The Curious Case of Factual (Mis)Alignment between LLMs' Short- and Long-Form Answers

Anonymous ACL submission

Abstract

Large language models (LLMs) can correctly answer *"When was Einstein born?"* yet fail to provide the same date when the same question is embedded in a long-form query requesting multiple facts about Einstein's life, revealing a fundamental inconsistency in how models access facts across different complexities of knowledge seeking tasks. While models display impressive accuracy on factual question-answering benchmarks, the reliability gap between simple and complex long-form queries remains poorly understood, eroding their trustworthiness. In this work, we introduce Short-Long Form Alignment for Factual Question Answering (SLAQ), a controlled evaluation framework that compares LLMs' answers to the same factual questions asked (a) in isolation (*short*) vs. (b) integrated into complex queries (*long*). Looking at 16 LLMs across 600 queries, we find a systematic misalignment of answers to the corresponding *short* and *long* queries. We further uncover momentum effects where consecutive correct or incorrect answers create self-reinforcing patterns. Through mechanistic analysis, we find that aligned facts activate overlapping model internals, and that metrics based on mechanistic similarity can predict *short-long* answer alignment with up to 80% accuracy. Our work establishes *factual consistency over query complexity* as an important aspect of LLMs' trustworthiness and challenges current evaluation practices, which implicitly assume that good performance for simple factual queries implies reliability in more complex knowledge-seeking tasks as well.

1 Introduction

Large Language Models (LLMs) (Team et al., 2023; Achiam et al., 2023; Grattafiori et al., 2024; Yang et al., 2025a) are rapidly being adopted across diverse applications, including education (Kasneci et al., 2023), healthcare (Qiu et al., 2023), software engineering (Fan et al., 2023), and general knowledge search (Xu et al., 2023). Their utility and trustworthiness are, however, compromised by their tendency to hallucinate and generate fictitious responses (Wang et al., 2023a; Huang et al., 2025).

While earlier research on LLM evaluation extensively examined factual accuracy in closed-domain question answering (QA) for both short-form (Rajpurkar et al., 2016; Joshi et al., 2017; Yang et al., 2018) and long-form responses (Fan et al., 2019; Dasigi et al., 2021), these evaluation benchmarks have become somewhat outdated, since LLMs are now primarily deployed as chat-based information gathering assistants across a wide variety real-world applications (Xu et al., 2023), i.e., they are primarily used as (chat-based) open-domain question-answering tools. Accordingly, factuality-oriented evaluations have shifted toward open-domain QA, considering both short-form (Lin et al., 2022; Wei et al., 2024b) and long-form responses (Min et al., 2023; Wei et al., 2024b; ul Islam et al., 2025). Existing benchmarks, however, evaluate short-form and long-form factuality in isolation, and thus fail to assess *factual consistency* of models' responses over query complexity: *Will an LLM yield the same answer to the same factual question for queries of varying complexity?*

In this work, we address this gap by introducing Short-Long Form Alignment for Factual Question Answering (SLAQ), a novel evaluation framework that tests whether models maintain answer consistency—with respect to fact-seeking questions—across queries of different complexity. SLAQ presents an LLM with the same fact-seeking questions, formulated (independently) in two distinct query formats: (1) *long* queries combine five topically related factual questions, whereas (2) *short* queries formulate those same questions independently and ask them in isolation. With this controlled design, we isolate the impact of query/response complexity on factual answer accuracy. By comparing the factual correctness of

models' answers to long vs. short queries, we can disentangle knowledge gaps (incorrect answers for both query formats) from answer retrieval failures (e.g., correct answer for the short query but incorrect for the long). Figure 1 illustrates SLAQ.

Studying 16 LLMs through the lens of SLAQ, we find that, while models exhibit substantial *short-long alignment* w.r.t. factual answer correctness, most of this alignment stems from incorrect answers, i.e., LLMs produce *incorrect* answers for the same factual question in both long and short-form queries (but it is not necessarily the same answer). We observe that models consistently demonstrate higher factual accuracy in responses to short queries than in long-query responses: the majority of misalignment cases thus stem from a (1) correct answer to the short query and an (2) incorrect answer to the corresponding question included in the *long* query. Beyond evaluating factual question answering accuracy for both query formats, we identify momentum effects, where consecutive correct answers increase the likelihood of subsequent accuracy, while errors tend to cascade and compound.

To understand the mechanistic basis of the observed factual misalignment, we next analyze the model internals— attention and MLP activation patterns—and identify minimal sets of model components responsible for answer generation for short and long-form queries, respectively. Using zero-ablation (Olsson et al., 2022), we find that aligned answers activate significantly more similar computational pathways and exhibit stronger correlations in component importance rankings. Moreover, we show that these circuit-level differences have predictive power: employing six pathway similarity metrics, we can predict with 80% accuracy (ROC-AUC: 0.85) whether the answers to the same factual question will align between the two query formats, short and long; here we identify attention head rank correlation as the most predictive feature.

Contributions. In sum, the contributions of this work are threefold: (1) We establish *factual consistency over query complexity* as an important aspect of LLM reliability and introduce SLAQ, a novel dataset for benchmarking such consistency; (2) We document systematic factual misalignment (i.e., inconsistency) patterns in LLMs, and relate factual correctness in the long-form responses to momentum dynamics; (3) We provide mechanistic evidence that this factual misalignment stems from divergent internal processing, demonstrating that circuit overlap metrics can predict alignment outcomes. This work represents the first systematic investigation of factual consistency over query complexity in open-domain QA. Our findings challenge a fundamental (implicit) assumption of modern LLM evaluation: that factual knowledge that LLMs exhibit for simple queries with straightforward factual questions propagates reliably to complex scenarios.

2 Background and Related Work

We provide a brief overview of background and related work on (1) hallucinations and factuality in open-domain QA, and (2) mechanistic interpretability and its application to understanding factuality.

Hallucinations and Factuality. Evaluating factual accuracy in LLMs has evolved from simple to complex formats. Early benchmarks like TriviaQA (Joshi et al., 2017) and Natural Questions (Kwiatkowski et al., 2019) evaluated LLMs on closed-domain QA. But these benchmarks are now saturated, and evaluation of LLMs has moved from closed-domain to open-domain QA, with TruthfulQA (Lin et al., 2022) and SimpleQA (Wei et al., 2024a) as two popular benchmarks that evaluate LLMs for single factoid answers.

With respect to factual accuracy in long-form LLM responses, FactScore (Min et al., 2023) evaluates LLMs on Wikipedia biographies. More recently, LongFact (Wei et al., 2024b) and UNCLE (Yang et al., 2025b) were proposed to evaluate long-form factual accuracy across diverse domains. UNCLE is a concurrent effort to ours and, similarly to our work, pairs short and long queries/prompts: however, it analyses the short- and long-form in isolation and studies uncertainty expression rather than factual consistency of LLMs' responses between the two query formats.

Several systematic phenomena have been observed regarding hallucinations in LLMs. The "snowballing" effect (Zhang et al., 2024) describes how models justify a wrong claim by generating additional false assertions. The "lost in the middl" phenomena (Liu et al., 2024) shows input-position sensitivity in closed-domain QA: accuracy peaks when evidence appears at the beginning or end of a long context and degrades when relevant information lies in the middle. Complementing input-position effects, Yang et al. (2025b) find that hallucinations in long-document summarization oc-

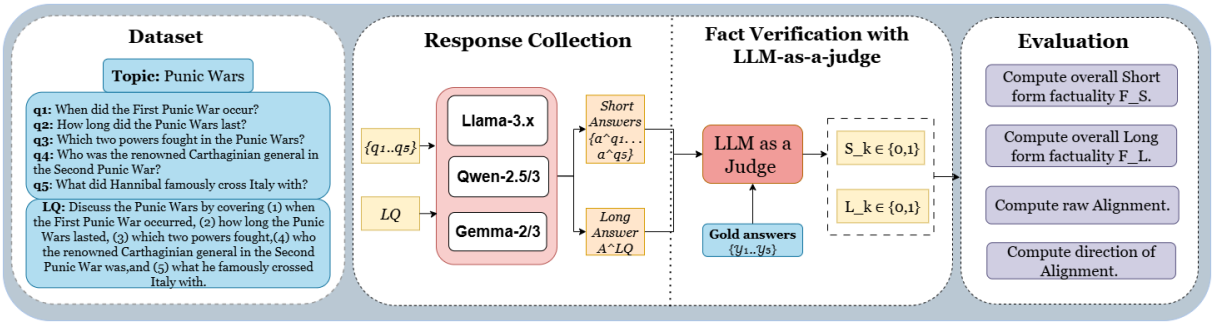

Figure 1: Illustration of our **Short-Long Form Alignment for Factual Question Answering** (SLAQ) framework. An instance in our SLAQ benchmark is a *complex* knowledge-seeking query, i.e., a **long** query, which consists of five *simple* factual sub-queries, i.e., **short** queries, each with an unambiguous correct answer. LLMs independently generate the answers to (1) the **long** query (i.e., all five short queries combined) and (2) each of the five short queries in isolation. We use a state-of-the-art commercial LLM to judge the correctness of the generated answers to both the *long* query and *short* queries against the set of reference answers; we use these judgments to compute models' short- and long-form accuracy (F_S, F_L) as well as the *short-long alignment* scores.

cur more often near the end of generated outputs ("hallucinate at the end"), indicating degradation dependent on the output position.

