.582/.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

Endel Tulving and Donald M. Thomson. Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, 80(5):352–373, 1973. doi: 10.1037/h0020071.

Mark Twain. *The Adventures of Tom Sawyer*. Project Gutenberg, 1876. URL https://www.gutenberg.org/ebooks/74.

Julie A Van Dyke and Richard L Lewis. Distinguishing effects of structure and decay on attachment and repair: A cue-based parsing account of recovery from misanalyzed ambiguities. *Journal of Memory and Language*, 49(3):285–316, 2003. ISSN 0749-596X. doi: https://doi.org/10.1016/S0749-596X(03)00081-0. URL https://www.sciencedirect.com/science/article/pii/S0749596X03000810.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Jesse Vig. A multiscale visualization of attention in the transformer model. In Marta R. Costa-jussà and Enrique Alfonseca (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 37–42, Florence, Italy, July

2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-3007. URL https://aclanthology.org/P19-3007/.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5797–5808, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1580. URL https://aclanthology.org/P19-1580/.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 11–20, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1002. URL https://aclanthology.org/D19-1002/.

Heng Xu, Tianqing Zhu, Lefeng Zhang, Wanlei Zhou, and Philip S. Yu. Machine unlearning: A survey. *ACM Comput. Surv.*, 56(1), August 2023. ISSN 0360-0300. doi: 10.1145/3603620. URL https://doi.org/10.1145/3603620.

Jin Yao, Eli Chien, Minxin Du, Xinyao Niu, Tianhao Wang, Zezhou Cheng, and Xiang Yue. Machine unlearning of pre-trained large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8403–8419, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.457. URL https://aclanthology.org/2024.acl-long.457/.

Ryo Yoshida, Shinnosuke Isono, Kohei Kajikawa, Taiga Someya, Yushi Sugimoto, and Yohei Oseki. If attention serves as a cognitive model of human memory retrieval, what is the plausible memory representation? In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9795–9812, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.483. URL https://aclanthology.org/2025.acl-long.483/.

Charles Yu, Sullam Jeoung, Anish Kasi, Pengfei Yu, and Heng Ji. Unlearning bias in language models by partitioning gradients. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 6032–6048, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.375. URL https://aclanthology.org/2023.findings-acl.375/.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. Explainability for large language models: A survey. *ACM Trans. Intell. Syst. Technol.*, 15(2), February 2024. ISSN 2157-6904. doi: 10.1145/3639372. URL https://doi.org/10.1145/3639372.

A GPT4-O PROMPT TO EXTRACT ANCHORED TERMS

```
[You are a helpful assistant. A long passage of text will be provided to you.
Your task is to extract a list of 20 (in total) expressions, names or
entities which are idiosyncratic to the text (try your best to keep in
between 1-2 words, THE SHORTER THE BETTER). Please extract exactly how
they appear in the text but in lowercase.]
```

Table 2: 1.0 ROUGE-L Recall samples count for all models.

