Disentangling Linguistic Competence and Factual Knowledge in LLMs: A Survey

Anonymous ACL submission

Abstract

Maintaining factual accuracy is becoming increasingly expensive as Large Language Models (LLMs) scale. This has spurred a modular perspective that decouples linguistic competence from factual knowledge in LLMs, which enables targeted fact updates without full retraining. Yet a coherent framework to guide this emerging line of work is still lacking. To fill this gap, we present a comprehensive survey through the lens of Linguistic–Knowledge Separation (LKS), consolidating methods and evaluations into a unified framework. We make four contributions: (1) We clarify the conceptual distinction between linguistic and factual knowledge. (2) We summarize representative benchmarks and metrics for the linguistic and knowledge sides, enabling side-specific evaluation. (3) We provide a comprehensive summary of LKS methodologies and develop a systematic taxonomy that organizes them into coherent categories. (4) Finally, we outline future directions and open challenges toward robust, generalizable LKS.

1 Introduction

With the continued scaling of large language models, maintaining factual accuracy through repeated end-to-end retraining becomes increasingly costly. This motivates a modularity-oriented perspective that separates linguistic competence and factual knowledge in LLMs (Mahowald et al., 2024; Collado-Montañez, 2024; Chen, 2023). In this paradigm, the model's linguistic competence for expression, composition, and reasoning is kept stable and reusable, whereas factual knowledge, including entities, relations, events, and their truth values, is made inspectable and readily revisable. Consequently, fact updates can be implemented through targeted interventions rather than prohibitively expensive retraining, a property that is particularly important in dynamic real-world settings where information evolves rapidly, and updates are required frequently.

This survey aims to provide an in-depth, systematic analysis of existing Linguistic–Knowledge Separation (LKS) approaches, emphasizing recent advancements to capture the rapidly evolving nature of this field. The main contributions of this paper are summarized as follows: (1) We define linguistic and factual knowledge separately, and analyze the mechanisms that lead to their entanglement in LLMs. (Section 2). (2) We compile representative benchmarks and metrics for both the linguistic and knowledge aspects, and propose a compact evaluation toolkit to enable consistent comparison across approaches (Section 3). (3) We review the literature from the perspective of linguistic–factual disentanglement in LLMs, outlining key approaches and trends (Section 4). Specifically, we group prior work into three main categories: (i) Isolating parametric knowledge within the model to make it explicit and auditable (Section 4.1); (ii) Overriding internal knowledge at inference time by injecting external evidence, thereby shifting facts outside the model (Section 4.2); (iii) Attenuating the model's reliance on stored knowledge via masking training, encouraging it to behave more like a pure language engine (Section 4.3). A comprehensive comparison of these three approach families is provided in Table 1 and Figure 1. (4) We discuss open challenges and propose potential solutions, offering actionable insights for future studies (Section 5).

Difference to Existing Surveys Existing surveys partially overlap with our scope in that they cover ingredients that separation methods often rely on, such as factuality and hallucination mitigation (Huang et al., 2025), retrieval-augmented generation (Fan et al., 2024; Gupta et al., 2024), knowledge updating and editing (Wang et al., 2024), and interpretability techniques for knowledge localization (Rai et al., 2024; Ranaldi, 2025). However,

Method Family	Representative Methods	Main Idea	L Intact	K Intact	K Conflict	Training Required	Drawback
Isolate	Dai et al. (2022) Deng et al. (2025b) Fedus et al. (2022)	Make parametric knowledge explicit and modularized inside LLMs	May be affected	✓	✗	Depends	Limited Transfer
Override	Lewis et al. (2020) Li et al. (2024b) Schick et al. (2023)	Override parametric memory with external evidence during generation.	✓	✓	✓	Depends	Override not assured
Attenuate	Boutet et al. (2025) Wang et al. (2025c) Eldan and Li (2023)	Target linguistic-only LLM by weakening parametric knowledge	May be affected	✗	Residual	✓	Implicit Factual Leakage

Table 1: A comprehensive comparison of separation approaches. L-intact: linguistic competence preserved; K-intact: parametric knowledge preserved; K-conflict: conflicts between parametric knowledge and external evidence; Training-required: requires training of the base model or auxiliary modules.