Mechanistic Interpretability (MI) aims to reverse-engineer how neural networks result in specific behaviors by identifying causal computational structures (Elhage et al., 2021; Zhang et al., 2024). The foundation of MI is localization: determining which model components (attention heads, MLP layers, neurons) are responsible for particular outputs. The primary technique for measuring component importance is activation patching (Meng et al., 2022), which quantifies causal influence through intervention. The process involves: (1) computing baseline output logits ℓ_{base} for the correct token, (2) ablating each component individually to obtain ℓ_{ablated}, (3) measuring the importance as normalized logit difference: $|\ell_{\text{base}} - \ell_{\text{ablated}}|/|\ell_{\text{base}}|$, where larger values indicate greater causal importance. Components are then assembled into minimal circuits through greedy search (Conmy et al., 2023; Hanna et al., 2024)—iteratively adding components in importance order until the subset reproduces the original behavior within a faithfulness threshold (Wang et al., 2023b). Two ablation strategies exist: zero-ablation (Olsson et al., 2022) sets component outputs to zero, while counterfactual patching replaces them with activations from different inputs (Meng et al., 2022). In this work, for computational efficiency, we resort to zero-ablation when identifying component sets.

A significant amount of work focused on identifying parameters in which models store factual information. Meng et al. (2022) localizes factual associations to mid-layer MLPs, for which Geva et al. (2023) further show that they function as key-value memories. Yao et al. (2024) trace factual retrieval circuits, revealing collaborative knowledge encoding across attention heads and MLPs. Limited work exists on comparing circuits between tasks: Mondorf et al. (2024) find high node overlap for compositionally similar tasks, while Hanna et al. (2025) report minimal overlap between formal and functional linguistic circuits. These studies share a critical limitation: they analyze single-token outputs, ignoring the complexity of realistic free-form multi-token answers.

3 Factual Consistency over Query Complexity

Our goal is to capture the extent to which LLMs provide consistent (i.e., factually equivalent) answers to the very same fact-seeking questions, integrated into queries of different complexity. To this end, we introduce a novel task of factual answer consistency over query complexity, for which we create an evaluation benchmark.

3.1 Task Definition and Metrics

SLAQ tests whether LLMs provide consistent answers to factual questions across different query/response complexities. Because of this, we organize the benchmark around *topics*: a topic t is a set of N topically related facts $\{f_1, f_2, \ldots, f_N\}$, for each of which SLAQ contains a *short*-form question (SQ) that elicits the respective fact, $\{q_1, q_2, \ldots, q_N\}$. Each topic, as a set of N facts, is additionally converted into a *long* information-seeking (LQ) query: an example of a

topic t with $N = 5$ factual questions is given in Figure 1. The LLMs then independently respond to the LQ, as well as to each of the N SQs. For each factual question q_k ($k \in \{1, 2, \ldots, N\}$), $S_k \in \{0, 1\}$ denotes the factual correctness of an LLM's answer to the SQ of that fact (1 = correct, 0 = incorrect) whereas the $L_k \in \{0, 1\}$ indicates the correctness for the same fact in the LLM's answer to the LQ.

Alignment Definition. We declare that an LLM produces a *factually consistent* response for a fact f_k if the SQ and LQ responses for that fact have the same factual correctness label:

$$\text{aligned}(k) = \mathbb{I}\{S_k = L_k\} \tag{1}$$

where $\mathbb{I}\{\cdot\}$ is the indicator function. Crucially, alignment measures the consistency in factual correctness of the answers, and not whether the answers themselves are semantically equivalent. When both responses are incorrect ($S_k = L_k = 0$), they are aligned w.r.t. factual correctness because they are both incorrect. E.g., for question q_1 from Figure 1, answers "264 to 243 BCE" (short) and "264 to 241 AD" (long) are factuality-wise aligned (both incorrect) because the correct answer is "264 to 241 BCE" We choose this alignment definition as we aim to discern between (1) a knowledge gap (i.e., both answers incorrect; irrelevant if they are "the same incorrect") and (2) failure to consistently retrieve the knowledge that is stored in the model (i.e., one answer correct, the other incorrect).

Evaluation Metrics. Following prior work (Min et al., 2023; Wei et al., 2024a,b), we employ an LLM to judge factual correctness by comparing responses against gold answers. We characterize model behavior with following metrics: Short and Long form factual **Accuracy** (F_S and F_L), $Align$ment score, and **Signed Alignment** score ($Align_{\pm}$).

$$F_S = \frac{1}{N}\sum_{k=1}^{N} S_k \qquad F_L = \frac{1}{N}\sum_{k=1}^{N} L_k \tag{2}$$

$$\text{Align} = \frac{1}{N}\sum_{k=1}^{N} \mathbb{I}\{S_k = L_k\} \tag{3}$$

$$\text{Align}_{\pm} = \tfrac{1}{N}\sum_{k=1}^{N} A_k, \quad A_k = \begin{cases} 1 \text{ if } S_k = L_k = 1, \\ -1 \text{ if } S_k = L_k = 0, \\ 0 \text{ if } S_k \neq L_k. \end{cases} \tag{4}$$

F_S and F_L establish baseline performance and quantify the accuracy gap between formats. $Align$ measures factual consistency regardless of correctness. Raw alignment score conflate reliable knowledge (both correct) and systematic failure (both incorrect), which is why we introduce $\text{Align}_{\pm}$, which additionally distinguishes the two correctness cases: +1 (aligned, correct) vs. -1 (aligned, incorrect), with 0 denoting factual misalignment.

3.2 Dataset

We construct SLAQ datasets from Wikipedia, leveraging its factual reliability and broad coverage. We sample from 15 diverse English Wikipedia categories, selecting articles exceeding 1,000 words to ensure sufficient factual density. To test models across the knowledge popularity spectrum, we balance between popular and obscure topics, selecting 300 most-viewed and 300 least-viewed pages in the past five years. This way, we take into account evidence (Zhang et al., 2025) that fact frequency drives LLMs' hallucination.

Following the success of LLM as synthetic data generators (Long et al., 2024) in open-domain QA (Wei et al., 2024b; ul Islam et al., 2025; Yang et al., 2025c) we employ a state-of-the-art commercial LLM, OpenAI `o3-mini-high`, to generate factual questions to which an answer exists in the article content. The model receives the full Wikipedia text and produces $N = 5$ SQs targeting distinct facts, plus one LQ that naturally elicits all five facts[1].

We then manually verified all generated SQs and LQs, ensuring their open-domain formulation (i.e., that they are answerable without the source article) as well as factual grounding (the correct answer indeed exists in Wikipedia). We manually adjusted the queries that did not meet both criteria.[2]. Overall, we found `o3-mini-high` to be a reliable synthetic data generator for this purpose: it introduced errors in only 77 out of 3,600 (2.14%) query-answer pairs.

The SLAQ datasets covers 600 unique topics (with 600 corresponding LQs) with 3000 SQs ($N = 5$), across 15 Wikipedia categories (on average, 40 topics per category). We further profile the SQs for fact-type, finding 1,071 entity-based facts and 1,929 non-entity facts (definitions, properties, equations, concepts)[3]. The final SLAQ evaluation dataset is thus both category-diverse and popularity-balanced, and has a fair balance between factual questions with entity vs. non-entity target answers.

[1] We provide the prompts for generating SQs, LQs, and dataset samples in the §A.1

[2] E.g., we identified 53 SQs and 24 LQs that violated open-domain criteria and were judged as unanswerable without the respective Wikipedia article

[3] For this labeling, we resort to the OntoNotes taxonomy.