Model	No. Samples (AT)
llama-2-13b	
Alice's Adventures in Wonderland (Carroll, 1865)	278
Dracula (Stoker, 1897)	218
Frankenstein (Shelley, 1818)	513
The Great Gatsby (Fitzgerald, 1925)	89
Moby-Dick (Melville, 1851)	41
Pride and Prejudice (Austen, 1813)	327
The Adventures of Sherlock Holmes (Doyle, 1892)	583
llama-2-7b	
Alice's Adventures in Wonderland (Carroll, 1865)	78
Frankenstein (Shelley, 1818)	51
The Great Gatsby (Fitzgerald, 1925)	43
Pride and Prejudice (Austen, 1813)	83
The Adventures of Sherlock Holmes (Doyle, 1892)	62
olmo-2-1124-13b	
Alice's Adventures in Wonderland (Carroll, 1865)	296
The Great Gatsby (Fitzgerald, 1925)	38
Moby-Dick (Melville, 1851)	66
The Odyssey (Homer, 2008)	84
Peter Pan (Barrie, 1920)	296
The Adventures of Tom Sawyer (Twain, 1876)	743
The Wonderful Wizard of Oz (Baum, 1900)	528
gpt-neox-20b	
Alice's Adventures in Wonderland (Carroll, 1865)	38
Frankenstein (Shelley, 1818)	113
Moby-Dick (Melville, 1851)	38
The Adventures of Sherlock Holmes (Doyle, 1892)	51
phi-4	
The Great Gatsby (Fitzgerald, 1925)	39
Pride and Prejudice (Austen, 1813)	86
Frankenstein (Shelley, 1818)	214
Moby-Dick (Melville, 1851)	36
Alice's Adventures in Wonderland (Carroll, 1865)	50
llama-3.1-8b	
The Adventures of Tom Sawyer (Twain, 1876)	82
Wuthering Heights (Brontë, 1847)	44
Alice's Adventures in Wonderland (Carroll, 1865)	811
Moby-Dick (Melville, 1851)	106
The Great Gatsby (Fitzgerald, 1925)	167
Dracula (Stoker, 1897)	347
The Odyssey (Homer, 2008)	48
The Adventures of Sherlock Holmes (Doyle, 1892)	253
Frankenstein (Shelley, 1818)	466
Pride and Prejudice (Austen, 1813)	458
The Wonderful Wizard of Oz (Baum, 1900)	54
Oliver Twist (Dickens, 1838)	37
qwen2.5-14b	
Alice's Adventures in Wonderland (Carroll, 1865)	120
Pride and Prejudice (Austen, 1813)	40
The Great Gatsby (Fitzgerald, 1925)	31
Frankenstein (Shelley, 1818)	36
Moby-Dick (Melville, 1851)	34

B MODELS AND THEIR DATASETS

See Table 2

C TOP MEMORY COEFFICIENT NEURONS FOR ALL OTHER MODELS

Figure 5, Figure 6, Figure 7, Figure 8, Figure 9, Figure 10, and Figure 11 show the top memory coefficient neurons for **Llama 2-7b**, **Llama 2-13b**, **Olmo 2-13b**, **GPT Neox 20b**, **Llama 3.1-8b**, **Phi 4**, and **Qwen 2.5-14b** respectively.

D EVALUATION OF INDIVIDUAL BOOKS FOR ALL MODELS

Figure 12: **Llama 2-7b**, Figure 13: **Llama 2-13b**, Figure 14: **Olmo 2-13b**, Figure 15: **GPT Neox-20b**, Figure 16: **Llama 3.1-8b**, Figure 17: **Phi 4**, and Figure 18: **Qwen 2.5-14b**.

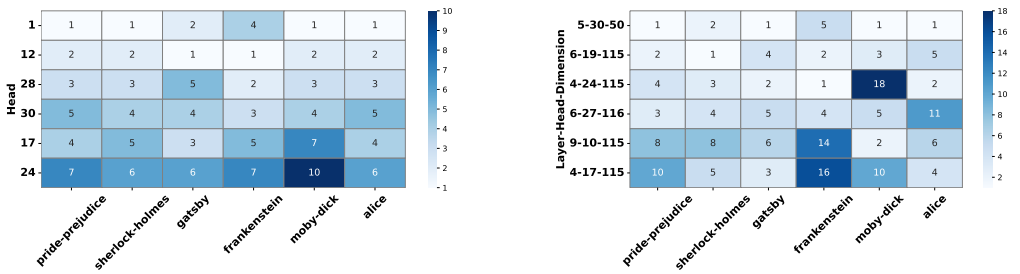

Figure 5: Top Memory Coefficient for **Llama 2-7b**. (**Left**): Top attention heads (**Right**): Top layer-head-dimension triplets