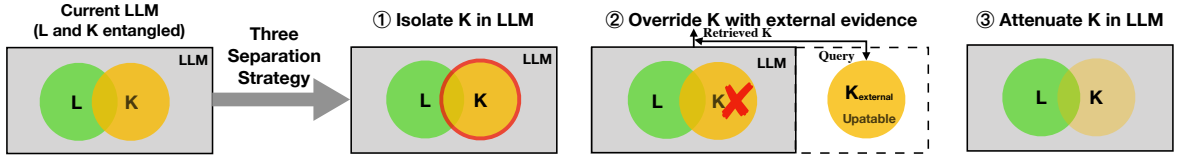

Figure 1: Three Strategies for Linguistic-Knowledge Separation: Isolate, Override, and Attenuate

these surveys typically treat the LLM as a monolithic generator and organize the literature by task outcomes (e.g., factual accuracy) or specific techniques (e.g., RAG, editing, probing). To the best of our knowledge, this is the first survey to systematically examine existing approaches through the lens of separating linguistic competence and factual knowledge in LLMs.

2 Background

What is Linguistic(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). Phonology characterizes the sound system (e.g., /p/ vs. /b/, stress patterns), while orthography describes writing conventions (e.g., spelling, punctuation, capitalization). Morphology and syntax govern how meaningful units form words and how words combine into well-formed sentences. Semantics concerns the meanings and truth conditions expressed by linguistic forms, and pragmatics explains how context shapes intended meaning (e.g., implicatures, politeness, indirect requests). Together, these components span linguistic form (phonology/orthography), grammatical organization (morphology/syntax), and interpretation in context (semantics/pragmatics).

What is Factual Knowledge (K)? Factual knowledge is justified true belief that can be verified or disproven with reliable sources from an epistemological standpoint (Lehrer and Paxson, 1969). 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 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.

Why are L and K entangled in LLM? Linguistic competence and factual knowledge become tightly entangled in LLMs because the model is trained as a single next-token predictor. The same parameters and hidden representations must do double duty: they learn to build coherent sentences, while also capturing statistical regularities about entities, relations, and events seen in the training data. Since the training signal only rewards making the next word likely, the model does not naturally store facts as clean, addressable records. Instead, it often encodes factual content together with the way that content is usually expressed, yielding internal features that mix *what is true* with *how it is said*.

3 Metrics and Benchmarks

3.1 Linguistic-side Evaluation

This section assesses the linguistic competence of LLMs while minimizing reliance on stored factual knowledge.

Grammatical Correctness The following resources evaluate whether models have internalized systematic word-formation rules. Inflection and lemmatization benchmarks (e.g., UniMorph; SIGMORPHON shared tasks; CELEX/Wiktionary-derived resources (Batsuren et al., 2022; Kodner et al., 2022; Baayen et al., 1996; Durrett and DeNero, 2013)) treat morphology as a bidirectional mapping between a lemma paired with a morphosyntactic feature bundle and its surface form, and evaluate models in either the generation direction (lemma+features → form; e.g., go + 3SG.PRES → goes) or the analysis direction (form → lemma/features; e.g., goes → go). Derivational morphology datasets (e.g., MorphyNet (Batsuren et al., 2021)) assess whether models can apply derivational rules to generate or recognize members of a derivational family. Morpheme segmentation datasets (e.g., MorphoChallenge (Kurimo et al., 2010)) probe internal word structure by requiring morpheme boundary and sequence recovery (roots and affixes). Standard metrics for these benchmarks include exact-match accuracy, average edit distance, and over- and understemming rates.

The following resources test the sentence-level grammatical competence. CoLA (Warstadt et al., 2019) and its multilingual counterparts ItaCoLA (Trotta et al., 2021), RuCoLA (Mikhailov et al., 2022), EsCoLa (Bel et al., 2024), QFrCoLA (Beauchemin and Khoury, 2025), NoCoLA (Jentoft and Samuel, 2023), CoLAC (Hu et al., 2023), and JCoLA (Someya et al., 2024) frame grammar as a linguistic acceptability problem, asking whether a sentence is well-formed or ungrammatical. MELA (Zhang et al., 2024b) further unifies CoLA-style resources into a single evaluation setup to support cross-lingual comparison. BLiMP (Warstadt et al., 2020) and its multilingual counterparts ZhoBLiMP (Liu et al., 2024), JBLiMP (Someya and Oseki, 2023), RuBLiMP (Taktasheva et al., 2024), BLiMP-NL (Suijkerbuijk et al., 2025), BLiMP-IT (Barbini et al., 2025), QFrBLiMP (Beauchemin et al., 2025) construct minimal grammatical contrast pairs (ill-formed vs. well-formed), and check whether the model prefers the well-formed sentence. They encompass a range of diverse syntactic and morphosyntactic phenomena, including subject–verb agreement, binding, and negative polarity item licensing. Syntactic parsing benchmarks require models to output explicit parse structures to test sentence-level grammar, including constituency treebanks datasets the Penn Treebank (Marcus et al., 1993), Penn Chinese Treebank (Xue et al., 2005), French Treebank (Abeillé et al., 2003), Arabic Treebank (Maamouri et al., 2004), and TIGER (Brants et al., 2002) that annotate phrase structure, as well as dependency benchmarks the Prague Dependency Treebank (Böhmová et al., 2003) and the multilingual Universal Dependencies treebanks (Nivre et al., 2016) that mark head–dependent relations between words. CoLA-style benchmarks report accuracy and Matthews Correlation Coefficient (MCC). BLiMP-style evaluations are scored by pairwise preference accuracy. Parsing performance is assessed using labeled bracketing F1 and unlabeled/labeled attachment scores.