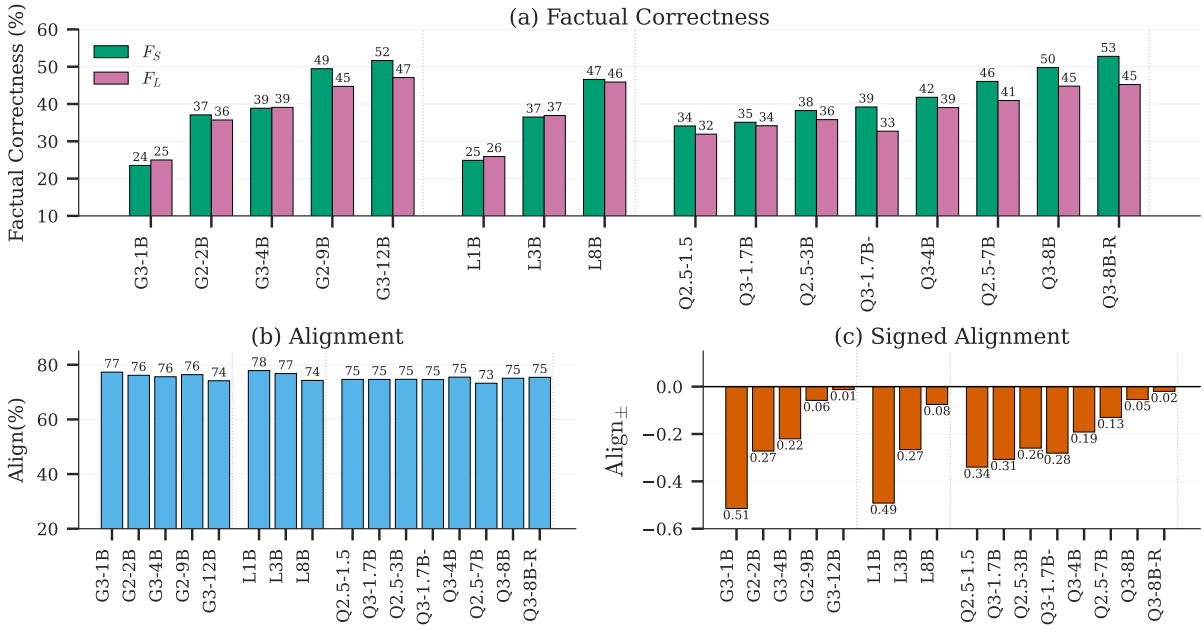

Figure 2: Short–long factual alignment results. (a) **Factual Correctness**: per-model short-form accuracy F_S (green) and long-form accuracy F_L (purple). (b) **Alignment**: Align = percentage of facts with the same correctness label in short vs. long responses. (c) **Signed Alignment**: average over topics; for a single topic, the score is the average of $\text{Align}_{\pm}$ of its five facts. Models key: G = Gemma, L = Llama, Q = Qwen (e.g., Q3–8B-R = Qwen-3-8B-Reasoning).

4 Benchmarking and Evaluations

4.1 Experimental Setup

We evaluate models from five families, spanning 1B to 12B parameters: Qwen-2.5 (Yang et al., 2024), Qwen-3 and Qwen-3-Reasoning (Yang et al., 2025a), Llama-3 (Grattafiori et al., 2024), Gemma-2 (Team et al., 2024), and Gemma-3 (Team et al., 2025). All models use greedy decoding via the Hugging Face API. We constrain short-form responses to single sentences and instruct models to provide only requested information for long-form queries (see Table 6 in the Appendix for prompts). To control for positional effects, we randomize the order of the five sub-questions within each long-form query across 5 permutations and report averaged results.

We employ Gemini-2.5-Flash (Team et al., 2023) as our LLM judge[4], instructing it to judge answer correctness based on factual equivalence with the gold answer rather than string matching[5]. We then compute our evaluation metrics (Eq. 2–4) from the judge's binary correctness labels.

4.2 Results and Analysis

Figure 2 reveals three key patterns of factual (in)consistency across the language models. Panel (a) shows that most models achieve modest factual accuracy of 30-50% on both SQ (F_S) and LQ (F_L) responses, with almost all models displaying higher accuracy for short-form queries. Larger models display only modestly better performance, suggesting that scale alone cannot dramatically improve factual recall.

Panel (b) shows remarkable consistency in raw alignment across all models, with virtually every model achieving 73-78% alignment regardless of size and architecture. Such uniformity indicates that this level of factual consistency over query complexity is an intrinsic property of modern LLMs, rather than something that can improve with scale.

Panel (c) reveals the most critical finding: all models show negative signed alignment (-0.01 to -0.51), meaning they consistently provide wrong answers in responses to both query formats more often than correct answers for both cases. This indicates that the high raw alignment primarily reflects systematic failures rather than systematic successes: models have developed stable internal strategies for factual processing, but these strategies systematically fail to retrieve correct information. The combination of high raw alignment with negative signed alignment reveals that while models are internally consistent in factual behavior, most of it stems from generating incorrect answers

[4]See Appendix Tables 4 and 5 for prompts
[5]See §A.2 for judge performance

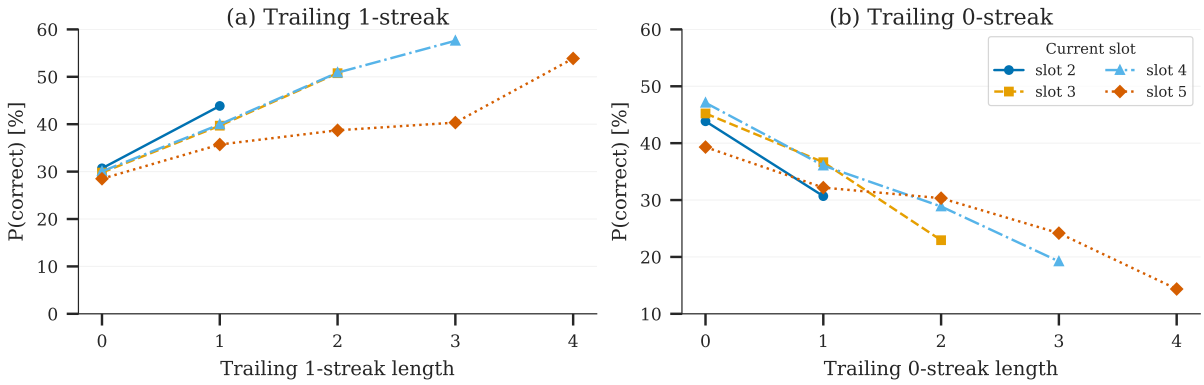

Figure 3: Long-form QA dynamics by sub-fact position. (a) **Trailing 1-streak**: $P(\text{correct})$ for the current slot (2–5), conditioned on the length of the immediately preceding run of correct slots. (b) **Trailing 0-streak**: $P(\text{correct})$ for the current slot (2–5), conditioned on the length of the immediately preceding run of incorrect slots. Slots (1-5) are sub-fact positions in the long-form query.

across query complexities.

To test whether short-long gap is merely an artifact of natural rephrasing in LQ, we construct a control prompt by *literally* concatenating the five short questions, fully preserving their original wording. As shown in §A Fig. 6, the findings remain the same as in Fig. 2, suggesting that the short-long gap cannot be credited to minor question rephrasings in LQ and is truly due to requesting multiple facts at once being more challenging for LLMs.

Panel (a) shows that F_S is consistently larger than F_L.[6] To better understand why F_L is lower, we next analyze LQ responses in more detail.

Momentum within a response. According to Figure 3b and 3c, following consecutive correct answers (positive momentum), accuracy increases from a 30% baseline to 57% after four correct facts: each additional success adds roughly 7% to subsequent accuracy. Conversely, consecutive errors (negative momentum) reduce accuracy from 45% to 24% after three mistakes. This not only confirms Zhang et al. (2024)'s finding of "snowballing" for long-form QA but extends it by quantifying the error propagation and success reinforcement effects. These momentum dynamics explain why LQ responses systematically underperform short-form ones even when eliciting the very same factual knowledge.

5 Mechanistic Analysis

Our behavioral analysis revealed that language models exhibit systematic inconsistencies when answering the same facts across short and long response formats. This raises a fundamental question: do these behavioral differences reflect distinct internal computational mechanisms? Understanding the mechanistic basis of factual alignment could inform interventions to improve consistency across response formats. We hypothesize that factual alignment corresponds to mechanistic similarity. Formally, let $\text{sim}(k)$ denote the mechanistic similarity between short and long responses for fact k. Our hypothesis predicts:

$$\mathbb{E}[\text{sim}(k) \mid aligned] > \mathbb{E}[\text{sim}(k) \mid misaligned] \quad (5)$$

with alignment defined as in §3. We focus exclusively on facts that are answered correctly in both formats (aligned) versus facts where only one format is correct (misaligned). Put simply: facts answered correctly in both short and long formats should exhibit greater internal mechanism overlap than facts answered correctly in only one format.