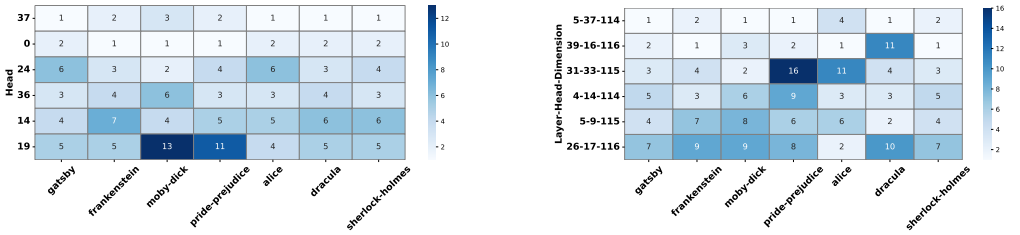

Figure 6: Top Memory Coefficient for **Llama 2-13b**. (**Left**): Top attention heads (**Right**): Top layer-head-dimension triplets.

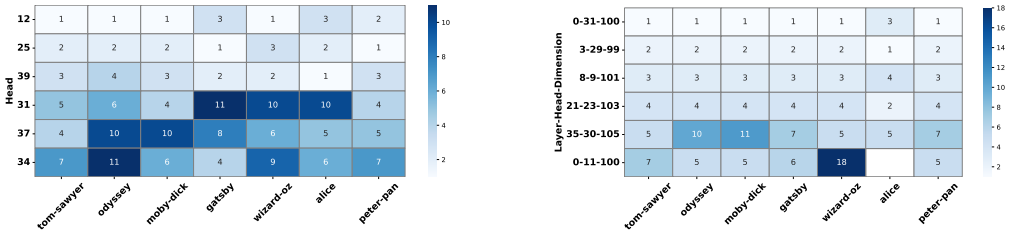

Figure 7: Top Memory Coefficient for **Olmo 2-13b**. (**Left**): Top attention heads (**Right**): Top layer-head-dimension triplets.

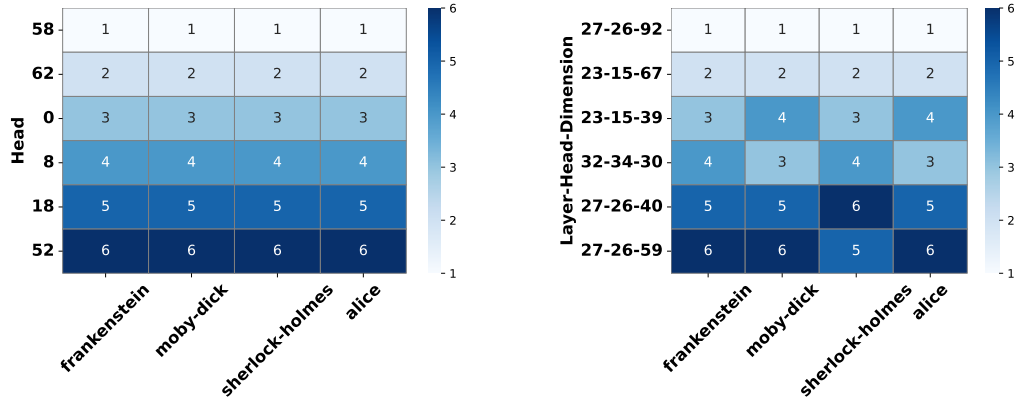

Figure 8: Top Memory Coefficient for **GPT-Neox**. (**Left**): Top attention heads (**Right**): Top layer-head-dimension triplets.

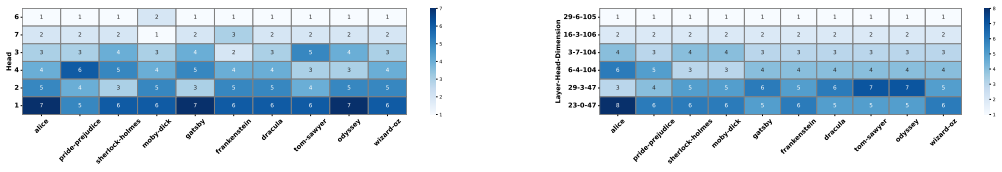

Figure 9: Top Memory Coefficient for **Llama 3.1-8b**. (**Left**): Top attention heads (**Right**): Top layer-head-dimension triplets