Fluency and Diversity Open-domain corpora, such as OpenWebText (Gokaslan and Cohen, 2019), C4 (Colossal Clean Crawled Corpus)[1] , The Pile (Gao et al., 2020), RedPajama (Weber et al., 2024), RefinedWeb (Penedo et al., 2023), Dolma (Soldaini et al., 2024), FineWeb (Penedo et al., 2024), and Datacomp-lm (Li et al., 2024a), serve as a natural source of diverse, in-the-wild prompts for open-ended continuation. Sampling generations from these corpora makes it straightforward to surface and quantify degeneration behaviors such as repetition, looping, loss of topical coherence, and incoherent drift, and to study how these failure modes depend on prompt type, length, decoding strategy, or model size.

ParaNMT-50M (Wieting and Gimpel, 2018), ParaBank/ParaBank2 (Hu et al., 2019; Hao et al., 2022), TaPaCo (Scherrer, 2020), and Opusparcus (Sjöblom) provide paraphrase pairs, sentences that express the same meaning with different wording, allowing evaluation of whether a model can vary surface form while preserving semantics. Complementary stress-test contrast sets, such as PAWS and PAWS-X (Zhang et al., 2019; Yang et al., 2019), focus on high-lexical-overlap pairs where minimal structural edits flip the meaning, probing whether models maintain semantic fidelity when surface cues are misleading. In addition, multi-reference data-to-text benchmarks such as E2E (Novikova et al., 2017) and WebNLG (Gardent et al., 2017)

[1] https://www.tensorflow.org/datasets/catalog/c4

Evaluation Dimension		Corpus	Metrics
L	Grammar	UniMorph; CoLA, BLiMP	EM; Pairwise preference accuracy
	Fluency	OpenWebText; The Pile	Repetition/loop rate; SLOR
	Diversity	ParaNMT-50M; WebNLG	Distinct-n; Self-BLEU
K	Faithfulness	FActScore, CiteCheck	Supported-claim rate; Citation precision/recall
	Truthfulness	TruthfulQA, Natural Questions	MC accuracy (MC1/MC2); EM (often with F1)

Table 2: A compact evaluation toolkit for language-side competence and knowledge-side reliability.

provide multiple human-written outputs for the same meaning representation. This enables the evaluation of whether a model can generate different valid phrasings of the same content, rather than being rewarded only for matching a single reference.

Fluency metrics include n-gram repetition rates, tail degeneration (Holtzman et al., 2019), and loop detectors (Welleck et al., 2019). SLOR (Lau et al., 2017) offers a length-normalized estimate of fluency based on log-probability deviations. Diversity is typically evaluated by Distinct-n (Li et al., 2016), Self-BLEU (Zhu et al., 2018), n-gram entropy (Pang et al., 2020), and MAUVE (Pillutla et al., 2021).

3.2 Knowledge-side Evaluation

This section reviews knowledge-side evaluation along two dimensions: faithfulness and truthfulness. Faithfulness measures how well a response matches the provided context, whereas truthfulness concerns an LLM's capability to tell the truth based on its parametric knowledge (see Appendix C).

Factual Faithfulness FactCC (Kryściński et al., 2020) reframes factuality checking as a sentence-level task: it treats each summary sentence as a claim and classifies whether that claim is consistent with the source document. DAE (Goyal and Durrett, 2020) refines this to dependency-arc–level entailment and aggregates arc decisions. SummaC (Laban et al., 2022) runs an off-the-shelf natural language inference (NLI) model on document–summary sentence pairs, then aggregates the resulting entailment scores into a single summary-level consistency metric. FActScore (Min et al., 2023) and CiteCheck (Xu et al., 2025) make the evaluation more explicitly claim-centric by first decomposing an answer into atomic statements and then verifying each statement against evidence (retrieved documents or cited passages), reporting the proportion of supported claims.