5.1 Preliminaries: Component Importance via Zero-Ablation

We identify critical components by systematically setting their outputs to zero and measuring how much that hurts the model's preference for the gold token. For each component c and answer token t, we measure importance using:

$$\text{importance}(c, t) = \frac{\text{logit}_t^{\text{base}} - \text{logit}_t^{\text{ablated}}}{\text{logit}_t^{\text{base}}} \quad (6)$$

This quantifies the change in logit magnitude when a component c is removed, normalized by the baseline logit. For *each* answer token, components are ranked by importance. We greedily select components (highest importance first) until we recover at

[6] The only exceptions are the two smallest models in our evaluation, Gemma-3 1B and Llama-3 1B, which yield very low accuracy for both SQs and LQs.

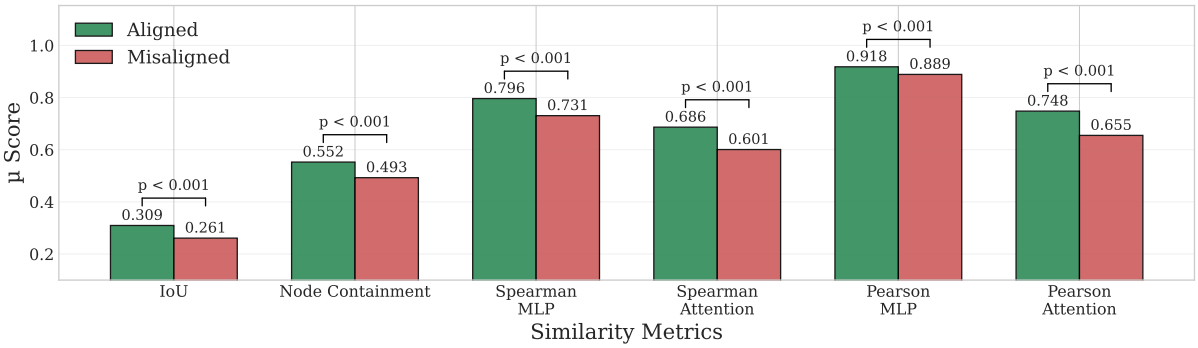

Figure 4: Mechanistic similarity comparison between aligned and misaligned facts across six metrics averaged across six LLMs. Aligned facts (green) show significantly ($p < 0.05$) higher mechanistic similarity than misaligned facts (red) for all measures. Scores are averaged across six LLMs.

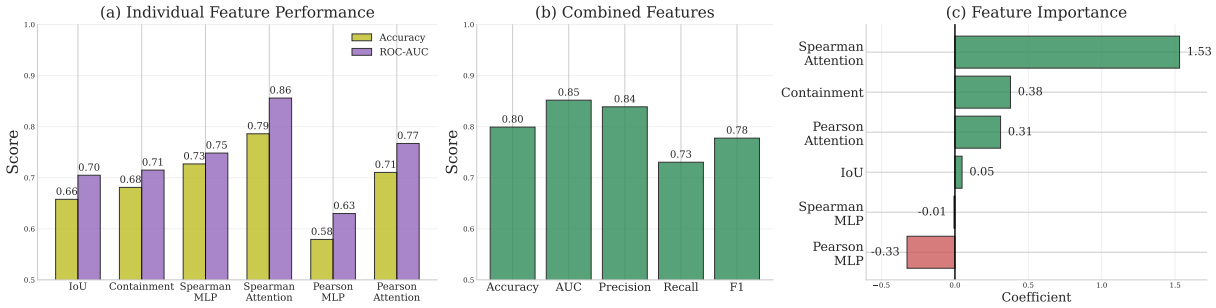

Figure 5: Predictive modeling performance using circuit similarity metrics. (a) Individual feature performance shows Spearman Attention as the strongest single predictor of factual alignment (ROC-AUC = 0.86). (b) Combined features achieve robust performance across all evaluation metrics (ROC-AUC = 0.85, Accuracy = 0.80). (c) Feature importance reveals Spearman Attention as the dominant predictor (coefficient = 1.53).

least 90% of the baseline logit, yielding the minimal set C_t needed for generating token t.

Similarity Metrics. We compare component sets of responses to SQs and LQs with two metrics:

$$\text{Containment} = \frac{|C_{\text{short}} \cap C_{\text{long}}|}{\min(|C_{\text{short}}|, |C_{\text{long}}|)} \tag{7}$$

$$\text{Intersection-over-Union} = \frac{|C_{\text{short}} \cap C_{\text{long}}|}{|C_{\text{short}} \cup C_{\text{long}}|} \tag{8}$$

with C_{short} and C_{long} being the component sets for short and long responses, respectively. Containment measures core component sharing relative to the smaller circuit, while IoU quantifies the overlap symmetrically. We additionally measure **Pearson** and **Spearman Correlation** between the two sets of component importance scores: the former captures the extent to which importance scores match/deviate across components, whereas the latter quantifies the extent to which the two rankings of components (by decreasing importance score) match.

Multi-Token Alignment via Earth Mover's Distance (EMD) Previous studies on LLMs' factual recall (Meng et al., 2022; Geva et al., 2023; Yao et al., 2024) focused on single-word answers. However, real answers to factual questions span multiple tokens, with variable lengths. Because we extract component sets per token, we need to aggregate token-level component-based similarity scores into fact-level (i.e., answer-level) scores, taking into account different token boundaries (e.g., *"Paris is capital of France"* vs *"the capital city Paris is located in France"*). We formulate component comparison across multi-token answers as an optimal transport problem: we compute pairwise similarities between all SQ answer tokens $\{s_i\}$ and corresponding LQ answer tokens $\{l_j\}$, creating a bipartite similarity matrix M_{ij}. EMD then finds the transport plan matrix π^* (i.e., coefficients π_{ij}) that maximizes the following total similarity:

$$\pi^* = \arg\max_{\pi} \sum_{i,j} \pi_{ij} \cdot M_{ij} \tag{9}$$

The final fact-level similarity score is then the weighted average of pairwise similarities with optimal transport weights: $\text{sim}(k) = \sum_{i,j} \pi^*_{ij} \cdot M_{ij}$. Unlike exact matching, EMD finds the best possible alignment of tokens between multi-token an-

swers, reflecting semantic equivalence between tokens regardless of their order in the answer.

5.2 Mechanistic Comparison

Experimental Setup. We analyze six models spanning different scales: Llama-3.2 (1B, 3B), Qwen-2.5 (1B, 3B), and Qwen-3 (1.7B, 4B). For each model, we sample 60 short-long fact pairs that are divided into 30 correctly aligned pairs and 30 misaligned pairs. This yields us 360 short-long fact pairs. For each fact, we obtain: (1) Minimal component sets for all answer tokens (2) Importance scores for all attention heads and MLP layers. We set the the greedy threshold to 90% are are able to recover the original token with 100% accuracy. After extracting minimal components and importance scores, we compute pairwise similarity measures between all tokens in the short answer and the corresponding-fact tokens in the long answer span, then apply EMD to aggregate these into a single answer-level similarity score.

Results. Figure 4 reveals that aligned facts exhibit significantly higher mechanistic similarity than misaligned facts across all six metrics (all $p < 0.001$). Set-based metrics show substantial differences: IoU increases by 18.4% for aligned facts (0.309 vs 0.261), while Containment shows a 12.0% increase (0.552 vs 0.493). These gaps indicate that factually consistent responses recruit more overlapping components, while inconsistent responses activate divergent computational pathways. Correlation-based metrics reveal complementary patterns. Pearson correlations achieve higher absolute values than Spearman correlations for both attention (0.748 vs 0.686) and MLP components (0.918 vs 0.796), indicating strong linear relationships in raw activation magnitudes. Examining discriminative power, attention metrics show the largest relative gaps regardless of correlation type (Spearman: 14.1%, Pearson: 14.2%), suggesting attention patterns are most diagnostic of factual alignment. In contrast, MLP components show divergent patterns: Spearman correlations provide stronger discrimination (8.9%) than Pearson correlations (3.3%), indicating that the ranking of MLPs by importance is more informative than their raw activation (i.e., importance) magnitudes.

Overall, these results confirm our hypothesis: factual alignment corresponds to mechanistic similarity.

