Separating Factual Knowledge from Linguistic Competence in LLMs: A Systematic Survey

Anonymous ACL submission

Abstract

Large language models (LLMs) encode facts and their linguistic expressions jointly in their parameters. This tight entanglement makes factual knowledge in LLMs difficult to inspect and update. To address these issues, recent research has called for the separation of factual knowledge from linguistic competence. However, existing work remains fragmented and lacks a unified framework to systematically characterize this direction. We bridge this gap by (1) systematically organizing prior work through the lens of linguistic–factual knowledge separation, (2) comprehensively reviewing evaluation protocols to quantify separation, and (3) highlighting key open questions to inspire future research in this area.

1 Introduction

Large Language Models (LLMs) simultaneously store factual knowledge and linguistic competence within shared parameters, a paradigm that introduces three fundamental limitations. First, it makes knowledge updating prohibitively costly, as the lack of a surgical editing mechanism forces resource-intensive retraining or fine-tuning for even minor factual corrections. Second, this parameter-wise storage hinders knowledge verification, creating a *black box* in which tracing information sources is difficult, thus increasing the risk of confident hallucinations. Third, entanglement restricts the induction of abstract knowledge. Because facts are tied to the linguistic contexts in which they appear, LLMs may fail to generalize factual associations to logically related forms. For example, due to the *reversal curse* (Berglund et al., 2024), a model trained on the statement *"Marie Curie discovered radium"* may still fail to correctly answer *"Who discovered radium?"*.

Driven by these critical bottlenecks, a growing body of research advocates for separating factual knowledge from linguistic competence in LLMs (Mahowald et al., 2024; Collado-Montañez, 2024; Chen, 2024). However, the literature on separation is scattered across several key areas, including knowledge editing (Meng et al., 2022, 2023), knowledge localization (Dai et al., 2022; Yao et al., 2024), and knowledge-limited pretraining (Zhao et al., 2025a), with disparate goals, assumptions, and evaluation protocols. This motivates a survey that connects these threads, synthesizes them within a unified separation framework, and provides a structural foundation for future research in this area.

Our main contributions are summarized as follows. (1) We systematically review prior work through the lens of factual–linguistic separation (Section 3). We categorize existing methods by where separation is realized, yielding two broad paradigms: internalized separation, which isolates factual knowledge within the model, and externalized separation, which shifts factual knowledge outside the model. (2) We survey evaluation protocols to quantify separation (Section 2). We formulate evaluation as a mutual-influence problem: whether factual knowledge remains stable under linguistic perturbations (e.g., paragraphs and multilingual variants) and whether linguistic competence (e.g., language understanding and generation) is preserved under factual suppression. (3) We outline key open questions (Section 4), including the narrow range of knowledge types currently addressed by separation methods, the lack of unified evaluation protocols, and the safety risks that separation may introduce.

Difference to Existing Surveys Prior surveys regarding factual knowledge in LLMs can be broadly grouped into: (1) Factuality and hallucination surveys (Wang et al., 2024e, 2023; Huang et al., 2025a; Dang et al., 2025) examine why LLMs generate fabricated and ungrounded statements and how to mitigate such errors. (2) Knowledge editing surveys (Yao et al., 2023; Zhang et al., 2024a) focus on

how to locally update factual knowledge without full retraining. (3) Knowledge probing and boundary surveys (Li et al., 2025; Wen et al., 2024; Zhang et al., 2024d) examine which knowledge LLMs actually contain. (4) Knowledge mechanism surveys (Rai et al., 2024; Ranaldi, 2025) focus on knowledge diagnosis and explanation, characterizing how LLMs store, access, and update knowledge. While diverse in focus, none of them takes the perspective of factual–linguistic separation. As far as we know, we are the first to review existing works from this angle.

2 Evaluation

In this section, we synthesize evaluation protocols to quantify separation. We evaluate separation along two axes: factual consistency under linguistic perturbations and preservation of linguistic competence after factual knowledge is separated from the model (see Appendix C for more details).

2.1 Factual Consistency

We examine the stability of factual knowledge under two forms of linguistic perturbations: (1) paraphrasing and (2) cross-lingual variation.

Paraphrases Factual Consistency. To evaluate factual consistency under paraphrases, we review question-answering (QA) datasets in which each question is associated with multiple paraphrased formulations. Consistent answers across these paraphrases indicate successful separation between factual knowledge and linguistic form. We categorize benchmarks based on how paraphrases are constructed: manually or automatically.

Manually. ComQA (Abujabal et al., 2019) manually groups 11,214 real-world questions from WikiAnswers into 4,834 semantically equivalent clusters, each associated with a single underlying answer. ParaRel (Elazar et al., 2021) writes 328 paraphrases for 38 binary relations to test knowledge consistency. PopQA-TP (Rabinovich et al., 2023) constructs semantically equal templates for each relation in PopQA (Mallen et al., 2023). SCORE (Nalbandyan et al., 2025) is a robustness evaluation framework that manually constructs 10 semantically equivalent prompts for the questions in MMLU, AGIEval, and MATH.

Automatically. PAQ (Lewis et al., 2021) employs a neural question-generation model to generate diverse question variants, resulting in a large-scale dataset of 65 million QA pairs. LPAQA (Jiang et al., 2020b) utilizes back-translation and corpus mining to expand LAMA (Petroni et al., 2019) templates. More recently, Raj et al. (2022) adopts 3 paraphrase models (DocT5Query, QC, and GPT-3) to generate 8,956 paraphrased questions, and then filters hard questions (the semantically inconsistent ones) with a trained paraphrase detection model, resulting in a curated set of 3,962 paraphrases spanning 817 original questions. Lunardi et al. (2025) extends this paradigm by leveraging LLM prompting (e.g., GPT-4o mini) to generate 5 alternative phrasings for questions in 6 established benchmarks (ARC-C, HellaSwag, etc.). RML-LAMA (Cao et al., 2026) employs Claude 3 Opus while MyriadLAMA (Zhao et al., 2024) utilizes GPT-4 for paraphrasing, using the manual template as the seed. RoParQ (Choi, 2025) also uses LLMs (Gemini 2.5 Flash Lite and Claude 3.5) to construct paraphrases for questions in MCQA.

The paraphrastic variation in these datasets is often limited, as it is shaped by annotators' rewriting habits or generator-specific stylistic preferences, and may therefore provide only a narrow view of the linguistic variation needed to rigorously evaluate knowledge–language separation.

Cross-Lingual Factual Consistency. To evaluate cross-lingual factual consistency, we review QA datasets in which the same question is available in multiple language variants. Consistently giving the same answer across these variants indicates successful separation between factual knowledge and linguistic form.

The following benchmarks instantiate the evaluation setting. XQuAD (Artetxe et al., 2020) translates 1,190 QA pairs from SQuAD v1.1 into 10 languages by professional translators. MLQA (Lewis et al., 2020a) manually translates English Wikipedia questions into 6 additional languages, but provides only partial alignment, with each question available in roughly 4 languages on average. In contrast, MKQA (Longpre et al., 2021a) and Mintaka (Sen et al., 2022) provide fully aligned multilingual questions, covering all 26 and 8 languages, respectively. mLAMA (Kassner et al., 2021) and X-FACTR (Jiang et al., 2020a) are constructed by translating cloze questions in LAMA to 53/25 languages, and Bi-ZsRE (Wang et al., 2024a) and MzsRE (Wang et al., 2024c) are constructed by translating ZsRE queries into 2/12 languages, followed by filtering and alignment to ensure that each query has the same answer. MMLU-ProX (Xuan et al., 2025) and Global-MMLU (Singh et al., 2025)

extend the MMLU (Hendrycks et al., 2021) to 29 and 42 languages, respectively, by using human-machine hybrid translations. More recently, WebFAQ (Dinzinger et al., 2025) automatically identifies semantically equivalent bilingual questions using multilingual embeddings, resulting in approximately 1.5M aligned question pairs across 1,001 language combinations. WebFAQ-2.0 (Dinzinger et al., 2026) further expands this to over 14.3M bilingual question pairs across 3,970 language combinations.

However, existing cross-lingual benchmarks suffer from limited language coverage, spanning only a small fraction of the world's 7,170 living languages and thereby undermining their validity for assessing factual consistency across natural languages.

2.2 Linguistic Competence Preservation

In this section, we evaluate separation from a complementary perspective: the stability of linguistic competence when factual knowledge is separated from the model. We organize existing protocols around two broad aspects of linguistic competence: language understanding and language generation.

Language Understanding. This section summarizes benchmarks that can be used to assess the model's language-understanding capability after separating factual knowledge. We focus on two core dimensions: semantics and syntax.