Other benchmarks stress-test faithfulness by explicitly introducing fact conflicts in context. Fakepedia [2] constructs counterfactual texts that conflict with a model's parametric knowledge, probing whether the model relies on the given evidence rather than defaulting to memorized beliefs when the two disagree. WikiContradict (Hou et al., 2024) presents mutually contradictory retrieved passages from Wikipedia and examines whether models faithfully acknowledge uncertainty or evidence conflict, rather than arbitrarily committing to one answer. FaithEval (Ming et al., 2025) injects counterfactual, inconsistent, or unanswerable contexts to check whether the model follows the given evidence instead of defaulting to parametric recall.

QA-based evaluators instead probe the output's content through generated questions. QAGS (Wang et al., 2020) and FEQA (Durmus et al., 2020) first derive questions from the system output, then use a QA model to answer these questions from the source, and finally score faithfulness by whether the context-derived answers agree with what the output implies. Q^2 (Honovich et al., 2021) couples QA probing with entailment-style checks. QuestEval (Scialom et al., 2021) and QAFactEval (Fabbri et al., 2022) improve QA-based evaluators with advanced question selection, answering, and aggregation strategies.

Faithfulness metrics generally share the same template: choose an evaluation unit (e.g., a sentence, an extracted claim, an atomic fact, a cited statement, a dependency arc, or a QA probe), apply a consistency checker between the unit and the context (e.g., a self-trained or an established NLI-style entailment model), and aggregate unit-level decisions into an overall score. The reported scores are typically (i) classification metrics such as exact match, and token-F1 when gold consistent and inconsistent labels are available, (ii) NLI-based scores that pool entailment and contradiction probabilities across unit–context pairs, and (iii) support rates computed as the fraction of units deemed entailed. Many works additionally report

[2] https://www.alphaxiv.org/benchmarks/epfl/fakepedia

Pearson or Spearman correlations with human faithfulness/factuality ratings to quantify how well the automatic metric aligns with human judgment.

Factual Truthfulness Factoid QA datasets such as Natural Questions (Kwiatkowski et al., 2019) and TriviaQA (Joshi et al., 2017) test truthfulness by measuring how accurately models answer encyclopedic questions. TruthfulQA (Lin et al., 2022) takes a different angle and uses *trap* questions that often trigger misconceptions, so truthfulness becomes the ability to avoid plausible-sounding falsehoods and refrain from confident hallucinations. In StrategyQA (Geva et al., 2021), models are judged truthful if they reach the correct yes/no decision through coherent multi-step inference rather than single-hop recall. ARC (Clark et al., 2018) mirrors grade-school science exams, where truthfulness is reflected in whether the model's answers match basic scientific facts and the simple reasoning needed to apply them. RealFactBench (Yang et al., 2025) measures LLM truthfulness via real-world claim verification: models output a calibrated true/false/unknown verdict (with rationale), scored against ground-truth labels and evidence. CommonsenseQA (Talmor et al., 2019), HellaSwag (Zellers et al., 2019), PIQA (Bisk et al., 2019), and SocialIQA (Sap et al., 2019) serve as proxy truthfulness tests by evaluating whether model outputs align with broadly shared everyday knowledge.

Knowledge probing benchmarks provide a direct, fine-grained test of truthfulness. LAMA (Petroni et al., 2019) (including its TREx subset derived from T-REx (Elsahar et al., 2018)) uses templated cloze prompts to probe whether models can recover the missing entity in a closed-book setting. LAMA-style variants refine this paradigm along several axes, including filtering easily guessable triples (e.g., LAMA-UHN (Poerner et al., 2020)), balancing answer distributions (e.g., WIKI-UNI (Cao et al., 2021)), extending coverage beyond English (e.g., mLAMA (Kassner et al., 2021) and X-FACTR (Jiang et al., 2020)), and adapting probing to specialized or shifting regimes (e.g., BioLAMA (Sung et al., 2021), MedLAMA (Meng et al., 2022b), and TempLAMA (Dhingra et al., 2022)).

Truthfulness evaluation depends on the format of the underlying task. For open-ended generation (Kwiatkowski et al., 2019), a common practice is to use LLM-based factuality judges (e.g., LLaMA-Judge (Zheng et al., 2023) and Skywork-Reward (Liu et al., 2025)) to score the factual correctness of model outputs automatically. For multiple-choice QA benchmarks (Talmor et al., 2019), performance is typically quantified by multi-class accuracy, i.e., whether the model selects the correct option among a set of distractors. For cloze-style probing tasks (Petroni et al., 2019), evaluation commonly uses exact-match accuracy or top-k accuracy (Hits@k), sometimes complemented by ranking-based metrics such as mean reciprocal rank (MRR) or mean rank to capture how highly the correct fact is prioritized in the model's output distribution.