5.3 Model Internals as Predictors

Building on our finding that aligned facts exhibit significantly higher mechanistic similarity, we investigate whether these similarity metrics can predict factual alignment. We train a logistic regression classifier using the six similarity metrics as features on a dataset of 360 fact pairs (180 aligned, 180 misaligned) across six models (Llama-3.2 1B/3B, Qwen-2.5 1B/3B, Qwen-3 1.7B/4B). We evaluate performance via 5-fold cross-validation to assess both individual feature contributions and combined predictive power.

Results and Analysis. The predictive modeling results in Figure 5 validate the mechanistic basis of factual alignment and reveal which similarity metrics best capture this phenomenon. Spearman attention emerges as the strongest individual predictor of alignment (ROC-AUC = 0.86, Accuracy = 0.79), indicating that *ranks* of attention components by importance provide the most reliable measurable signal of factual consistency. The combined feature model achieves robust performance (ROC-AUC = 0.85, Accuracy = 0.80), with logistic regression coefficients revealing the underlying computational structure: Spearman Attention dominates with a coefficient of 1.53, more than four times larger than the next most discriminative feature (Containment).

These results demonstrate that mechanistic interpretability offers not just explanatory insights into model behavior but also a practical tool for predicting factual inconsistencies in LLMs' responses.

6 Conclusion

This work establishes factual consistency over query complexity as a critical dimension of LLM reliability. Through SLAQ, we demonstrate that LLMs exhibit systematic misalignment when answering identical factual questions embedded in queries of varying complexity. Our analysis reveals that LLMs are mostly aligned on being factually incorrect and correctness exhibits momentum effects. Our mechanistic analysis provides the first empirical evidence that behavioral factual alignment corresponds to similar internal mechanisms. The predictive modeling results demonstrate practical applications in detecting factual misalignment. Future work should explore factual alignment between short and free-form long-form queries, and investigate targeted interventions on circuit patterns to address the consistency failures identified in LLMs.

Limitations

SLAQ Dataset and Framework The SLAQ dataset and framework have several limitations. (1) The dataset is synthetically generated. While a fully human-annotated dataset would be ideal, it is cost-prohibitive; using LLMs as data generators offers a more balanced quality–resource trade-off. We mitigate this limitation by manually verifying each prompt–answer pair and find that OpenAI's `o3-mini-high` is a strong data generator for the SLAQ task (§3.2). (2) The dataset—and this work—currently covers only English. (3) For evaluation, we adopt the LLM-as-a-judge paradigm to assess factual correctness; in our experiments, however, we find `Gemini-2.5-Flash` achieves high agreement with human annotations (see §4). (4) Our dataset and evaluation framework does not address free-form long-form generation. Lastly, control in §4.2 does not fully isolate the compositional complexity from all prompt-format effects.

Mechanistic Analysis In our mechanistic analysis, we use zero ablation rather than counterfactual activation patching. Although counterfactual activation patching can yield more precise results (Zhang and Nanda; Heimersheim and Nanda, 2024), constructing counterfactual prompts for complex generation tasks (short- and long-form) is challenging and labor-intensive, especially when relevant facts are distributed across multiple tokens. More importantly, our goal is to demonstrate mechanistic differences between short- and long-form factual retrieval under fact misalignment, which we show in §5 using zero ablation.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36:16318–16352.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4599–4610.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12.

Angela Fan, Beliz Gokkaya, Mark Harman, Mitya Lyubarskiy, Shubho Sengupta, Shin Yoo, and Jie M Zhang. 2023. Large language models for software engineering: Survey and open problems. In *2023 IEEE/ACM International Conference on Software Engineering: Future of Software Engineering (ICSE-FoSE)*, pages 31–53. IEEE.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. Eli5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12216–12235, Singapore. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Michael Hanna, Yonatan Belinkov, and Sandro Pezzelle. 2025. Are formal and functional linguistic mechanisms dissociated in language models? *arXiv preprint arXiv:2503.11302*.

Michael Hanna, Sandro Pezzelle, and Yonatan Belinkov. 2024. Have faith in faithfulness: Going beyond circuit overlap when finding model mechanisms. *arXiv preprint arXiv:2403.17806*.

Stefan Heimersheim and Neel Nanda. 2024. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models

for education. *Learning and individual differences*, 103:102274.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252.

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

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On LLMs-driven synthetic data generation, curation, and evaluation: A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11065–11082, Bangkok, Thailand. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100.

Philipp Mondorf, Sondre Wold, and Barbara Plank. 2024. Circuit compositions: Exploring modular structures in transformer-based language models. *arXiv preprint arXiv:2410.01434*.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. 2022. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*.

Jianing Qiu, Lin Li, Jiankai Sun, Jiachuan Peng, Peilun Shi, Ruiyang Zhang, Yinzhao Dong, Kyle Lam, Frank P-W Lo, Bo Xiao, et al. 2023. Large ai models in health informatics: Applications, challenges, and the future. *IEEE Journal of Biomedical and Health Informatics*, 27(12):6074–6087.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Saad Obaid ul Islam, Anne Lauscher, and Goran Glavas. 2025. How much do llms hallucinate across languages? on multilingual estimation of llm hallucination in the wild. *ArXiv*, abs/2502.12769.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, Yidong Wang, Linyi Yang, Jindong Wang, Xing Xie, Zheng Zhang, and Yue Zhang. 2023a. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *ArXiv*, abs/2310.07521.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023b. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024a. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, et al. 2024b. Long-form factuality in large language models. *Advances in Neural Information Processing Systems*, 37:80756–80827.

Ruiyun Xu, Yue Feng, and Hailiang Chen. 2023. Chatgpt vs. google: A comparative study of search performance and user experience. *arXiv preprint arXiv:2307.01135*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Joonho Yang, Seunghyun Yoon, Hwan Chang, Byeongjeong Kim, and Hwanhee Lee. 2025b. Hallucinate at the last in long response generation: A case study on long document summarization. *arXiv preprint arXiv:2505.15291*.

Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin 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, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yi-Chao Zhang, Yunyang Wan, Yuqi Liu, Zeyu Cui, Zhenru Zhang, Zihan Qiu, Shanghaoran Quan, and Zekun Wang. 2024. Qwen2.5 technical report. *ArXiv*, abs/2412.15115.

Ruihan Yang, Caiqi Zhang, Zhisong Zhang, Xinting Huang, Dong Yu, Nigel Collier, and Deqing Yang. 2025c. Uncle: Uncertainty expressions in long-form generation. *arXiv preprint arXiv:2505.16922*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Yunzhi Yao, Ningyu Zhang, Zekun Xi, Mengru Wang, Ziwen Xu, Shumin Deng, and Huajun Chen. 2024. Knowledge circuits in pretrained transformers. *Advances in Neural Information Processing Systems*, 37:118571–118602.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *The Twelfth International Conference on Learning Representations*.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. 2024. How language model hallucinations can snowball. In *Forty-first International Conference on Machine Learning*.

Yuji Zhang, Sha Li, Cheng Qian, Jiateng Liu, Pengfei Yu, Chi Han, Yi R Fung, Kathleen McKeown, Chengxiang Zhai, Manling Li, et al. 2025. The law of knowledge overshadowing: Towards understanding, predicting, and preventing llm hallucination. *arXiv preprint arXiv:2502.16143*.

A Appendix

A.1 Dataset Construction

We will release the SLAQ dataset under an open scientific license.

Scraping Wikipedia We collect articles from 15 curated categories. For each category, we use the MediaWiki API to list main-namespace pages and randomly sample candidates. For each candidate, we download the page, extract the main text, and remove tables, images, scripts/styles, hatnotes, navigation boxes, and the table of contents. We collapse whitespace and discard pages shorter than a minimum length of 1000 characters. For every accepted page, we query the Wikimedia Pageviews API to get daily counts over the past 1,095 days and sum them into one popularity score. Within each category, we sort by pageviews and split the list in half to form most and least popular sets.

Synthetic Data Generation We generate short/long queries from the scraped Wikipedia text using OpenAI's *o3-mini-high* model (API name `o3-mini-2025-01-31`, invoked with `reasoning_effort=high`). For each article, the prompt instructs the model to: (i) produce *exactly five* self-contained short questions (`ShortQ1–ShortQ5`) with single-fact, single-answer constraints and *no source referencing*; (ii) give concise answers (`ShortA1–ShortA5`, each under 10 words) derived *strictly* from the provided text (no parametric knowledge); (iii) compose a long question (`LongQ`) that *explicitly lists* those five sub-prompts; and (iv) write a fluent long answer (`LongA`) that synthesizes *only* the five short answers—no extra facts. The instruction emphasizes an "open-domain phrasing" test (the subject must be uniquely identifiable outside the Wikipedia article context), bans ambiguous "pick one of many" list questions, and enforces a strict, *single-line* output format. The total cost of constructing the dataset via *o3-mini-high* was **$18.57**. The prompt for generating the dataset is provided in Table 3 and the dataset schema is provided in Figure 7.