Semantics. WiC (Pilehvar and Camacho-Collados, 2019) targets word-level semantic disambiguation by asking whether a target word conveys the same sense across different contexts. MRPC (Dolan and Brockett, 2005) and QQP [1] evaluate whether two sentences express equivalent meanings, testing the model's ability to capture nuanced semantic similarities and differences. PAWS (Zhang et al., 2019) creates challenging sentence pairs with high lexical overlap, where minimal structural edits flip the meaning. PARAPHRASUS (Michail et al., 2025) aggregates 10 datasets to capture a wide range of domains, facilitating a more comprehensive assessment of semantic understanding. Xcomps (He et al., 2025) probes conceptual understanding, such as *taxonomic*, while CxMP (Oba and Sugawara, 2026) evaluates their ability to handle nuanced constructions, such as *let alone* and *comparative correlatives*. High performance on these datasets suggests that the model can capture semantic distinctions even when factual knowledge is suppressed.

Syntax. The Corpus of Linguistic Acceptability (CoLA) (Warstadt et al., 2019) contains sentences annotated as grammatically acceptable or unacceptable, covering diverse syntactic phenomena such as agreement, word order, and long-distance dependencies. BLiMP (Warstadt et al., 2020) consists of minimal sentence pairs that isolate specific grammatical contrasts, with one well-formed sentence and one containing a targeted violation. This design enables precise evaluation of grammatical competence. Strong performance—reflected in high Matthew's Correlation Coefficient on CoLA and high accuracy on BLiMP—suggests that the model's grammatical competence remains intact even after factual knowledge has been suppressed.

Language Generation. This section surveys protocols for evaluating language capabilities after the separation of factual knowledge.

Reference-based benchmarks, such as text summarization (e.g., CNN/DailyMail (Nallapati et al., 2016)) and machine translation (e.g., WMT (Bojar et al., 2016)), offer practical test beds for assessing language-generation quality. High ROUGE (Lin, 2004), BLEU (Papineni et al., 2002), MAUVE (Pillutla et al., 2021), and BARTScore (Yuan et al., 2021) suggest that the model's outputs remain close to human-written references and thus are fluent, natural, and well-formed.

Reference-free metrics provide another way to assess language-generation quality after separating factual knowledge. For example, Schafhalter et al. (2024) uses next-token prediction accuracy, assuming that a model with preserved language generation capability should remain capable of predicting natural continuations in text. GPTScore (Fu et al., 2024) adopts GPT-based log-likelihoods as the evaluation signal. Higher log-probabilities indicate closer alignment between the separated model's outputs and GPT's learned linguistic distribution, suggesting better well-formedness, fluency, and coherence. Similarly, Qi et al. (2025) uses perplexity on WikiCorpus as a proxy for linguistic competence evaluation, under the assumption that lower perplexity indicates better alignment with natural language distributions. LLM-as-a-judge (Liu et al., 2023a; Kim et al., 2023; Gu et al., 2024b), such as G-Eval (Liu et al., 2023b), offers a more up-to-date evaluation paradigm. By role-based promoting (e.g., "You are an expert linguistic evaluator. Rate the fluency, coherence, style (for-

[1] https://huggingface.co/datasets/SEACrowd/id_qqp

mal or informal) of the text on a scale of 1–5"), top-edged LLMs (e.g., Llama3-70B) can be employed to provide holistic assessments of generation quality. Despite its cost, human evaluation (Elangovan et al., 2024; Tam et al., 2024) remains the most reliable approach for assessing language-generation quality, typically covering dimensions such as fluency (Manning et al., 2020; Gao et al., 2023b), readability, coherence (Van der Lee et al., 2021), and logical consistency (De Wynter et al., 2023).

Besides these individual evaluation protocols, BenchMAX (Huang et al., 2025b) and SEA-HELM (Susanto et al., 2025) provide the timely and comprehensive evaluation suites to evaluate LLM linguistic competency.

However, it remains difficult to evaluate linguistic competence independently of factual knowledge using existing benchmarks, since many language understanding and generation tasks also require world knowledge and commonsense reasoning.

3 Methodology

In this section, we systematically synthesize prior work through the lens of the linguistic-knowledge separation. We categorize prior methods based on where the separation resides: (1) *Internalized separation* disentangles factual knowledge from linguistic competence inside the model. (2) *Externalized separation* moves factual knowledge into external resources, allowing the model to focus mainly on language understanding and generation. Our literature search strategy are provided in Appendix A.

3.1 Internalized Separation

Internalized separation can be realized through two main strategies: (i) post-hoc localization, which identifies factual knowledge units inside the model, and (ii) training-time isolation, which confines factual knowledge to specific modules via additional training.

3.1.1 Post-hoc Localization

Post-hoc localization aims to identify which units inside LLMs (e.g., neurons, layers, attention heads, and circuits) store factual knowledge. Existing methods typically estimate the attribution of these units to a given fact by formulating factual knowledge as a subject–relation–object recall problem. For example, given the prompt *Paris is the capital of* __., where Paris is the subject s, capital is the relation r, and France is the object o, the attribution of an internal unit can be measured by its contribution to the probability of predicting the correct object, $P(o \mid s, r)$.

Gradient-based Attribution. This line of work leverages Integrated Gradients (IG) to track which internal neurons most strongly influence a model's factual recall. Compared to standard gradients, IG mitigates saturation effects, a limitation that causes vanilla gradients to vanish and consequently underestimate neuron significance, thereby providing a more faithful estimate of factual importance.

Dai et al. (2022) represents one of the earliest and most influential works in this line of research. They scale the neuron's activation (a weight dimension in a feed-forward network (FFN) layer) from 0 to its original value, then compute the integrated gradients of $P(o|s, r)$ along this path. Highly integrated gradients indicate that the neuron strongly influences recall of the fact and thus should be labeled a knowledge neuron. To filter out false positives, they retain only those neurons that exhibit consistently high attribution scores (i.e., integrated gradients) across semantically equivalent but linguistically diverse phrasings. Building on this work, QRNCA (Chen et al., 2025) reveals that domain knowledge is primarily localized in the middle layers of LLMs, whereas linguistic information is more diffusely distributed across neurons, suggesting that linguistic processing is engaged throughout all layers. AMIG (Chen et al., 2024) and MKE (Zhang et al., 2025) detect knowledge neuron sets for the same fact in different languages and derive *language-independent knowledge* neurons by intersecting these sets. To further improve robustness, Cao et al. (2026) moves beyond simple intersection and explicitly models attribution uncertainty, selecting neurons that contribute strongly and consistently across paraphrases and languages.

Gradient-based attribution mainly captures correlation rather than causation. A neuron may receive a high attribution score simply because it is correlated with factual recall, even if the model does not causally depend on it; removing or perturbing the neuron may leave the prediction unchanged.

Causal Intervention. Causal intervention addresses this limitation by directly testing the unit's necessity for factual recall. It validates the necessity by perturbing selected units via ablation or activated patching and measuring whether factual recall deteriorates.

Ablation-based methods suppress neurons or other units and examine whether the associated

Method	Separation Method	Knowledge Location	Knowledge Type	Train Req.	Benchmark
Dai et al. (2022)	Gradient	Neurons	Entity	✗	PARAREL
Chen et al. (2024)	Gradient	Neurons	Entity	✗	mLAMA
Chen et al. (2025)	Gradient	Neurons	Entity, QA Facts	✗	MMLU, mLAMA
Zhang et al. (2025)	Gradient	Neurons	Entity	$\checkmark^{aux}$	Bi-ZsRE, MzsRE
Cao et al. (2026)	Gradient	Neurons	Entity	$\checkmark^{aux}$	RML-LAMA
Yu and Ananiadou (2024)	Ablation	Neurons	Entity	✗	TriviaQA
Geva et al. (2023)	Ablation	FFNs, Heads	Entity	✗	CounterFact
Yao et al. (2024)	Ablation	Circuits	Entity	✗	LRE
Meng et al. (2022)	Patching	FFNs	Entity	✗	CounterFact
Zhang and Nanda (2024)	Patching	FFNs, Heads	Entity	✗	CounterFact
Wang et al. (2021)	Adapter	Sub-Modules	Entity, Relation	✓	T-REx, FewRel
Schafhalter et al. (2024)	Adapter	Sub-Modules	Domain Knowledge	✓	Math, Code, English
Neeman et al. (2023)	Counterfact Fine-Tune	External	Entity	✓	NQ
Li et al. (2023)	Counterfact Fine-Tune	External	Entity, QA Facts	✓	SQuAD 2.0, T-REx, QASC, TriviaQA
Shi et al. (2024)	Steering	External	Entity, Contextual Facts	✗	CNN-DM, XSum, NQ-Swap, MemoTrap
Santosh et al. (2025)	Steering	External	Domain Knowledge Contextual Facts	✗	ECtHR, ECHR,LegalBench BillSum, MultiLexSum
Yu et al. (2023)	Steering	External	Entity	✗	Country-Capitals
Zhao et al. (2025b)	Steering	External	Entity	$\checkmark^{aux}$	NQSwap,Macnoise
Cohen et al. (2026)	Entity-Anon. Pretraining	External	Entity	✓	SuperGLUE, LAMA SQuAD, NQ, FEVER
Zhao et al. (2025a)	Entity-Masked Pretraining	External Database	Entity	✓	FactScore, T-REx, PopQA, NLU
Li et al. (2022)	Differential Private	External	Personal Information	✓	GLUE, E2E,DART
Wang et al. (2024b)	Differential Private	External	Personal Information	✓	SST-2, QQP,QNLI