Summary Table 2 provides a lightweight yet informative checklist to encourage consistent evaluation across studies.

4 Methodology

4.1 Isolating

Isolating separates L and K by making parametric knowledge explicit within the model. This is achieved through one of three strategies: locating knowledge-sensitive units (Section 4.1.1); decomposing hidden representations into interpretable feature subspaces (Section 4.1.2); or decoupling knowledge-related computation into dedicated, identifiable modules such as adapters (Section 4.1.3).

4.1.1 Units Localization

Units localization identifies the internal components (e.g., neurons, attention heads, and MLP/FFN blocks) that causally mediate knowledge.

At the neuron level, knowledge neurons (Dai et al., 2022) compute attribution scores on cloze-style queries to isolate a sparse set of MLP neurons that support a given factual continuation, and validate the localization by amplifying those neurons to strengthen the corresponding fact. This line has been extended to broader settings: AMIG (Chen et al., 2024) adapts integrated gradients to multilingual and architecture-specific conditions for more robust knowledge-neuron discovery, while a scalable attribution framework (Yu and Ananiadou, 2024) identifies influential value neurons and traces the query neurons that activate them to analyze where different knowledge types concentrate across attention and FFN components. At the module level, ROME-style causal tracing (Meng et al., 2022a) contrasts a corrupted run with a clean run and patches activations back piece by piece; if restoring hidden states at a particular mid-layer FFN/MLP block recovers the correct object predic-

tion, that block is implicated as a causal bottleneck for fact recall. More generally, activation patching and causal tracing (Zhang and Nanda, 2023) provide a reusable toolkit for attributing factual behaviors to specific layers, heads, or modules via counterfactual restoration rather than correlation. Several works apply the same intervention logic to attention heads. PH3 (Jin et al., 2024) uses path patching under conflict prompts to identify late-layer heads necessary for producing the correct factual object, and validates them by ablating the heads to alter factual outputs. Head-level attribution (Yu et al., 2023) reaches similar conclusions and demonstrates controllability by scaling the value contribution of individual heads. Going further, edge-level interventions (Geva et al., 2023) trace how factual signals propagate through attention and which heads act as key carriers along the recall pathway. Circuit-level approaches, on the other hand, generalize localization from single units to coordinated mechanisms. Knowledge Circuits (Yao et al., 2024) extract multi-component circuits spanning heads and MLPs that jointly support factual completion, and validate them with targeted ablations or restorations while tracking collateral effects. Related circuit-tracing and attribution-graph method (Olah et al., 2020) similarly outputs behavior-relevant computational subgraphs, offering a structured view of factual recall as cooperative pathways rather than isolated units. Targeted unlearning (Guo et al., 2024; Ravfogel et al., 2020; Belrose et al., 2023; Turner et al., 2023; Rimsky et al., 2024) can also serve as a stress test for localization quality: if the proposed units truly mediate a fact family, erasing them should remove that knowledge efficiently with minimal side effects.

Discussion Unit localization faces two main limitations: (1) Limited transferability: the localized *knowledge units* often change with prompt wording and context, making findings hard to carry across domains and datasets; (2) Narrow coverage: localization is tied to the specific facts probed, and may not generalize beyond those probe instances.

4.1.2 Feature Decomposition

Feature decomposition transforms the hidden representations in LLM into a more interpretable feature base and analyzes which features encode target knowledge or attributes.

Sparse feature is the widely used approach to decompose the hidden representation (Bricken et al., 2023; Cunningham et al., 2023; Shu et al., 2025). They utilize sparse autoencoders (SAEs) and dictionary learning to decompose activations into a sparse set of interpretable features, and then manipulate the features, either by ablation or scaling, to define a subspace for constrained updates. Farrell et al. (2024) identifies SAE features correlated with domain knowledge (e.g., biology) and performs unlearning by inhibiting those features, selectively attenuating the targeted knowledge. Wang et al. (2025d) (SAE subspace projection) uses SAE-derived relevant and irrelevant subspaces to constrain unlearning updates, restricting parameter changes to directions associated with the knowledge to be forgotten. Deng et al. (2025b) group SAE features into higher-level causal semantic modules (e.g., concept vs. relation) via feature coactivation patterns and validate these modules with interventions, showing compositional control over factual content. Meanwhile, Deng et al. (2025a) finds that SAE features are tied to specific languages and shows that scaling these features can steer the output language without requiring retraining. Deng et al. (2025b) (SASFT) utilizes SAE features to regularize fine-tuning, thereby suppressing unwanted language features and reducing code-switching while preserving multilingual ability.