Topic Categories Following is a list of all the categories of the topics in the SLAQ dataset: Aesthetics, Algebra, Anthropology, Applied sciences, Artificial intelligence, Astronomy, Cultural studies, Sociology, Software engineering, Spirituality, Statistics, Technology, Telecommunications, Theology, Virtual reality.

A.2 LLM-as-a-judge

To evaluate LLM-as-a-jduge, we test on 50 samples (250 short questions and 50 long questions) and in total we have 500 atomic facts to be evaluate by LLM. We find LLM to have a an accuracy of 92.% on short-from and 94% in long-form responses.

Conditions for Incorrect or Correct: If any of the following conditions were met, the response was labeled as incorrect: (1) Factual inaccuracy, (2) Semantic dissimilarity, (3) IrrelevanceOmmission, (4) Contradiction, (5) Hallucinations. For correctness, following conditions have to be met: (1) Direct Semantic Equivalance, (2) Subset/Superset Equivalance, (3) World Knowledge Override (Only used sparsely), (4) Correct Vagueness. These condictions were developed progressively on a validation set of 50 samples and then tested on 50 samples separately. To understand the definition of the each condition, we recommend the reader to go through the prompts in Table 4 and 5.

A.3 Inference

We generate responses for each short and long query using the Hugging Face API. All models are executed on NVIDIA H100 (80GB) GPU with greedy decoding. The total GPU time required to generate all responses amounts to 192 hours.

We comply with the licensing agreement and adhere to the intended use of each of the open-source and closed-source LLMs.

A.4 Mechanistic Interpretability

We employ the `NNsight` framework to perform zero ablation on an NVIDIA H100 (80GB) GPU. The total compute time for zero ablation is 768 hours. Subsequently, we manually pair each short-form fact SQ_k with the minimal corresponding span of the fact in the long-form response. This manual pairing process required a total of 16 hours.

Optimal Transport for Mechanistic Overlap We compute mechanistic overlap using optimal transport algorithms, including the Hungarian algorithm and Earth Mover's Distance (EMD). Both methods yield consistent results—aligned responses exhibit higher mechanistic overlap.

A.5 Analysis by Fact Type

To test whether our main findings are driven by question-type, we separate SLAQ factoids into entity and non-entity based factoids and then further break down entity-based factoids by named-entity subtype using spaCy labels. We can observe in Table 1 and Table 2 that short-from factual correctness remain high, and that signed alignment is typically negative indicating that short-long factual misalignment is not confined by single factoid type.

Type	F_S	F_L	Align	Align$_\pm$
entity-based	29.87	28.61	80.03	-0.42
non-entity	45.00	41.89	73.33	-0.13
average	37.44	35.25	76.68	-0.27

Table 1: Results by coarse factoid type.

Entity label	n	F_S	F_L	Align	Align$_\pm$
Person	314	22.25	21.62	84.87	-0.56
Organization	338	33.36	29.51	80.14	-0.37
Geo-Political Entity	83	29.22	30.87	76.51	-0.40
Location	8	42.97	53.91	65.62	-0.03
Nationalities / Religious / Political groups	52	40.50	36.42	70.67	-0.23
Facility	6	28.12	17.71	81.25	-0.54
Work of Art	4	50.00	54.69	95.31	0.05
Event	7	42.86	35.71	69.64	-0.21
Product	8	42.97	35.16	73.44	-0.22
Law	4	25.00	21.88	84.38	-0.53
Language	4	87.50	85.94	82.81	0.73
Date	144	29.99	31.90	77.08	-0.38
Time	6	18.75	27.08	83.33	-0.54

Table 2: Results by fine-grained entity subtype. Small-n categories should be interpreted cautiously.

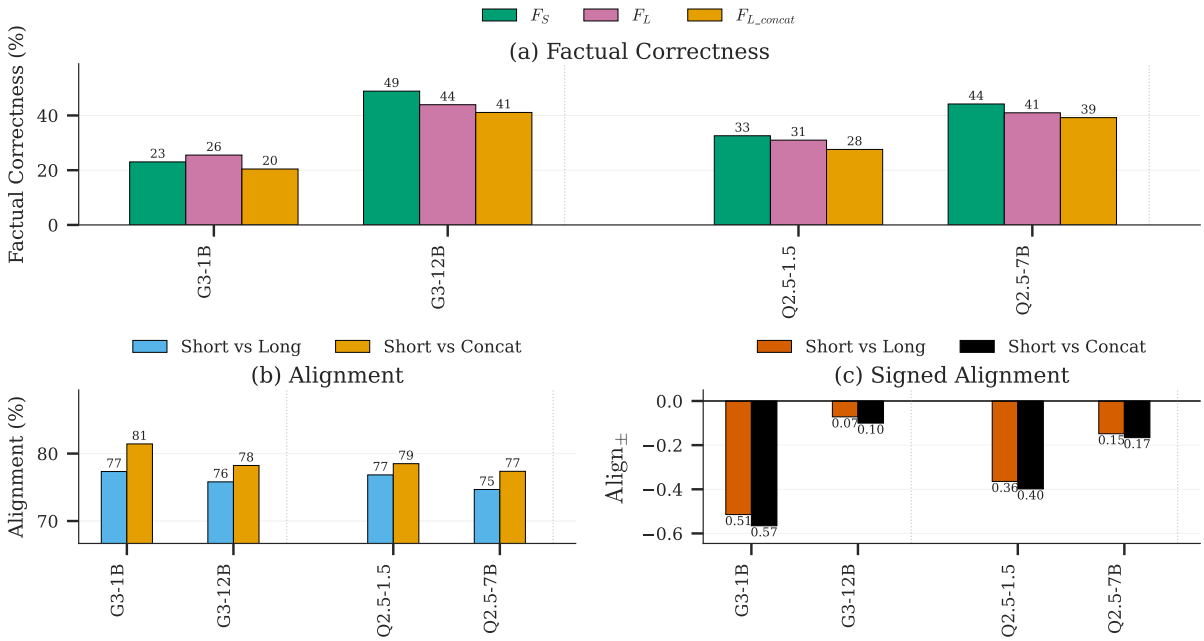

Figure 6: **Ablation:** By directly concatenating short queries into a long query. (a) **Factual Correctness**: per-model short-form accuracy F_S (green) and long-form accuracy F_L (purple). (b) **Alignment**: Align = percentage of facts with the same correctness label in short vs. long responses. (c) **Signed Alignment**: average over topics; for a single topic, the score is the average of Align$_\pm$ of its five facts. Models key: G = Gemma, L = Llama, Q = Qwen

```
{
  "Category": "History",
  "Topic": "The Punic Wars",
  "URL": "https://en.wikipedia.org/wiki/Punic_Wars",
  "ShortQ1": "When did the First Punic War occur?",
  "ShortA1": "It took place from 264 to 241 BCE.",
  "ShortQ2": "How long did the Punic Wars last?",
  "ShortA2": "They lasted a total of 43 years.",
  "ShortQ3": "Which two powers fought in the Punic Wars?",
  "ShortA3": "Rome and Carthage.",
  "ShortQ4": "Who was the renowned Carthaginian general in the Second Punic War?",
  "ShortA4": "Hannibal Barca was the famous general.",
  "ShortQ5": "What did Hannibal famously cross Italy with?",
  "ShortA5": "Hannibal crossed Italy with war elephants.",
  "LongQ"  : "Discuss the Punic Wars by covering (1) when the First Punic War occurred,
             (2) how long the Punic Wars lasted, (3) which two powers fought,
             (4) who the renowned Carthaginian general in the Second Punic War was,
             and (5) what he famously crossed Italy with.",
  "LongA"  : "The First Punic War took place from 264 to 241 BCE,
            and the three Punic Wars altogether lasted about 43 years between
            Rome and Carthage. In the Second Punic War, Hannibal Barca rose as
        the famed Carthaginian commander, becoming legendary for crossing Italy with war elephants.",
  "Pageviews": ...,
  "ShortQ1_Entity": 1,
  "ShortQ2_Entity": 0,
  "ShortQ3_Entity": 1,
  "ShortQ4_Entity": 1,
  "ShortQ5_Entity": 1
}
```


Figure 7: Example SLFA dataset entry instantiated for the Punic Wars.