Table 1: A Taxonomy of Linguistic–Knowledge Separation. *Train. Req.* indicates training required or not: ✗= no training; $\checkmark^{aux}$ = auxiliary training; ✓= direct training or optimization for separation.

factual information is disrupted. Specifically, Yu and Ananiadou (2024) suppresses selected neurons by setting their activations to zero and quantifies the resulting drop in factual recall, measured by the decrease in the probability of the correct object: $P_{\text{orig}}(o|s,r) - P_{\text{sup}}(o|s,r)$. Neurons that exhibit large performance degradation are considered knowledge neurons. Geva et al. (2023) uses zero ablation to detect knowledge-related attention heads. Given an attention head h, they set $atte(i \leftarrow j)$ to 0 to cut off token i from attending to token j. If the intervention reduces $P(o|s,r)$, the head is considered to play a key role in predicting o and is thus labeled a knowledge-related head. Yao et al. (2024) extends knowledge localization from individual attention heads to knowledge circuits—a set of interacting components (e.g., attention heads and MLPs) (Olah et al., 2020). They ablate edges between components by setting their contributions to zero, and identify those whose removal leads to significant degradation as causal pathways underlying factual prediction.

In contrast, activation patching (Meng et al., 2022, 2023) tests causal relevance from the opposite direction. It first corrupts the model's computation, causing factual recall to fail, and then restores the activation of selected units to their values from a clean run. If restoring a unit recovers the correct prediction, this provides evidence that the unit carries information necessary for the factual association. Zhang and Nanda (2024) argue that Gaussian noising (GN) used in standard activation patching leads to out-of-distribution perturbations in corrupted runs, and propose to use Symmetric Token Replacement (STR) to construct corruption, which swaps the subject with a semantically related but wrong entity (e.g., France → Italy in "Paris is the capital of _").

Operations such as zero ablation, Gaussian corruption, and activation patching can push hidden states outside the model's natural activation distribution. As a result, the observed degradation may

partly reflect the effect of unnatural perturbations, rather than the removal of a genuinely knowledge-bearing unit.

3.1.2 Training-Time Isolation

Training-time isolation promotes factual–linguistic separation by assigning factual knowledge and linguistic competence to different trainable components during training, such as dedicated adapters, experts, or modules.

K-ADAPTER (Wang et al., 2021) promotes separation by assigning factual and linguistic information to different adapter modules. The factual adapter is trained on relation classification, encouraging it to capture entity-relation knowledge, while the linguistic adapter is trained on dependency relation prediction to capture syntactic structure. MAD-X (Pfeiffer et al., 2020) extends this modular view by separating language-specific linguistic knowledge into dedicated language adapters, each trained with masked language modeling, and then adding a task-specific knowledge adapter on top using named entity recognition. DEMix (Gururangan et al., 2022) takes another direction by further decomposing knowledge modules. They train domain-specific knowledge modules (experts) by replacing the standard single feed-forward network (FFN) with a set of parallel FFNs, and then optimizing each FFN on its respective domain corpus (e.g., news, legal, and medical) to enable they have different domain knowledge. Similarly, MoDE (Schafhalter et al., 2024) extends this idea by independently training domain-specialized knowledge modules and then composing them through a lightweight training process.

Training-time isolation augments the base model with additional knowledge- or language-specific modules while keeping the model's original parameters frozen. Although this avoids disrupting the base model's existing linguistic capabilities, it does not achieve full separation, as factual knowledge and linguistic competence remain entangled within the base model's parameters.

3.1.3 Limitations of Internalized Separation

First, internalized separation is sensitive to the probing setup: the identified knowledge-bearing units may change with the prompt template, language, phrasing, or domain corpus, making localization results unstable and hard to generalize. Second, internalized separation typically covers only a limited subset of facts, rather than the full body of factual knowledge, resulting in partial rather than complete separation. Third, although internalized separation reduces the search space for knowledge updates, updating still requires parameter-level interventions rather than direct edits to an explicit knowledge store. This makes updates costly and may introduce side effects, such as over-editing, interference with neighboring facts, or degradation of unrelated knowledge and linguistic behavior (Gu et al., 2024a).

3.2 Externalized Separation

Externalized separation places factual knowledge outside the model (Lewis et al., 2020b; Guu et al., 2020; Karpukhin et al., 2020; Fan et al., 2024), establishing a clearer boundary between linguistic competence and factual storage. There are two ways to implement the external separation: (1) Post-hoc Overriding, and (2) Training-time Attenuating.

3.2.1 Post-hoc Overriding

Post-hoc overriding achieves separation by steering the model to prioritize contextual knowledge over parametric knowledge. Directly prompting the model to use contextual knowledge is highly intuitive; however, parametric knowledge is stubborn: LLMs tend to cling to their parametric knowledge (Xie et al., 2023; Augenstein, 2025) despite explicitly being instructed to use contextual knowledge, especially when there is a conflict between the two kinds of knowledge (Cheng et al., 2024).

Counterfactual Fine-tuning. To address this issue, counterfactual fine-tuning trains models on examples whose answers are entailed by the provided context but contradict the model's parametric knowledge. By training on such counterfactual data, this approach encourages the model to follow contextual knowledge even when it conflicts with parametric knowledge.

The key challenge is *how to construct counterfactual training data*. Existing work mainly follows three strategies. First, Longpre et al. (2021b), DisentQA (Neeman et al., 2023), and KAFT (Li et al., 2023) constructs counterfactual data through entity swapping. They replace entities in the context with alternative ones, creating an explicit conflict between contextual and parametric knowledge. Second, Zhou et al. (2024) constructs counterfactual data through filtering. Instead of manually editing entities, they discard instances that base LLMs can already answer correctly and retain cases where these models fail. Such failure cases are more

likely to expose conflicts between contextual evidence and parametric knowledge, and therefore provide useful training signals for improving context reliance. Third, MQuAKE (Zhong et al., 2023) constructs compositional counterfactuals by editing multi-hop fact chains. Rather than changing a single entity, it modifies linked facts in a reasoning chain, replacing an original chain A-B-C with a counterfactual chain A-B'-C'. This creates contexts in which the final answer depends on following the edited chain rather than on memorized associations.

While counterfactual fine-tuning can train models to prioritize contextual evidence under seen patterns, it does not guarantee out-of-distribution robustness.

Decoding-Based Steering. Decoding-based steering promotes separation at inference time by biasing next-token selection toward context-supported tokens. For instance, context-aware decoding (CAD) and its adaptive variants (Shi et al., 2024; Bi et al., 2024; Wang et al., 2025a; Khandelwal et al., 2025) contrast token distributions computed with and without the provided context. Tokens whose probabilities increase when the context is included are treated as context-supported and are amplified, while tokens that remain likely even without the context are treated as more dependent on parametric knowledge and are down-weighted. Other decoding-time control methods provide complementary ways to bias generation toward contextual knowledge. Controller-guided decoding (Yang and Klein, 2021; Komorowski et al., 2026) provides a more general framework, which steers generation by adding an auxiliary scoring signal to the decoding objective. Copy-biased decoding (Santosh et al., 2025) softly increases the likelihood of generating salient entities and details from the provided passages by interpolating the model distribution with a context-derived copying distribution. Lexically constrained decoding (Hokamp and Liu, 2017; Lu et al., 2021) requires that specified evidence-derived items, such as entities, dates, or numbers, appear in the final answer. We also summarize mechanism-based steering approaches in Appendix B.

These steering methods provide only local and probabilistic control, rather than a principled separation between parametric and contextual knowledge. As a result, such methods cannot guarantee that the final output is globally faithful to the provided context, nor that unsupported parametric priors are fully suppressed.