Besides SAEs, other approaches also localize knowledge by decomposing activations into a small set of interpretable components. Valentin et al. (2025) uses sparse coding (learning a dictionary of atoms D and sparse coefficients z with $h \approx Dz$), so knowledge can be traced to the atoms that reliably receive high coefficients for particular entities or relations. Leask et al. (2025) performs encoder-free, inference-time decomposition, identifying knowledge features as the few components repeatedly selected to explain factual patterns, and validating them by removing or swapping those components during forward passes. Collins et al. (2018); Fel et al. (2023) factorize activation tensors into additive concept bases, where knowledge appears as parts-like factors aligned with relational attributes; localization then reduces to selecting the most attribute-predictive factors and intervening on their activations. Tamkin et al. (2023) constraints hidden states to be a sum of a few learned codes, so knowledge-related features become identifiable codes that can be switched on and off during generation.

Discussion Feature decomposition also suffers from limited transferability: features learned in one setting often do not generalize across tasks.

4.1.3 Module Decoupling

Module decoupling trains additional modules to encapsulate knowledge and linguistic abilities, respectively. K-ADAPTER (Wang et al., 2021) provides a canonical example of capability-specific modularization by attaching multiple adapters in parallel in RoBERTa: a factual adapter trained with Wikidata supervision to acquire encyclopedic knowledge, and a linguistic adapter trained on dependency parsing signals to capture syntactic structure. A similar factorization is adopted in MAD-X (Pfeiffer et al., 2020), which separates adapters into language-specific and task-specific modules, with the former handling cross-lingual adaptation and the latter encoding task semantics. AdapterFusion (Pfeiffer et al., 2021) further decomposes linguistic competence by training multiple expert adapters specializing in distinct structural cues such as POS tags, morphology, or dependency relations, and dynamically combining them via learned attention weights over frozen adapters. SWITCH TRANSFORMERS (Fedus et al., 2022) and GLAM (Du et al., 2022) use a learned router to dynamically select sparse experts, allowing different experts to implicitly specialize across heterogeneous knowledge domains. Building on this MoE idea, domain-mixture architectures such as DEMix (Gururangan et al., 2022), MoDE (Schafhalter et al., 2024), and MoDEM (Simonds et al., 2024) make the specialization more explicit by partitioning experts by domain (e.g., legal, medical, financial) and conditioning activation on domain signals.

Discussion Module decoupling scales naturally, as new experts can be added to encode additional knowledge or linguistic capabilities. However, this scalability relies on the assumption that linguistic competence and factual knowledge can be cleanly isolated at the level of model components. In practice, the two are tightly intertwined within shared computational trajectories, often at token-level granularity, so enforcing fixed module boundaries can limit expressivity and lead to unstable routing as contextual demands change.

4.2 Overriding

Due to the instability of isolating, overriding takes an alternative route: they prioritize updatable external evidence over parametric memory during generation, effectively relocating factual grounding outside the model (Samuel et al., 2024).

Training-free retrieval-augmented generation (RAG) (Lewis et al., 2020; Guu et al., 2020; Karpukhin et al., 2020; Fan et al., 2024) is the most common instantiation, which first turns the user request into a retrieval query, then retrieves the top-k passages from a pre-indexed corpus using BM25 or dense retrieval, and finally appends these passages to the prompt as supporting context for an LLM to generate the answer. Beyond text retrieval, graph-based RAG (Peng et al., 2024) equips LLMs with reliable knowledge bases. Specifically, SubgraphRAG (Li et al., 2024b) retrieves a compact, query-specific KG subgraph by scoring triples and feeds it to the LLM as structured evidence. KG-RAG (Sanmartin, 2024) uses LLM-guided graph exploration (chain of explorations) to traverse nodes and relations, gathering KG evidence for grounded answering. Tool-augmented RAGs (Nakano et al., 2021; Schick et al., 2023; Qu et al., 2025) further delegate factual access to external tools, such as web engines, QA retrievers, query APIs (e.g., entity and relation lookup), and functional calls (e.g., calculators and code execution). The performance of these methods largely depends on the retrieval triggering strategy (always-on vs. on-demand), the relevance of selected evidence, and the effectiveness of evidence integration. Appendix B provides a detailed overview of optimizations.