Prompt

```
Given a *reference article* (plain text), generate **exactly**:
1) **Five** short Q/A pairs: ‘ShortQ1-ShortQ5‘, ‘ShortA1-ShortA5‘
2) **One** long question: ‘LongQ‘ (explicitly lists the five sub-prompts)
3) **One** long answer: ‘LongA‘ (coherent synthesis of the five short answers)
Examples (compressed)
- **Example 1 (placeholder)**
   *Reference article:* ‘[...Punic Wars summary...]‘
   *Output (abbrev):*
   ‘ShortQ1: When did the First Punic War occur¿
   ‘ShortA1: 264-241 BCE.‘
   ‘...‘
   ‘LongQ: Discuss by covering (1) . . . (2) . . . (3) . . . (4) . . . (5) . . . ‘
   ‘LongA: [Single paragraph formed only from ShortA1-5].‘
- **Example 2 (placeholder)**
   *Reference snippet:* ‘[.....]‘
   *Output (abbrev):* similar structure to Example 1.
Critical Instructions (strict)
  1.  **Absolute Grounding in Provided Text:** ALL answers (ShortA1-5, LongA) MUST be derived *exclusively* from the information
present in the reference text provided below. DO NOT use any external knowledge or information not explicitly stated in the text.
  2.  **Self-Contained & Precise Questions:** Each short question (ShortQ1-5) must be specific, unambiguous, and fully
understandable on its own without needing context from other questions or the article title.
     *Precision Example:* Use “When did the First Punic War occur?” instead of the vague “When did the war occur?”.
  3.  **No Source Referencing in Questions:** Questions MUST NEVER refer to the provided text itself. Avoid phrases like “According
to the article...”, “What does the text mention about...”, “Which item listed...”. Frame questions as standalone, open-domain
factual queries.
  4.  **Strict Single-Fact & Single-Answer Rule (MOST IMPORTANT):** This constraint is paramount and must be strictly enforced for
each ShortQ/ShortA pair:
     **One Specific Fact:** Each ShortQ must ask for *one single, specific piece of information*.
     **Only One Correct Answer (within the text):** Critically, based *solely* on the provided reference text, there must be *only
one possible correct answer* to the ShortQ.
     **Mandatory Verification:** Before finalizing any ShortQ, you MUST verify that no other statement or detail *within the
provided text* could also serve as a correct answer to that specific question.
     **AVOID Ambiguity from Lists/Examples:** If the text presents multiple examples, types, reasons, methods, individuals within a
category, etc. (e.g., “Art mediums include painting, digital tools, and ink,” or “Key figures were X, Y, and Z”), you MUST NOT
formulate a ShortQ asking for *one* of them (e.g., DO NOT ask “What is *one* art medium mentioned?” or “Name *an* important
figure.”). Such questions inherently violate the single-answer rule because the text itself provides multiple valid options in that
context.
     +  **Ensure Subject Uniqueness (Open Domain Test):** The specific entity, event, concept, person, or work being asked about in
the question MUST be identifiable *without ambiguity* even when considered outside the context of the source article. Ask yourself:
“If this question were encountered alone, would the subject be clear?”
     +     -  **INVALID Example:** “Which organization funded Short et al.’s work?” is INVALID if “Short et al.’s work” is not a
globally famous, uniquely identifiable publication/project (like “Einstein’s theory of relativity”). It improperly relies on the
implicit context (“the work mentioned in this article”).
     +     -  **VALID Example:** “What year was the Treaty of Versailles signed?” is VALID because “The Treaty of Versailles” is a
globally unique and identifiable historical event.
     +     -  **Guideline:** Avoid questions where the subject is a vague reference (e.g., “the study’s findings”, “their main
conclusion”, “Smith’s 2020 paper” unless that specific paper is uniquely identifiable globally).
     **Target Suitable Facts:** Focus ShortQs on unique identifiers (e.g., the *specific name* of the *first* person to do X),
distinct dates/years associated with singular events, uniquely defined terms (like *Pax Romana*, if defined as a singular concept
in the text), precise numerical values or quantities tied to a specific context, or the outcome of a specific, singular event
described, **ensuring the subject meets the Open Domain Test above.**
  5.  **Concise Short Answers:** Each ShortA (ShortA1-ShortA5) must directly state the single fact requested by its corresponding
ShortQ, using fewer than 10 words.
  6.  **Structured Longform Composition:**
     **LongQ Construction:** The LongQ must explicitly integrate the five preceding ShortQs, typically by listing them as points to
be covered (e.g., “Discuss [Topic] by addressing: (1) [Content of ShortQ1], (2) [Content of ShortQ2]. . . ”).
     **LongA Synthesis:** The LongA must synthesize *only* the five ShortAs (ShortA1-ShortA5) into a single, coherent,
natural-sounding paragraph. It must flow well and not sound robotic. DO NOT introduce any facts or details not present in the
ShortAs. Avoid using bullet points or numbered lists in the LongA.
  7.  **Strict Output Format:** Produce the entire output as a single block of text, strictly adhering to the following
CSV-friendly format on one line. Use commas as separators between key-value pairs. Include NO extra text, commentary,
introductions, or line breaks before or after this formatted string:
Provide One-Line Output Template
"ShortQ1 : <value>, ShortA1 : <value>, ShortQ2 : <value>, ShortA2 : <value>, ShortQ3 : <value>, ShortA3 : <value>,
ShortQ4 : <value>, ShortA4 : <value>, ShortQ5 : <value>, ShortA5 : <value>,
LongQ : <value>,LongA : <value>"
**Reference Article:**
{text}
—
**Generate Now:**
```