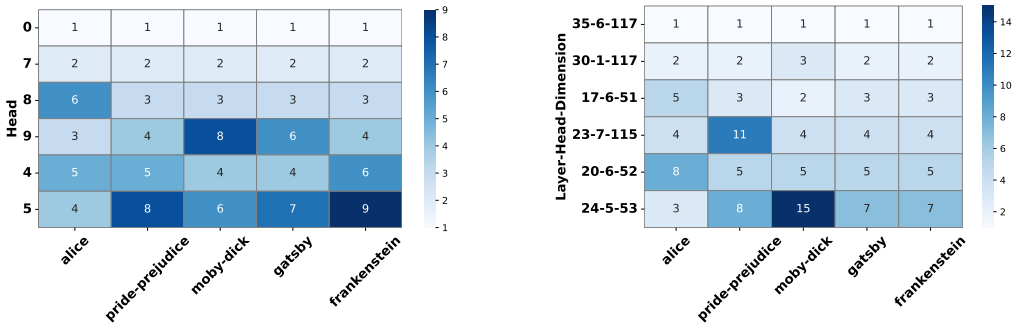

Figure 10: Top Memory Coefficient for **Phi 4**. (**Left**): Top attention heads (**Right**): Top layer-head-dimension triplets

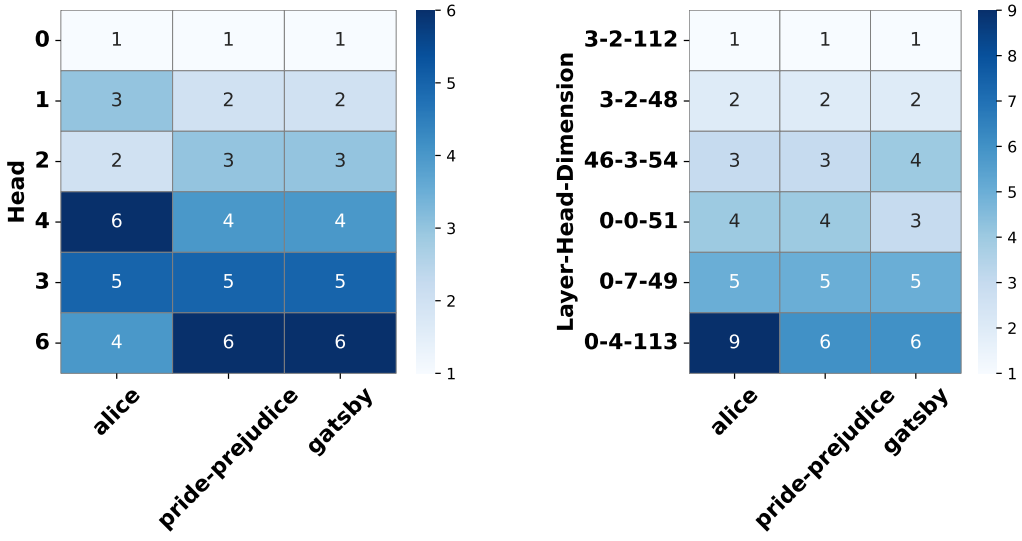

Figure 11: Top Memory Coefficient for **Qwen 2.5-14b**. (**Left**): Top attention heads (**Right**): Top layer-head-dimension triplets