Ensuring that generations remain grounded in retrieved, up-to-date evidence rather than defaulting to the LLM's parametric memory, which may be stale, is key to the success of these methods in separating K and L. Fine-grained citation supervision (Gao et al., 2023c) promotes evidence reliance by redefining the optimization target: the model is rewarded not only for producing a plausible response, but also for producing claims that can be explicitly attributed to specific evidence spans. Memory-based answers are difficult to justify with correct sentence- or span-level citations and therefore receive lower likelihood during training. Asai et al. (2024) interleaves retrieval with self-reflection: the model explicitly critiques whether intermediate statements are supported by the retrieved passages, and if not, modifies the generation process accordingly. Building on a similar idea, *research-and-revise* pipeline (Gao et al., 2023a) first drafts an initial answer and then retrieves targeted evidence to edit away unsupported content, yielding a final response that better matches the evidence. Another line of work enforces evidence dominance at inference time by modifying decoding. Context-aware

decoding (CAD) and its adaptive variants (Shi et al., 2024; Wang et al., 2025a; Khandelwal et al., 2025) use a contrastive objective between generation with vs. without retrieved context to suppress parametric priors under conflict. Other decoding-time control can be imposed by (i) controller-guided decoding that uses auxiliary signals to steer continuations toward context-consistent outputs (Yang and Klein, 2021; Komorowski et al., 2025), (ii) copy-biased decoding that promotes salient entities and details from retrieved passages (Santosh et al., 2025), and (iii) lexically constrained decoding that enforces the inclusion of key evidence-derived items (e.g., entities or numbers) in the final answer (Hokamp and Liu, 2017; Lu et al., 2021).

Discussion Parametric knowledge exhibits stubbornness. Despite being trained to retrieve before generating, evidence overriding cannot be fully guaranteed: LLMs may still fall back on parametric priors, especially when retrieved evidence is missing, noisy, or conflicting.

4.3 Attenuating

Due to the fall-back risk of override methods, attenuating methods weaken parametric knowledge during training to reduce the model's reliance on parametric memory.

The most direct strategy is corpus-level de-factualization through identifier obfuscation: Boutet et al. (2025) utilizes a mix of rule-based detectors and NER tools to identify direct- and quasi-identifiers, 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., 2023) enforces entity-level consistency by combining NER with coreference resolution so that all mentions of the same entity map to a stable but semantically meaningless pseudonym, preserving discourse coherence while stripping retrievable factual identity; Cabrera-Diego and Gheewala (2024) provides standardized setups for evaluating these transformations. Beyond data obfuscation, a more model-centric route uses differentially private (DP) pretraining to limit memorization and leakage of training examples (Hoory et al., 2021; Li et al., 2021; Wang et al., 2025c), which can also suppress retention of fact-bearing spans in parameters, albeit without targeting factual knowledge specifically. Finally, synthetic-only pretraining (e.g., Eldan and Li (2023)) reduces exposure to real-world facts by training on fully generated corpora with tightly controlled content (limited vocabulary and minimal named-entity grounding), so the model's parameters primarily capture linguistic regularities rather than open-world factual knowledge.

Discussion Implicit factual leakage is the issue. Even when explicit entity mentions are masked, models can still internalize factual regularities from contextual cues, such as numerical patterns and high-level relational schemas. Moreover, training on de-factualized corpora can bias models toward idealized data distributions, reducing robustness under real-world linguistic noise variation.

Summary Isolation is the most direct separation strategy. Overriding is more controllable and, therefore, typically easier to implement in practice. Attenuation carries a higher risk but can deliver a more complete separation.

5 Open Challenges and Future Directions

Data bottleneck is one of the biggest obstacles to separating L and K in LLM. Existing resources rarely satisfy scale and quality simultaneously: web text is abundant but tightly entangles linguistic form with real-world facts, whereas resources explicitly targeting linguistic competence (e.g., BLiMP-style minimal pairs) or factual knowledge (e.g., Wikipedia) are often limited in coverage and uneven across domains and long-tail entities. A promising direction is to use LLM self-supervision to build large-scale separation corpora. For example, converting unstructured text into structured triples makes the factual content explicit and allows the knowledge-side component to be trained directly on these structured targets, rather than being implicitly learned through next-token prediction. Another approach is to use attention-guided perturbations to counterfactually shuffle factual slots in text while preserving linguistic form, and to train with contrastive objectives that prevent the model from recovering facts from contextual regularities.

6 Conclusion

We survey linguistic-knowledge separation methods in LLMs, while synthesizing relevant evaluation benchmarks and metrics. By organizing existing approaches into a unified framework, we provide conceptual clarity on how knowledge and language can be disentangled within or beyond the model.

7 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.