3.2.2 Training-Time Attenuation

Post-hoc overriding only suppresses the use of parametric knowledge, whereas training-time attenuation goes towards a true separation by directly weakening parametric knowledge. In this sense, the LLM is transformed from a versatile model into a pure language model (Collado-Montañez et al., 2025).

Entity-Anonymized Pretraining. Entity-anonymized pretraining weakens entity-centered factual knowledge in LLMs by masking or replacing entity mentions during pretraining. This design naturally aligns with knowledge graphs, where facts are organized around entities and relations. KLLMs (Cohen et al., 2026) uses an OntoNotes NER model to detect entities and replace them with their types, in a total of 18 classes. It then trains the model on the resulting de-factualized corpus to reduce memorization of entity-related knowledge. Follow the same idea, Boutet et al. (2026) utilizes a mix of rule-based detectors and NER tools to identify direct- and quasi- entities, then replaces them with either type-level placeholders (e.g., <NAME>, <DATE>) before continuing masked language modeling (MLM) training on the sanitized text, thereby weakening linkable identity and fact cues that models tend to memorize. Pseudonymization (Vakili et al., 2024; Cabrera-Diego and Gheewala, 2024) further enforces entity-level consistency by conducting co-reference resolution after NER, mapping all mentions of the same entity to a stable but semantically meaningless pseudonym, thereby preserving discourse coherence while stripping retrievable factual identity. Moving beyond pure masking, LMLM (Zhao et al., 2025a) shifts from parametric memorization to the retrieval of externalized knowledge. It extracts structural triples $\langle e, r, v \rangle$ into an external database and trains the model to dynamically query facts, thereby decoupling language modeling from internal factual storage.

A limitation of these approaches is the distribution shift introduced by replacing natural text with placeholders, pseudonyms, or lookup calls, which makes sanitized pretraining less compatible with ordinary real-world language use.

Gradient Penalization Gradient penalization weakens the memorization of personally identifiable information (PII) by penalizing its gradients during training. Although originally developed for privacy protection, these methods closely align

with our framework because PII, such as phone numbers, email addresses, and residential locations, constitutes a specific form of factual knowledge about individuals. Reducing its memorization can therefore be viewed as selectively weakening factual content within the model's parameters while preserving general linguistic learning.

One line of work is unlearning. Eldan and Russinovich (2023) and Jang et al. (2023) execute unconstrained gradient ascent (negative updates) on targeted documents to minimize the likelihood of generating fact-bearing tokens. To bound this destructive process, Yao and Xu (2024) propose a joint optimization framework that anchors negative gradients on a "forget set" alongside standard gradient descent on a benign "retain set." More recently, Negative Preference Optimization (NPO) (Zhang et al., 2024b) computes token-level factual priors and dynamically penalizes the optimization only when the model exhibits an overconfident bias towards memorized facts.

Another line of work is differential privacy (DP) (Abadi et al., 2016; Du et al., 2025). DP-based approaches clip individual gradients and add calibrated noise to PII tokens, preventing them from dominating parameter updates and thereby suppressing the retention of sensitive personal information. To make this mechanism feasible for large-scale Transformers, subsequent works focus on two practical dimensions of optimization. The first controls where to apply privacy: Li et al. (2022) and Yu et al. (2022) combine DP with parameter-efficient fine-tuning (PEFT), restricting privacy perturbations to a small subset of weights. The second limits what to protect: Shi et al. (2022) and Wang et al. (2024b) introduce selective local DP, safeguarding only the specific PII tokens within a sequence rather than the entire text. Together, these methods selectively weaken personal factual content by controlling how individual sensitive instances alter model parameters.

However, directly manipulating gradients may introduce substantial training instability and increase the risk of catastrophic forgetting (Maini et al., 2024), where the model degrades in its general language and reasoning capabilities.

3.2.3 Limitation of Externalized Separation

Externalized separation also introduces new challenges: (1) Sensitivity to retrieval quality. Performance is highly dependent on the quality of retrieved evidence. When relevant information is missing, incorrectly retrieved, or poorly ranked, the model is likely to produce low-quality outputs. We discuss strategies for improving retrieval in Appendix F. (2) High engineering complexity. Externalized separation relies on additional retrieval and knowledge-integration modules, which increase system complexity, latency, and cost.

4 Open Challenges

Despite recent progress, linguistic-factual separation in LLMs remains far from solved.

Improved coverage of the separated knowledge. Most existing methods primarily target entity-centered facts, such as names, dates, and simple relations. However, LLMs also encode procedural knowledge (e.g., *the conditional execution steps of bubble sort*) and conditional realities (e.g., *if an asset's supply spikes under fixed demand, its market value drops*). It remains unclear whether current separation methods can generalize beyond entity-level lookups to decouple abstract, consensus-based forms of factual knowledge from the model's linguistic core.

Lack of a unified evaluation framework. A major open challenge is the lack of unified evaluation protocols to quantify the separation between linguistic and factual knowledge. Existing evaluations typically capture only one aspect of separation: paraphrase and multilingual benchmarks test factual stability under linguistic variation, while linguistic benchmarks assess whether language understanding and generation are preserved after separation. Such fragmentation makes systematic comparison across methods difficult.

Safety and security concerns. Separation may improve the transparency and explainability of LLMs, but it may also introduce new risks. For example, external knowledge sources can be poisoned, retrieved evidence can be adversarially manipulated, and weakened parametric knowledge may reduce the model's resilience when external information is unreliable. Understanding the safety trade-offs of separation is therefore crucial.

5 Conclusion

We survey literature in linguistic–factual separation, classifying existing methods into two primary paradigms: internalized and externalized separation. Furthermore, we synthesize evaluation protocols to quantify separation and highlight open questions to inspire future research in this area.

6 Limitations

Given the pace of progress and the breadth of the literature, this survey may not be exhaustive, and some very recent or less visible but relevant work may have been omitted.

Beyond directly measuring, separation can also be indirectly evaluated through downstream-task metrics, such as editing success rate (how well the targeted fact update is applied) (Meng et al., 2022), overriding robustness (how reliably new information overrides stale parametric beliefs) (Ouyang et al., 2025), and conflict-resolution behavior (whether models follow provided evidence when it contradicts memorized knowledge) (Ming et al., 2025). In order to maintain focus, we do not systematically review these complementary evaluation axes.

7 Acknowledgments

AI tools were used only for language polishing and readability improvement. The authors retain full responsibility for the content, analysis, and conclusions of this work.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318.

Abdalghani Abujabal, Rishiraj Saha Roy, Mohamed Yahya, and Gerhard Weikum. 2019. ComQA: A community-sourced dataset for complex factoid question answering with paraphrase clusters. In *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)*, pages 307–317, Minneapolis, Minnesota. Association for Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 4623–4637. Association for Computational Linguistics.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *Preprint*, arXiv:2310.11511.

Isabelle Augenstein. 2025. Understanding the interplay between llms' utilisation of parametric and contextual knowledge: A keynote at ecir 2025. *ACM SIGIR Forum*, 59(2):1–11.

Lukas Berglund, Meg Tong, Maximilian Kaufmann, Mikita Balesni, Asa Stickland, Tomek Korbak, and Owain Evans. 2024. The reversal curse: Llms trained on "a is b" fail to learn "b is a". In *International Conference on Learning Representations*, volume 2024, pages 18623–18642.

Baolong Bi, Shenghua Liu, Lingrui Mei, Yiwei Wang, Pengliang Ji, and Xueqi Cheng. 2024. Decoding by contrasting knowledge: Enhancing llms' confidence on edited facts. *arXiv preprint arXiv:2405.11613*.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, and 1 others. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198.

Antoine Boutet, Lucas Magnana, and Juliette Sénéchal. 2026. Towards the anonymization of the language modeling. *Preprint*, arXiv:2501.02407.

Luis-Adrián Cabrera-Diego and Akshita Gheewala. 2024. Psilence: A pseudonymization tool for international law. In *Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization (CALD-pseudo 2024)*, pages 25–36.

Boxi Cao, Hongyu Lin, Xianpei Han, Le Sun, Lingyong Yan, Meng Liao, Tong Xue, and Jin Xu. 2021. Knowledgeable or educated guess? revisiting language models as knowledge bases. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1860–1874.

Pengfei Cao, Yuheng Chen, Zhuoran Jin, Yubo Chen, Kang Liu, and Jun Zhao. 2026. One mind, many tongues: A deep dive into language-agnostic knowledge neurons in large language models. *Artificial Intelligence*, 353:104490.