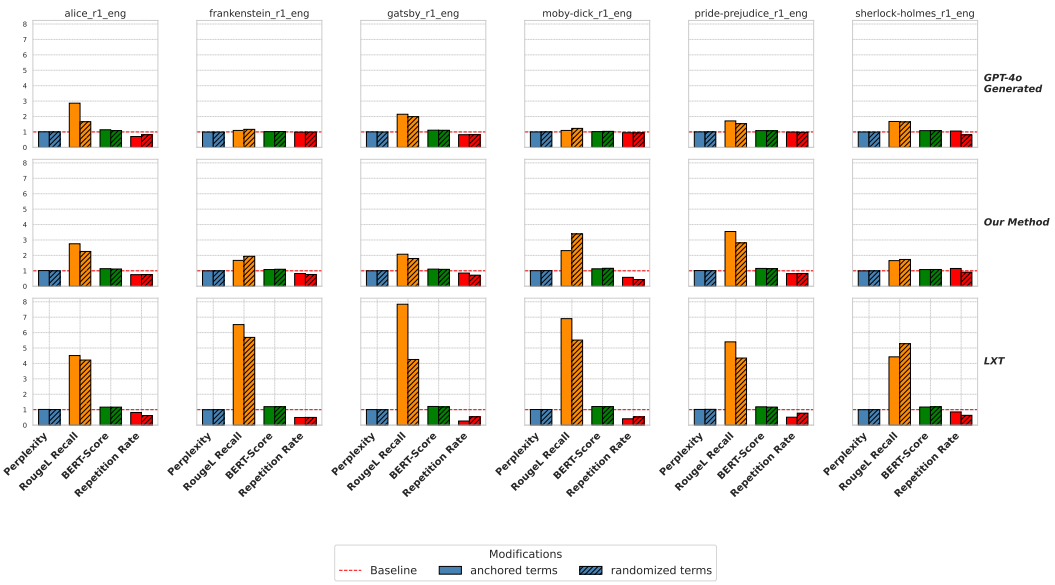

Figure 12: Evaluation on individual books for Llama 2-7b.

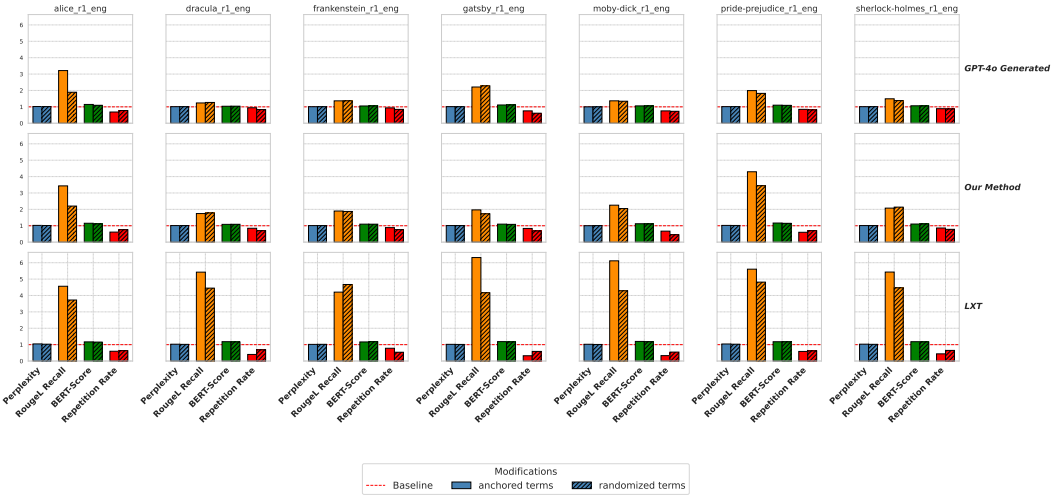

Figure 13: Evaluation on individual books for Llama 2-13b.