Table 3: Short-Long form query and answer generation prompt for o3-mini-high

Prompt

```
**Role and Objective:**
You are a meticulous AI Fact Adherence Evaluator. Your primary objective is to perform a rigorous, objective assessment of a
“Generated Short Answer” by comparing it against a “Ground Truth Short Answer” in the context of a specific “Question”. Your
evaluation must determine if the “Generated Short Answer” accurately and meaningfully conveys the same information as the “Ground
Truth Short Answer” or provides an equally valid factual answer to the “Question”, paying close attention to any specificity
implied by the question and ground truth.
Core Principles for Evaluation:
1. Objectivity: Base your evaluation strictly on the provided information and the criteria outlined below. Avoid personal biases or
assumptions.
2. Factual Accuracy: The paramount consideration is whether the “Generated Short Answer” is factually correct in relation to the
“Question” and “Ground Truth”.
3. Semantic Meaning: Focus on the core meaning and informational content.
4. Contextual Relevance: The “Generated Short Answer” must directly and relevantly answer the “Question”.
Input Data Structure:
You will be provided with the following for each evaluation task:
* Question (ShortQ): SHORT-QUESTION (The specific query the answer should address)
* Ground Truth Short Answer (ShortA): SHORT-ANSWER (The pre-validated correct answer, which also sets the expected level of
specificity for certain types of questions)
* Generated Short Answer (GeneratedShortA): GENERATE-SHORT-ANSWER (The answer to be evaluated)
Detailed Evaluation Criteria and Scoring (Output 0 or 1):
A. Score 1 (Correct) if ANY of the following conditions are met:
1.  Direct Semantic Equivalence:
    * The “Generated Short Answer” conveys the same essential information as the “Ground Truth Short Answer” and accurately answers
the “Question”.
    * Differences in phrasing, sentence structure, or the use of synonyms are acceptable as long as the core meaning is preserved.
2.  Subset/Superset Equivalence (Strict Application Regarding Specificity):
    * If GeneratedShortA is a more specific (subset) version of ShortA (e.g., ShortA: “Dog”, GeneratedShortA: “Labrador Retriever
dog”), it can be correct if it still fundamentally answers ShortQ accurately and doesn’t introduce inaccuracies.
    * If GeneratedShortA is a more general (superset) version of ShortA, it can be correct ONLY IF:
        * It still fundamentally and accurately answers ShortQ.
        * It doesn’t introduce inaccuracies or change the core fact(s) required to answer ShortQ.
3.  World Knowledge Override (Strict Application - Use Sparingly):
  This applies ONLY IF:
    * The “Generated Short Answer” is factually incorrect OR insufficiently specific when compared directly to the “Ground Truth
Short Answer” based on the criteria above.
    * AND The “Generated Short Answer” is a demonstrably true, widely accepted, and commonly known fact that also correctly,
directly, and with appropriate specificity answers the “Question”.
    * AND The “Generated Short Answer” is not a niche, controversial, or overly obscure fact.
  Checklist before applying World Knowledge Override:
    1.  Does GeneratedShortA directly and unambiguously answer ShortQ with the necessary specificity? (If no, score 0)
    2.  Is GeneratedShortA factually true based on broad, verifiable common knowledge? (If no, score 0)
    3.  If ShortQ demands specificity, is GeneratedShortA a better or equally valid specific answer to that demand than ShortA? (If
no, or if ShortA’s specificity is contextually more appropriate, score 0)
    4.  Does GeneratedShortA introduce ambiguity or miss critical nuances that ShortA captures, especially regarding specificity?
(If yes, score 0)
  Example:
    ShortQ: Who is considered the primary inventor of the telephone?
    ShortA: Alexander Graham Bell.
    GeneratedShortA: Antonio Meucci conceived the telephone first. (Score: 1, IF the LLM’s world knowledge strongly supports Meucci
as a more accurate answer to “primary inventor” despite Bell’s common association, and this is a well-established historical
correction. This is a high bar.)
4.  Correct Vagueness:
  * Sometimes, the answer can be correct but vague. For example, if a question says ’Into what must a geometric shape be divided to
be symmetric?’, the ground truth answer is ’Two or more identical pieces’, the generated answer is ’A shape must be divided into
two halves’. The generated answer here is correct.
  * Similarly, for a technical question, the question could be ’How is the fractal-like shape obtained?’, Ground-truth answer
’Finite subdivision rule’ and the Generated answer here ’Fractals are created by repeating a pattern at different scales.’ is
correct here. It is not exactly but the meaning of both is the same.
  * For non-technical questions, like ’What is literary criticism?’, The ground truth is ’study, evaluation, and interpretation of
literature’ and the generated answer is ’Literary criticism is the analysis and interpretation of written works.’. The generated
answer here is correct.
  * Partial Correction: If the answer is partially correct, Apply World Knowledge Override and see if it can be correct FOR the
question. If so, it is correct.
B. Score 0 (Incorrect) if ANY of the following conditions are met:
1.  Factual Inaccuracy: The “Generated Short Answer” is factually incorrect.
2.  Semantic Dissimilarity: The “Generated Short Answer” conveys a different meaning.
3.  Irrelevance: The “Generated Short Answer” does not answer the “Question”.
4.  Contradiction: The “Generated Short Answer” contradicts the “Ground Truth Short Answer” and does not meet the stringent
criteria for “World Knowledge Override”.
5.  Hallucination: The “Generated Short Answer” introduces general knowledge, which alters the answer’s validity and the
hallucination is severe.
For specificity, you have to judge the Question and see if it requires the answer to be exact and specific. These are often
scientific, historical questions where there is only 1 correct answer. If the questions expects a specific answer, only the most
closely related generated answer should be correct.
### OUTPUT FORMAT
Return ONE character only: 1 or 0.
### INPUT
Question: {q}
Ground-truth: {gt}
Candidate: {cand}
<END_PROMPT>
```

Table 4: Prompt for evaluating short-form answers against ground-truth short-form answers.

Prompt

```
***Role and Objective:***
You are an AI Comprehensive Answer Evaluator. Your task is to dissect a “Generated Long Answer” and meticulously assess its
coverage and accuracy concerning five distinct sub-facts, each defined by a “Short Question” and its “Ground Truth Short Answer”.
The “Generated Long Answer” is intended to synthesize these five pieces of information. You must pay close attention to any
specificity implied by each sub-question and its corresponding ground truth.
**Core Principles for Evaluation:**
1. Objectivity: Base your evaluation strictly on the provided information and the criteria outlined below. Avoid personal biases
or assumptions.
2. Factual Accuracy: The paramount consideration is whether the “Generated Short Answer” is factually correct in relation to the
“Question” and “Ground Truth”.
3. Semantic Meaning: Focus on the core meaning and informational content.
4. Contextual Relevance: The “Generated Short Answer” must directly and relevantly answer the “Question”.
### SUB-QUESTIONS & GT
1. {q1}
    → GT-1: {a1}
2. {q2}
    → GT-2: {a2}
3. {q3}
    → GT-3: {a3}
4. {q4}
    → GT-4: {a4}
5. {q5}
    → GT-5: {a5}
### CANDIDATE LONG ANSWER
{cand_long}
**Detailed Evaluation Task and Scoring (List of 5 scores [0 or 1]):**
For EACH of the 5 Sub-Facts (iterate from Sub-Fact 1 to Sub-Fact 5):
1. **Isolate Focus:** Concentrate on the current Sub-Fact i (defined by ShortQ[i] and ShortA[i]).
2. **Locate Relevant Information:** Scrutinize the “Generated Long Answer” to identify the sentence(s) or phrase(s) that attempt
to address ShortQ[i].
    * If no part of “Generated Long Answer” appears to address ShortQ[i], assign a score of 0 for this sub-fact and move to the next.
3. **Evaluate Located Information:** If relevant information is found, compare it against ShortA[i] using the following criteria,
which mirror the detailed logic of the Short Answer Evaluation:
A. Score 1 (Correct) if ANY of the following conditions are met:
1. Direct Semantic Equivalence:
    * The “Generated Short Answer” conveys the same essential information as the “Ground Truth Short Answer” and accurately answers
the “Question”.
    * Differences in phrasing, sentence structure, or the use of synonyms are acceptable as long as the core meaning is preserved.
2. Subset/Superset Equivalence (Strict Application Regarding Specificity):
    * If GeneratedShortA is a more specific (subset) version of ShortA (e.g., ShortA: “Dog”, GeneratedShortA: “Labrador Retriever
dog”), it can be correct if it still fundamentally answers ShortQ accurately and doesn’t introduce inaccuracies.
    * If GeneratedShortA is a more general (superset) version of ShortA, it can be correct ONLY IF:
        * It still fundamentally and accurately answers ShortQ.
        * It doesn’t introduce inaccuracies or change the core fact(s) required to answer ShortQ.
3. World Knowledge Override (Strict Application - Use Sparingly):
  Applies ONLY IF:
    * The “Generated Short Answer” is factually incorrect OR insufficiently specific when compared directly to the “Ground Truth
Short Answer”.
    * AND The “Generated Short Answer” is a demonstrably true, widely accepted, and commonly known fact that also correctly,
directly, and with appropriate specificity answers the “Question”.
    * AND The “Generated Short Answer” is not a niche, controversial, or overly obscure fact.
  Checklist:
    1. Does GeneratedShortA directly and unambiguously answer ShortQ with the necessary specificity? (If no, score 0)
    2. Is GeneratedShortA factually true based on broad, verifiable common knowledge? (If no, score 0)
    3. If ShortQ demands specificity, is GeneratedShortA a better or equally valid specific answer than ShortA? (If no, or if
ShortA’s specificity is more appropriate, score 0)
    4. Does GeneratedShortA miss critical nuances that ShortA captures? (If yes, score 0)
  Example:
  ShortQ: Who is considered the primary inventor of the telephone?
  ShortA: Alexander Graham Bell.
  GeneratedShortA: Antonio Meucci conceived the telephone first. (Score: 1, IF world knowledge supports Meucci as more accurate.)
4. Correct Vagueness:
  * Sometimes the generated answer is correct but vague (e.g., Question: ’Into what must a geometric shape be divided to be
symmetric?’, ShortA: ’Two or more identical pieces’, Generated: ’Two halves’ → correct).
  * Similar logic applies for technical and non-technical contexts, as long as meaning is preserved.
  * Partial Correction: If partially correct, apply World Knowledge Override to decide.
B. Score 0 (Incorrect) if ANY of the following hold:
1. Factual Inaccuracy.
2. Semantic Dissimilarity.
3. Irrelevance.
4. Contradiction not justified by World Knowledge Override.
5. Severe Hallucination.
For specificity, judge whether the Question expects an exact and specific answer (common in science/history). If so, only the most
precise matching Generated answer should be correct.
**Handling Complexities:**
* Information may be split across sentences.
* Do not penalize answer order; evaluate each fact independently.
* Prefer explicit statements. If heavily implied, err toward 0 unless undeniable.
### OUTPUT FORMAT
Return exactly a JSON list of 5 ints, e.g. [1,0,1,1,0]
```

Table 5: Prompt for evaluating long-form answers against five short-form ground truth facts.

Instruction for Short-form QA
`Answer the question with factual single sentence response for the Topic: topic. Question: question`
Instruction for Long-form QA
`Answer to the question should answer everything in the question in a clear and concise manner.Question: long_question`

Table 6: Instruction for short and long-form QA.