Huajun Chen. 2024. Large knowledge model: Perspectives and challenges. *Data Intelligence*, 6(3):587–620.

Lihu Chen, Adam Dejl, and Francesca Toni. 2025. Identifying query-relevant neurons in large language models for long-form texts. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(22):23595–23604.

Yuheng Chen, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. 2024. Journey to the center of the knowledge neurons: discoveries of language-independent knowledge neurons and degenerate

knowledge neurons. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press.

Aileen Cheng, Alon Jacovi, Amir Globerson, Ben Golan, Charles Kwong, Chris Alberti, Connie Tao, Eyal Ben-David, Gaurav Singh Tomar, Lukas Haas, Yonatan Bitton, Adam Bloniarz, Aijun Bai, Andrew Wang, Anfal Siddiqui, Arturo Bajuelos Castillo, Aviel Atias, Chang Liu, Corey Fry, and 46 others. 2025. The facts leaderboard: A comprehensive benchmark for large language model factuality. *Preprint*, arXiv:2512.10791.

Sitao Cheng, Liangming Pan, Xunjian Yin, Xinyi Wang, and William Yang Wang. 2024. Understanding the interplay between parametric and contextual knowledge for large language models. *Preprint*, arXiv:2410.08414.

Minjoon Choi. 2025. Roparq: Paraphrase-aware alignment of large language models towards robustness to paraphrased questions. *arXiv preprint arXiv:2511.21568*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Roi Cohen, Yvan Carré, Nick Lechtenbörger, Hendrik Droste, Lucas Kerschke, Russa Biswas, Gerard de Melo, and Jan Buys. 2026. Knowledgeless language models: Decoupling linguistic competence and factual knowledge.

Jaime Collado-Montañez. 2024. Separating linguistic competence from factual knowledge in large language models. In *NLP-DS@ SEPLN*.

Jaime Collado-Montañez, L Alfonso Ureña-López, and Arturo Montejo-Ráez. 2025. Towards fundamental language models: Does linguistic competence scale with model size? *arXiv preprint arXiv:2509.02225*.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502.

Hoang Anh Dang, Vu Tran, and Le-Minh Nguyen. 2025. Survey and analysis of hallucinations in large language models: attribution to prompting strategies or model behavior. *Frontiers in Artificial Intelligence*, 8:1622292.

Adrian De Wynter, Xun Wang, Alex Sokolov, Qilong Gu, and Si-Qing Chen. 2023. An evaluation on large language model outputs: Discourse and memorization. *Natural Language Processing Journal*, 4:100024.

Michael Dinzinger, Laura Caspari, Kanishka Ghosh Dastidar, Jelena Mitrović, and Michael Granitzer. 2025. Webfaq: A multilingual collection of natural q&a datasets for dense retrieval. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3802–3811.

Michael Dinzinger, Laura Caspari, Ali Salman, Irvin Topi, Jelena Mitrović, and Michael Granitzer. 2026. Webfaq 2.0: A multilingual qa dataset with mined hard negatives for dense retrieval. *arXiv preprint arXiv:2602.17327*.

William B Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the third international workshop on paraphrasing (IWP2005)*.

Hao Du, Shang Liu, and Yang Cao. 2025. Can differentially private fine-tuning llms protect against privacy attacks? In *IFIP Annual Conference on Data and Applications Security and Privacy*, pages 311–329. Springer.

Xeron Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, King Zhu, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, Chujie Zheng, Kaixin Deng, Shuyue Guo, Shian Jia, Sichao Jiang, Yiyan Liao, Rui Li, Qinrui Li, Sirun Li, and 76 others. 2026. SuperGPQA: Scaling LLM evaluation across 285 graduate disciplines. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Aparna Elangovan, Ling Liu, Lei Xu, Sravan Babu Bodapati, and Dan Roth. 2024. Considers-the-human evaluation framework: Rethinking human evaluation for generative large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1137–1160.

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

Ronen Eldan and Mark Russinovich. 2023. Who's harry potter? approximate unlearning in llms. *Preprint*, arXiv:2310.02238.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pages 6491–6501.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. Splade: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on*

Research and Development in Information Retrieval, pages 2288–2292.

Jinlan Fu, See Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. Gptscore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023a. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023b. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore. Association for Computational Linguistics.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Jia-Chen Gu, Hao-Xiang Xu, Jun-Yu Ma, Pan Lu, Zhen-Hua Ling, Kai-Wei Chang, and Nanyun Peng. 2024a. Model editing harms general abilities of large language models: Regularization to the rescue. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16801–16819.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2024b. A survey on llm-as-a-judge. *The Innovation*.

Esmail Gumaan. 2025. Expertrag: Efficient rag with mixture of experts–optimizing context retrieval for adaptive llm responses. *arXiv preprint arXiv:2504.08744*.

Suchin Gururangan, Mike Lewis, Ari Holtzman, Noah A Smith, and Luke Zettlemoyer. 2022. Demix layers: Disentangling domains for modular language modeling. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5557–5576.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Linyang He, Ercong Nie, Sukru Samet Dindar, Arsalan Firoozi, Van Nguyen, Corentin Puffay, Riki Shimizu, Haotian Ye, Jonathan Brennan, Helmut Schmid, and 1 others. 2025. Xcomps: A multilingual benchmark of conceptual minimal pairs. In *Proceedings of the 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 75–81.

Roee Hendel, Mor Geva, and Amir Globerson. 2023. In-context learning creates task vectors. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9318–9333.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Chris Hokamp and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1546, Vancouver, Canada. Association for Computational Linguistics.

Arian Hosseini, Siva Reddy, Dzmitry Bahdanau, R Devon Hjelm, Alessandro Sordoni, and Aaron Courville. 2021. Understanding by understanding not: Modeling negation in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1301–1312.

Xuming Hu, Junzhe Chen, Xiaochuan Li, Yufei Guo, Lijie Wen, Philip S Yu, and Zhijiang Guo. 2024. Towards understanding factual knowledge of large language models. In *The twelfth international conference on learning representations*.

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

Xu Huang, Wenhao Zhu, Hanxu Hu, Conghui He, Lei Li, Shujian Huang, and Fei Yuan. 2025b. Benchmax: A comprehensive multilingual evaluation suite for large language models. *Preprint*, arXiv:2502.07346.

Shayekh Bin Islam, Md Asib Rahman, K S M Tozammel Hossain, Enamul Hoque, Shafiq Joty, and Md Rizwan Parvez. 2024. Open-RAG: Enhanced retrieval augmented reasoning with open-source large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14231–14244, Miami, Florida, USA. Association for Computational Linguistics.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Towards unsupervised dense information retrieval with contrastive learning. *CoRR*, abs/2112.09118.

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. 2023. Knowledge unlearning for mitigating privacy risks in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14389–14408.

Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2023a. LLMLingua: Compressing prompts for accelerated inference of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13358–13376, Singapore. Association for Computational Linguistics.

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2024. Longllmlingua: Accelerating and enhancing llms in long context scenarios via prompt compression. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1658–1677.

Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. 2020a. X-FACTR: Multilingual factual knowledge retrieval from pretrained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5943–5959, Online. Association for Computational Linguistics.

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

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023b. Active retrieval augmented generation. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 7969–7992.

Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. 2024. Cutting off the head ends the conflict: A mechanism for interpreting and mitigating knowledge conflicts in language models. *arXiv preprint arXiv:2402.18154*.

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*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, pages 6769–6781.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. Multilingual LAMA: Investigating knowledge in multilingual pretrained language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3250–3258, Online. Association for Computational Linguistics.

Anant Khandelwal, Manish Gupta, and Puneet Agrawal. 2025. Cocoa: Confidence-and context-aware adaptive decoding for resolving knowledge conflicts in large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6846–6866.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and 1 others. 2023. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*.

Piotr Komorowski, Elena Golimblevskaia, Reduan Achtibat, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. 2026. Attribution-guided decoding. In *The Fourteenth International Conference on Learning Representations*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, and 1 others. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Keith Lehrer and Thomas Paxson. 1969. Knowledge: Undefeated justified true belief. *The Journal of Philosophy*, 66(8):225–237.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020a. Mlqa: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 7315–7330.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Patrick Lewis, Yuxiang Wu, Linqing Liu, Pasquale Minervini, Heinrich Küttler, Aleksandra Piktus, Pontus Stenetorp, and Sebastian Riedel. 2021. PAQ: 65 million probably-asked questions and what you can do with them. *CoRR*, abs/2102.07033.

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. 2023. Large language models with controllable working memory. In *Findings of the association for computational linguistics: ACL 2023*, pages 1774–1793.