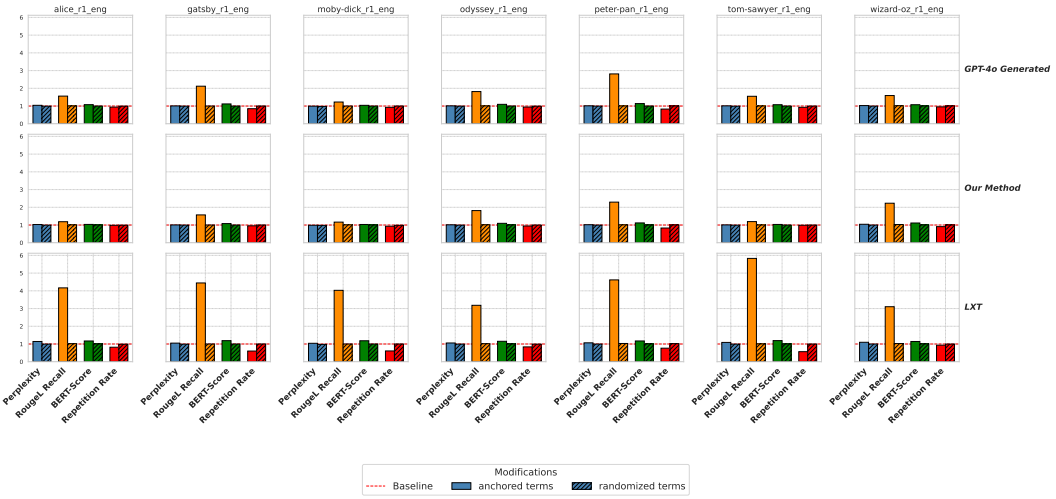

Figure 14: Evaluation on individual books for Olmo 2-13b.

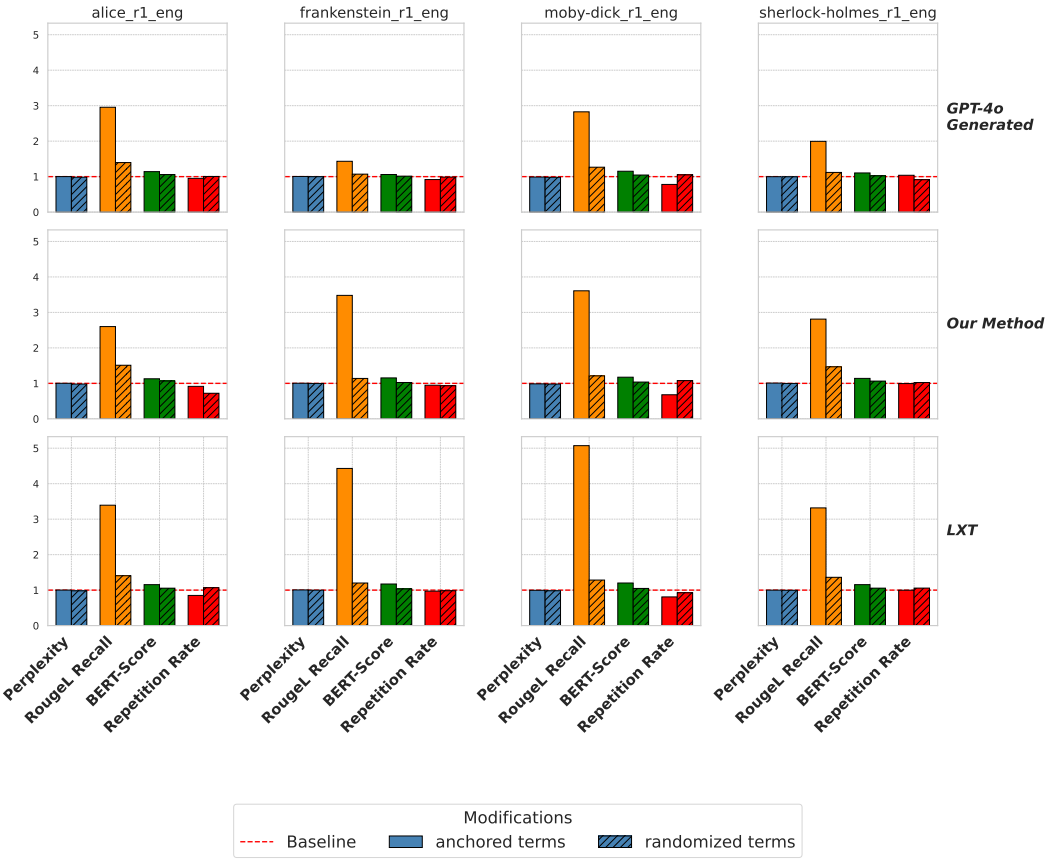

Figure 15: Evaluation on individual books for GPT Neox-20b.