Moxin Li, Yong Zhao, Wenxuan Zhang, Shuaiyi Li, Wenya Xie, See Kiong Ng, Tat-Seng Chua, and Yang Deng. 2025. Knowledge boundary of large language models: A survey. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5131–5157.

Xuechen Li, Florian Tramèr, Percy Liang, and Tatsunori Hashimoto. 2022. Large language models can be strong differentially private learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023a. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: Nlg evaluation using gpt-4 with better human alignment. *Preprint*, arXiv:2303.16634.

Shayne Longpre, Yi Lu, and Joachim Daiber. 2021a. MKQA: A linguistically diverse benchmark for multilingual open domain question answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021b. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 7052–7063.

Ximing Lu, Peter West, Rowan Zellers, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Neurologic decoding:(un) supervised neural text generation with predicate logic constraints. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies*, pages 4288–4299.

Riccardo Lunardi, Vincenzo Della Mea, Stefano Mizzaro, and Kevin Roitero. 2025. On robustness and reliability of benchmark-based evaluation of llms. *arXiv preprint arXiv:2509.04013*.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query rewriting in retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5303–5315.

Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in cognitive sciences*, 28(6):517–540.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary Chase Lipton, and J Zico Kolter. 2024. TOFU: A task of fictitious unlearning for LLMs. In *First Conference on Language Modeling*.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Emma Manning, Shira Wein, and Nathan Schneider. 2020. A human evaluation of amr-to-english generation systems. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4773–4786.

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.

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

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and 1 others. 2022. Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147*.

Andrianos Michail, Simon Clematide, and Juri Opitz. 2025. Paraphrasus: A comprehensive benchmark for evaluating paraphrase detection models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8749–8762.

Yifei Ming, Senthil Purushwalkam, Shrey Pandit, Zixuan Ke, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. 2025. Faitheval: Can your language model stay faithful to context, even if "the moon is made of marshmallows". *Preprint*, arXiv:2410.03727.

Charles William Morris. 1938. Foundations of the theory of signs. In *International encyclopedia of unified science*, pages 1–59. Chicago University Press.

Grigor Nalbandyan, Rima Shahbazyan, and Evelina Bakhturina. 2025. Score: Systematic consistency and robustness evaluation for large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 470–484.

Ramesh Nallapati, Bowen Zhou, Cicero Dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Proceedings of the 20th SIGNLL conference on computational natural language learning*, pages 280–290.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2023. Disentqa: Disentangling parametric and contextual knowledge with counterfactual question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10056–10070.

Miyu Oba and Saku Sugawara. 2026. Cxmp: A linguistic minimal-pair benchmark for evaluating constructional understanding in language models. *arXiv preprint arXiv:2602.21978*.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001.

Jie Ouyang, Tingyue Pan, Mingyue Cheng, Ruiran Yan, Yucong Luo, Jiaying Lin, and Qi Liu. 2025. Hoh: A dynamic benchmark for evaluating the impact of outdated information on retrieval-augmented generation. *arXiv preprint arXiv:2503.04800*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 2463–2473.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Long Phan, Alice Gatti, Nathaniel Li, Adam Khoja, Ryan Kim, Richard Ren, Jason Hausenloy, Oliver Zhang, Mantas Mazeika, Dan Hendrycks, Ziwen Han, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, Michael Choi, Anish Agrawal, and 1081 others. 2026. A benchmark of expert-level academic questions to assess ai capabilities. *Nature*, 649(8099):1139–1146.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *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)*, pages 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Yejin Choi, and Zaïd Harchaoui. 2021. MAUVE: human-machine divergence curves for evaluating open-ended text generation. *CoRR*, abs/2102.01454.

Max Ploner, Jacek Wiland, Sebastian Pohl, and Alan Akbik. 2025. Lm-pub-quiz: A comprehensive framework for zero-shot evaluation of relational knowledge in language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 29–39.

Nina Poerner, Ulli Waltinger, and Hinrich Schütze. 2020. E-bert: Efficient-yet-effective entity embeddings for bert. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 803–818.

Siyuan Qi, Bangcheng Yang, Kailin Jiang, Xiaobo Wang, Jiaqi Li, Yifan Zhong, Yaodong Yang, and Zilong Zheng. 2025. In-context editing: Learning knowledge from self-induced distributions. In *The Thirteenth International Conference on Learning Representations*.

Yifu Qiu, Zheng Zhao, Yftah Ziser, Anna Korhonen, Edoardo M Ponti, and Shay B Cohen. 2024. Spectral editing of activations for large language model alignment. *Advances in Neural Information Processing Systems*, 37:56958–56987.

Ella Rabinovich, Samuel Ackerman, Orna Raz, Eitan Farchi, and Ateret Anaby Tavor. 2023. Predicting question-answering performance of large language models through semantic consistency. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 138–154.

Zackary Rackauckas. 2024. Rag-fusion: A new take on retrieval augmented generation. *International Journal on Natural Language Computing*, 13(1):37–47.

Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. 2024. A practical review of mechanistic interpretability for transformer-based language models. *arXiv preprint arXiv:2407.02646*.

Harsh Raj, Domenic Rosati, and Subhabrata Majumdar. 2022. Measuring reliability of large language models through semantic consistency. *arXiv preprint arXiv:2211.05853*.

Leonardo Ranaldi. 2025. Survey on the role of mechanistic interpretability in generative ai. *Big Data and Cognitive Computing*, 9(8):193.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2024. Gpqa: A graduate-level google-proof q&a benchmark. In *First conference on language modeling*.

TYSS Santosh, Youssef Tarek Elkhayat, Oana Ichim, Pranav Shetty, Dongsheng Wang, Zhiqiang Ma, Armineh Nourbakhsh, and Xiaomo Liu. 2025. Cocolex: Confidence-guided copy-based decoding for grounded legal text generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19002–19018.

Peter Schafhalter, Shun Liao, Yanqi Zhou, Chih-Kuan Yeh, Arun Kandoor, and James Laudon. 2024. Scalable multi-domain adaptation of language models using modular experts. *arXiv preprint arXiv:2410.10181*.

Priyanka Sen, Alham Fikri Aji, and Amir Saffari. 2022. Mintaka: A complex, natural, and multilingual dataset for end-to-end question answering. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1604–1619, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 783–791.

Weiyan Shi, Ryan Shea, Si Chen, Chiyuan Zhang, Ruoxi Jia, and Zhou Yu. 2022. Just fine-tune twice: Selective differential privacy for large language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6327–6340.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, and 1 others. 2025. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799.

Yosephine Susanto, Adithya Venkatadri Hulagadri, Jann Railey Montalan, Jian Gang Ngui, Xian Bin Yong, Weiqi Leong, Hamsawardhini Rengarajan, Peerat Limkonchotiwat, Yifan Mai, and William Chandra Tjhi. 2025. Sea-helm: Southeast asian holistic evaluation of language models. *Preprint*, arXiv:2502.14301.

Marzieh Tahaei, Aref Jafari, Ahmad Rashid, David Alfonso-Hermelo, Khalil Bibi, Yimeng Wu, Ali Ghodsi, Boxing Chen, and Mehdi Rezagholizadeh. 2024. Efficient citer: Tuning large language models for enhanced answer quality and verification. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4443–4450.

Thomas Yu Chow Tam, Sonish Sivarajkumar, Sumit Kapoor, Alisa V Stolyar, Katelyn Polanska, Karleigh R McCarthy, Hunter Osterhoudt, Xizhi Wu, Shyam Visweswaran, Sunyang Fu, Piyush Mathur, Giovanni E. Cacciamani, Cong Sun, Yifan Peng, and Yanshan Wang. 2024. A framework for human evaluation of large language models in healthcare derived from literature review. *Preprint*, arXiv:2405.02559.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2024. Steering language models with activation engineering. *Preprint*, arXiv:2308.10248.

Peter Unger. 1967. Experience and factual knowledge. *The Journal of Philosophy*, 64(5):152–173.

Peter Unger. 1968. An analysis of factual knowledge. *The Journal of Philosophy*, pages 157–170.

Thomas Vakili, Aron Henriksson, and Hercules Dalianis. 2024. End-to-end pseudonymization of fine-tuned clinical bert models: privacy preservation with maintained data utility. *BMC Medical Informatics and Decision Making*, 24.

Chris Van der Lee, Albert Gatt, Emiel Van Miltenburg, and Emiel Krahmer. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, 67:101151.

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. 2023. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *Preprint*, arXiv:2310.07521.

Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. 2025a. Adacad: Adaptively decoding to balance conflicts between contextual and parametric knowledge. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11636–11652.

Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. 2025b. Retrieval-augmented generation with conflicting evidence. In *Second Conference on Language Modeling*.

Jiaan Wang, Yunlong Liang, Zengkui Sun, Yuxuan Cao, Jiarong Xu, and Fandong Meng. 2024a. Cross-lingual knowledge editing in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11676–11686.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuan-Jing Huang, Jianshu Ji, Guihong Cao, Daxin Jiang, and Ming Zhou. 2021. K-adapter: Infusing knowledge into pre-trained models with adapters. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1405–1418.

Teng Wang, Lindong Zhai, Tengfei Yang, Zhucheng Luo, and Shuanggen Liu. 2024b. Selective privacy-preserving framework for large language models fine-tuning. *Information Sciences*, 678:121000.

Weixuan Wang, Barry Haddow, and Alexandra Birch. 2024c. Retrieval-augmented multilingual knowledge editing. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 335–354.

Xiaozhi Wang, Kaiyue Wen, Zhengyan Zhang, Lei Hou, Zhiyuan Liu, and Juanzi Li. 2022. Finding skill neurons in pre-trained transformer-based language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11132–11152.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, and 1 others. 2024d. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290.

Yuxia Wang, Minghan Wang, Muhammad Arslan Manzoor, Fei Liu, Georgi Nenkov Georgiev, Rocktim Jyoti Das, and Preslav Nakov. 2024e. Factuality of large language models: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19519–19529.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2020. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Preprint*, arXiv:1805.12471.

Zhihua Wen, Zhiliang Tian, Zexin Jian, Zhen Huang, Pei Ke, Yifu Gao, Minlie Huang, and Dongsheng Li. 2024. Perception of knowledge boundary for large language models through semi-open-ended question answering. *Advances in Neural Information Processing Systems*, 37:88906–88931.

Jacek Wiland, Max Ploner, and Alan Akbik. 2024. Bear: A unified framework for evaluating relational knowledge in causal and masked language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2393–2411.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2023. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*.

Guangzhi Xiong, Qiao Jin, Xiao Wang, Yin Fang, Haolin Liu, Yifan Yang, Fangyuan Chen, Zhixing Song, Dengyu Wang, Minjia Zhang, Zhiyong Lu, and Aidong Zhang. 2026. Supervising the search process produces reliable and generalizable information-seeking agents. *Preprint*, arXiv:2502.13957.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjue Wang, Fan Gao, Jinghui Lu, Yuang Jiang, Huitao Li, Xin Li, Kunyu Yu, Ruihai Dong, Shangding Gu, Yuekang Li, Xiaofei Xie, and 13 others. 2025. MMLU-ProX: A multilingual benchmark for advanced large language model evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1513–1532, Suzhou, China. Association for Computational Linguistics.

Kevin Yang and Dan Klein. 2021. FUDGE: Controlled text generation with future discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3511–3535, Online. Association for Computational Linguistics.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.

Yuanshun Yao and Xiaojun Xu. 2024. Large language model unlearning. *Advances in Neural Information Processing Systems*, 37:105425–105475.

Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2023. Editing large language models: Problems, methods, and opportunities. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10222–10240, Singapore. 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.

Paul Youssef, Osman Koraş, Meijie Li, Jörg Schlötterer, and Christin Seifert. 2023. Give me the facts! a survey on factual knowledge probing in pre-trained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15588–15605.

Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, Sergey Yekhanin, and Huishuai Zhang. 2022. Differentially private fine-tuning of language models. In *International Conference on Learning Representations*.

Qinan Yu, Jack Merullo, and Ellie Pavlick. 2023. Characterizing mechanisms for factual recall in language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Zeping Yu and Sophia Ananiadou. 2024. Neuron-level knowledge attribution in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3267–3280.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in neural information processing systems*, 34:27263–27277.

Fred Zhang and Neel Nanda. 2024. Towards best practices of activation patching in language models: Metrics and methods. In *The Twelfth International Conference on Learning Representations*.

Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, Siyuan Cheng, Ziwen Xu, Xin Xu, Jia-Chen Gu, Yong Jiang, Pengjun Xie, Fei Huang, Lei Liang, Zhiqiang Zhang, and 3 others. 2024a. A comprehensive study of knowledge editing for large language models. *Preprint*, arXiv:2401.01286.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. 2024b. Negative preference optimization: From catastrophic collapse to effective unlearning. *arXiv preprint arXiv:2404.05868*.

Tianjun Zhang, Shishir G. Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E. Gonzalez. 2024c. Raft: Adapting language model to domain specific rag. *Preprint*, arXiv:2403.10131.

Xue Zhang, Yunlong Liang, Fandong Meng, Songming Zhang, Yufeng Chen, Jinan Xu, and Jie Zhou. 2025. Multilingual knowledge editing with language-agnostic factual neurons. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5775–5788, Abu Dhabi, UAE. Association for Computational Linguistics.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. Paws: Paraphrase adversaries from word scrambling. In *North American Chapter of the Association for Computational Linguistics*.

Zhen Zhang, Xinyu Wang, Yong Jiang, Zhuo Chen, Feiteng Mu, Mengting Hu, Pengjun Xie, and Fei Huang. 2024d. Exploring knowledge boundaries in large language models for retrieval judgment. *arXiv preprint arXiv:2411.06207*.

Linxi Zhao, Sofian Zalouk, Christian K. Belardi, Justin Lovelace, Jin Peng Zhou, Ryan Thomas Noonan, Dongyoung Go, Kilian Q. Weinberger, Yoav Artzi, and Jennifer J. Sun. 2025a. Pre-training limited memory language models with internal and external knowledge. *Preprint*, arXiv:2505.15962.

Wayne Xin Zhao, Naoki Yoshinaga, and Daisuke Oba. 2024. What matters in memorizing and recalling facts? multifaceted benchmarks for knowledge probing in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13186–13214.

Yu Zhao, Alessio Devoto, Giwon Hong, Xiaotang Du, Aryo Pradipta Gema, Hongru Wang, Xuanli He, Kam-Fai Wong, and Pasquale Minervini. 2025b. Steering knowledge selection behaviours in llms via sae-based representation engineering. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5117–5136.

Zexuan Zhong, Zhengxuan Wu, Christopher D Manning, Christopher Potts, and Danqi Chen. 2023. Mquake: Assessing knowledge editing in language models via multi-hop questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15686–15702.

Sizhe Zhou, Sha Li, Yu Meng, Yizhu Jiao, Heng Ji, and Jiawei Han. 2024. Establishing knowledge preference in language models. *Preprint*, arXiv:2407.13048.

A Literature Search and Scope Definition

We use the following keywords to collect related papers: linguistic and knowledge separation, knowledge location, knowledge manipulation, knowledge editing, knowledge neurons, unlearning, adapters for internalized separation, contextual knowledge overriding, context-aware decoding, knowledge selection steering, privacy protection pre-training, and knowledge-limited memory pre-training for externalized separation. Rather than limiting our scope to works that explicitly frame themselves as separation, we also include methods that can be interpreted through this lens. This broader scope is intended to inspire future work by offering a more comprehensive set of ideas for separating factual knowledge from linguistic competence.

We exclude *skill neurons* (Wang et al., 2022) from internalized separation because they are designed to locate neurons associated with specific

task labels, such as the positive label in sentiment analysis and the entailment label in natural language inference, rather than neurons associated with factual knowledge. In other words, skill neurons target task competence, whereas our focus is on localizing factual knowledge; thus, the two goals do not align.

B Mechanism-based Steering

Mechanism-based steering methods first distinguish units (e.g., features, attention heads) associated with parametric knowledge from those associated with contextual knowledge. After the distinction, they apply different interventions to them to steer the model toward using contextual rather than parametric knowledge.

TaskVec (Hendel et al., 2023), ActAdd (Turner et al., 2024), and SEA (Qiu et al., 2024) first calculate the average residual stream $\overline{h_C}$ over samples that generate answers based on context rather than parametric knowledge, and then replace the residual stream h with $\overline{h_C}$ to steer the model to follow the context. Yu et al. (2023) first constructs a controlled knowledge-selection benchmark by treating original country–capital pairs (e.g., France–Paris) as parametric knowledge and counterfactual remixed pairs (e.g., France–London) as contextual knowledge. To identify context-steering attention heads, they compare each head's embedding to the unembedding matrix of London and Paris, respectively. Heads whose projections are more aligned with London than Paris are considered to preferentially promote contextual knowledge. Finally, they amplify the value of these contextual-steering heads, enabling the model to leverage contextual knowledge. Jin et al. (2024) replaces the target head's activation with the corrupted activation, and then quantifies how much this corruption reduces the model's preference for the context-supported answer over the memory-based answer. Heads with lower scores are considered not to support the use of contextual knowledge and are therefore suppressed by setting their activations to zero. Zhao et al. (2025b) uses mutual information between feature activations and knowledge-selection labels to identify knowledge-relevant features, and steer model behavior by suppressing features aligned with parametric knowledge while amplifying those aligned with contextual knowledge.

Knowledge selection steering provides a more mechanistic form of steering, but its effectiveness depends heavily on the robustness and causal validity of the identified knowledge-selection components.

C Separation Evaluation

C.1 Factual Knowledge Retention

To assess factual knowledge retention after suppression, prior evaluation protocols can be broadly grouped into two categories: (1) factoid QA benchmarks and (2) knowledge probing datasets. A substantial performance drop on these datasets indicates low factual retention, suggesting successful separation between factual knowledge and linguistic form.

C.1.1 Factoid QA Benchmarks

Natural Question (Kwiatkowski et al., 2019) contains 307,373 natural questions collected from the Google search engine. We suggest using the short-answer subset of NQ, which is more fact-oriented. TriviaQA (Joshi et al., 2017) contains 650K question–answer pairs from trivia websites, with 92.85% of answers corresponding to Wikipedia titles. PopQA (Mallen et al., 2023) uses manual templates to automatically convert 14,000+ Wikidata triples into factoid question-answer pairs. Multi-subject QA benchmarks such as ARC (Clark et al., 2018), GPQA (Rein et al., 2024), MMLU (Hendrycks et al., 2021), MMLU-Pro (Wang et al., 2024d), Humanity's Last Exam (HLE) (Phan et al., 2026), and SuperGPQA (Du et al., 2026) can be employed to assess the breadth of the parametric factual knowledge. Low F1 on factoid QA benchmarks (closed-book setting) demonstrates successful forgetting of factual knowledge after suppression, indicating that factual knowledge has been effectively separated from the model.

C.1.2 Knowledge Probing

LAMA (Petroni et al., 2019) is one of the earliest and most influential benchmarks. It generates fill-in-the-blank cloze questions to examine parametric factual knowledge. Concretely, given "Dante was born in [MASK]", LAMA measures whether the model can correctly rank the answer "Florence" among its top-k predictions (p@K). To prevent the probed model from leveraging statistical shortcuts, LAMA-UHN (Poerner et al., 2020) filters out easily guessable triples; WikiUni (Cao et al., 2021) enforces a uniform distribution over ground-truth objects; and Negated LAMA (Hosseini et al., 2021)

evaluates whether models can reject correct facts when they are presented in negated form. To mitigate the limitations of synthetic queries in LAMA, FACTS-Parametric (Cheng et al., 2025) constructs a dataset of 2,104 cloze-style questions derived from real-world Google Search queries. BEAR (Wiland et al., 2024) extends LAMA to support multi-token answers and introduces a larger benchmark comprising 60 relations and 7,731 Wikidata-derived instances. LM-PUB-QUIZ (Ploner et al., 2025) provides a framework to implement BEAR evaluation. Correct answers that are not ranked among the top-k predictions (LAMA-style), or that are assigned lower log-likelihood than incorrect answers (BEAR-style), indicate low retention of factual knowledge.

Neither factoid QA nor knowledge probing is a flawless indicator of factual knowledge retention. Factoid QA inevitably includes questions that require reasoning, making it difficult to determine whether a model's failure stems from missing factual knowledge or from an inability to reason over that knowledge. Knowledge-probing is relatively small and primarily relies on carefully curated subsets of Wikidata, limiting its coverage of factual knowledge and potentially biasing estimates of knowledge retention.

D What is Linguistic Competency (L)?

Linguistic competency refers to the ability to produce and interpret well-formed, fluent language, governed by phonology, orthography, morphology, syntax, semantics, and pragmatics (Morris, 1938; Grice, 1975; Mahowald et al., 2024).

Specifically, phonology describes the sound system of a language. It includes contrasts between sounds, such as /p/ versus /b/ in pat and bat, prosodic patterns, such as stress placement (e.g., REcord and reCORD), and regular pronunciation rules, such as the aspiration of /p/ in pin but not in spin. Orthography describes how a language is written and the conventions used in writing. This includes spelling rules (e.g., in English, "i before e except after c"), punctuation (such as using a question mark for questions), and capitalization (e.g., capitalizing the first letter of a sentence). Morphology explains how words are built from smaller meaningful parts, such as roots, prefixes, and suffixes, and how these parts express grammatical features like tense, number, or case. Syntax concerns how words are arranged into phrases and sentences, and which word orders and structures are considered well-formed in a language. Semantics studies how linguistic expressions convey meaning, including reference, truth conditions, and the composition of sentence meaning from the meanings of its parts. Pragmatics focuses on how context affects meaning, explaining how the same sentence can be used to express different intentions, such as making indirect requests, showing politeness, or conveying implied meanings beyond what is said explicitly.

E What is Factual Knowledge (K)?

From an epistemological standpoint, knowledge is a justified true belief (Lehrer and Paxson, 1969), which should arise from a reliable method that consistently leads to truth, and should not be true merely by accident (Unger, 1967, 1968). For example, reading the correct time from a stopped clock at the exact right moment is epistemic luck, rather than reliable, repeatable knowledge. Factual knowledge, one kind of knowledge, is typically understood as a form of declarative world knowledge whose truth can be checked against authoritative external sources (Youssef et al., 2023).

Within NLP, instead, factual knowledge is typically characterized in terms of entity- or concept-centered attributes and relations with specific values (e.g., Dante was born in Florence; Oslo is the capital of Norway). Formally, it is cast as queryable fact units (e.g., knowledge triples $t = (s, r, o)$ or equivalent question–answer pairs $< q, a >$) and is tested via cloze prompts to see whether the model can recover the correct o. Hu et al. (2024) further note that facts in LLMs can be misgeneralized and become outdated. Consequently, evaluating factual knowledge should go beyond *single-fact retrieval* and systematically cover dimensions such as source, temporal validity, domain, and geography.

F RAG Optimization

RAG is generally constrained by three bottlenecks: how well retrieval is triggered, how well evidence is selected, and how effectively evidence is integrated into generation. (i) When to retrieve is a delicate design choice: over-triggering wastes compute and injects noise, whereas under-triggering increases hallucinations. To make triggering explicit during generation, FLARE (Jiang et al., 2023b) and Active RAG (Jiang et al., 2023b) rely on uncertainty- or refinement-based signals to decide when to retrieve, while ReAct (Yao et al.,

2022) triggers retrieval through an iterative reasoning–acting trajectory that interleaves external actions (e.g., search/lookup) with intermediate reasoning. OPEN-RAG (Islam et al., 2024) further discusses using an external judge (e.g., GPT-4 or FlanT5) to predict retrieval necessity. ExpertRAG (Gumaan, 2025) treats retrieval as an optional expert invoked by a learned gate. (ii) Evidence selection errors are a dominant failure mode in RAG, particularly for long-tail entities, compositional questions, and multi-hop reasoning. To mitigate this issue, recent work has improved the retrieval stage through stronger dense and hybrid retrievers (Karpukhin et al., 2020; Izacard et al., 2021), late-interaction re-rankers (Khattab and Zaharia, 2020), and sparse expansion re-rankers (Formal et al., 2021). Complementary approaches Gao et al. (2023a); Ma et al. (2023) rewrite the query to mitigate the mismatch between natural user questions and retrievable evidence, thereby enhancing the quality of evidence selection. (iii) Evidence integration remains challenging even when retrieval succeeds: top-k passages are frequently redundant, noisy, or mutually inconsistent. This bottleneck is typically addressed via inference-time evidence management: Jiang et al. (2023a, 2024) score and prune low-utility spans to fit the prompt within the context window. Rackauckas (2024); Ma et al. (2023) generate multiple diverse queries and then merge and rerank the pooled candidates, thereby reducing the brittleness of single queries. Wang et al. (2025b) detects conflicting claims across sources and guides the model to either reconcile the evidence or report the conflict with appropriate uncertainty, improving consistency under the conflicting context. Another line of work improves evidence integration by post-training LLMs to better condition on retrieved context. Zhang et al. (2024c); Tahaei et al. (2024) utilize instruction tuning to enable LLMs to ignore distractor documents while citing supporting spans. Menick et al. (2022) rewards evidence-backed answers and penalizes hallucinated generations through preference optimization (e.g., RL from human feedback). Asai et al. (2023); Xiong et al. (2026) focus on process supervision, enforcing the selection of intermediate evidence before final generation.