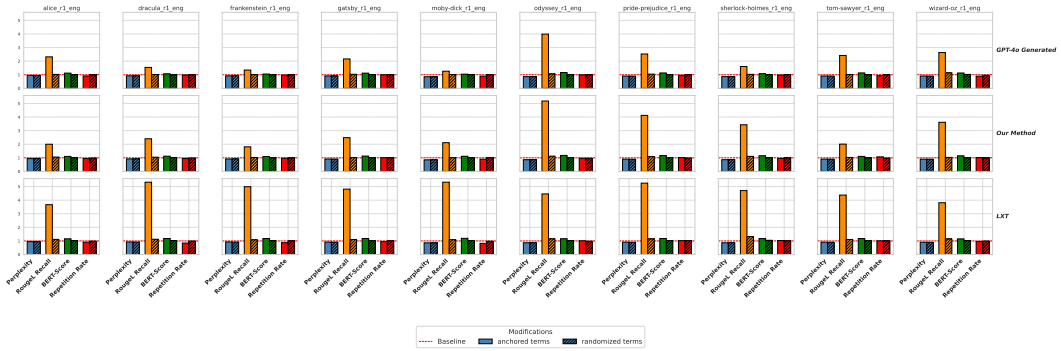

Figure 16: Evaluation on individual books for Llama 3.1.

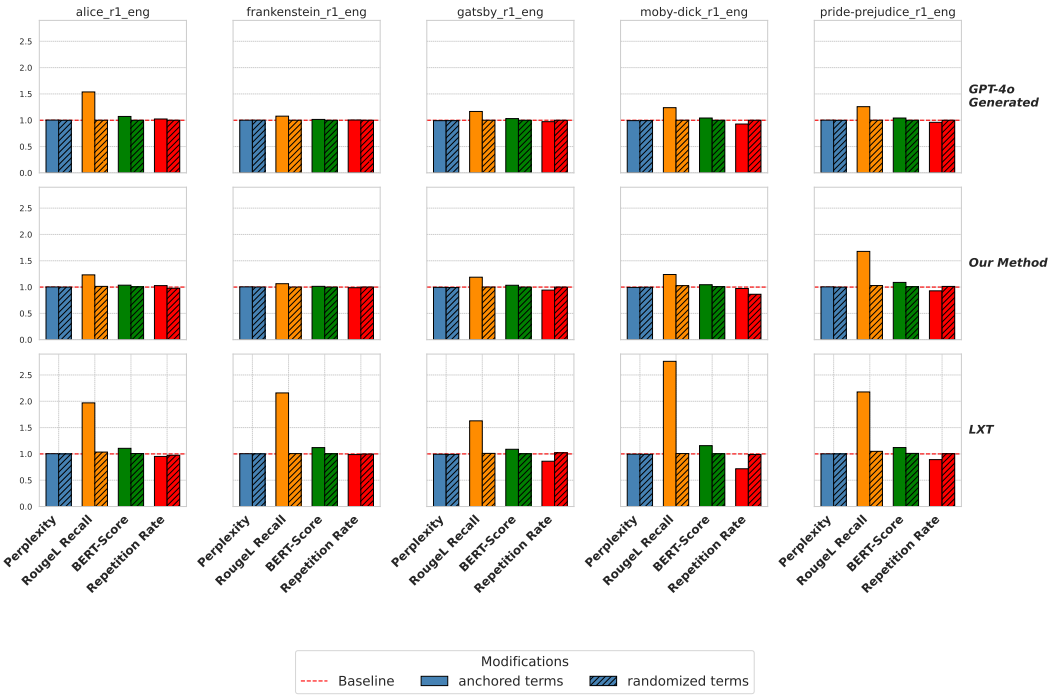

Figure 17: Evaluation on individual books for Phi 4.

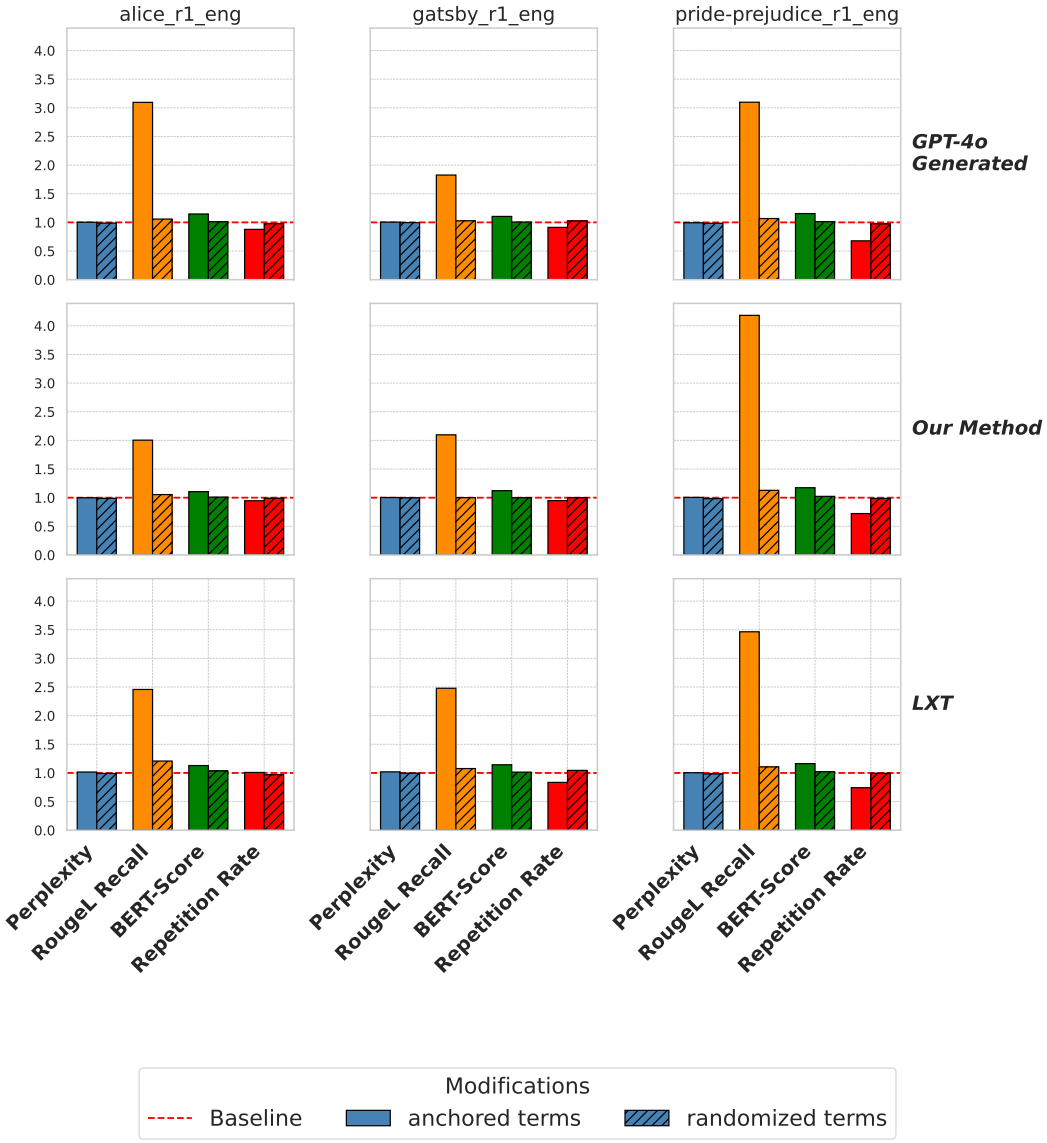

Figure 18: Evaluation on individual books for Qwen 2.5-14b.