We also note that separation can be evaluated beyond our primary linguistic-side and knowledge-side benchmarks; for example, one may assess separation via downstream criteria such as knowledge-editing effectiveness (how successfully a targeted update is applied while minimizing collateral changes) (Meng et al., 2022a), time-sensitive updating (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). Due to space constraints, we do not systematically review these complementary evaluation axes.

8 Acknowledgements

We used the AI assistant to help with language polishing and phrasing.

References

Anne Abeillé, Lionel Clément, and François Toussenel. 2003. Building a treebank for french. In *Treebanks: Building and using parsed corpora*, pages 165–187. Springer.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection.

R Harald Baayen, Richard Piepenbrock, and Leon Gulikers. 1996. The celex lexical database (cd-rom).

Matilde Barbini, Maria Letizia Piccini Bianchessi, Veronica Bressan, Achille Fusco, Sofia Neri, Sarah Rossi, Tommaso Sgrizzi, and Cristiano Chesi. 2025. Blimp-it: Harnessing automatic minimal pair generation for italian language model evaluation. In *CEUR WORKSHOP PROCEEDINGS*. CEUR.

Khuyagbaatar Batsuren, Gábor Bella, Aryaman Arora, Viktor Martinovic, Kyle Gorman, Zdeněk Žabokrtský, Amarsanaa Ganbold, Šárka Dohnalová, Magda Ševčíková, Kateřina Pelegrinová, Fausto Giunchiglia, Ryan Cotterell, and Ekaterina Vylomova. 2022. The SIGMORPHON 2022 shared task on morpheme segmentation. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 103–116, Seattle, Washington. Association for Computational Linguistics.

Khuyagbaatar Batsuren, Gábor Bella, and Fausto Giunchiglia. 2021. Morphynet: a large multilingual database of derivational and inflectional morphology. In *Proceedings of the 18th sigmorphon workshop on computational research in phonetics, phonology, and morphology*, pages 39–48.

David Beauchemin and Richard Khoury. 2025. Qfrcola: a quebec-french corpus of linguistic acceptability judgments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 119–130.

David Beauchemin, Pier-Luc Veilleux, Richard Khoury, and Johanna-Pascale Roy. 2025. Qfrblimp: a quebec-french benchmark of linguistic minimal pairs. *arXiv preprint arXiv:2509.25664*.

Nuria Bel, Marta Punsola, and Valle Ruíz-Fernández. 2024. Escola: Spanish corpus of linguistic acceptability. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6268–6277.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. 2023. Leace: Perfect linear concept erasure in closed form. *Advances in Neural Information Processing Systems*, 36:66044–66063.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2019. PIQA: reasoning about physical commonsense in natural language. *CoRR*, abs/1911.11641.

Alena Böhmová, Jan Hajič, Eva Hajičová, and Barbora Hladká. 2003. The prague dependency treebank: A three-level annotation scenario. In *Treebanks: building and using parsed corpora*, pages 103–127. Springer.

Antoine Boutet, Lucas Magnana, Juliette Sénéchal, and Hélain Zimmermann. 2025. Towards the anonymization of the language modeling. *arXiv preprint arXiv:2501.02407*.

Sabine Brants, Silvia Hansen, and 1 others. 2002. The tiger treebank. In *In Proceedings of the Workshop on Treebanks and Linguistic Theories*.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, and 1 others. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2.

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. *arXiv preprint arXiv:2106.09231*.

Huajun Chen. 2023. Large knowledge model: Perspectives and challenges. *arXiv preprint arXiv:2312.02706*.

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 AAAI Conference on Artificial Intelligence*, volume 38, pages 17817–17825.

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*.

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

Edo Collins, Radhakrishna Achanta, and Sabine Susstrunk. 2018. Deep feature factorization for concept discovery. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 336–352.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*.

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.

Boyi Deng, Yu Wan, Baosong Yang, Yidan Zhang, and Fuli Feng. 2025a. Unveiling language-specific features in large language models via sparse autoencoders. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4563–4608.

Ruixuan Deng, Xiaoyang Hu, Miles Gilberti, Shane Storks, Aman Taxali, Mike Angstadt, Chandra Sripada, and Joyce Chai. 2025b. Sparse feature coactivation reveals composable semantic modules in large language models. *arXiv preprint arXiv:2506.18141*.

Bhuwan Dhingra, Jeremy R Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W Cohen. 2022. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, 10:257–273.

Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, and 1 others. 2022. Glam: Efficient scaling of language models with mixture-of-experts. In *International conference on machine learning*, pages 5547–5569. PMLR.

Esin Durmus, He He, and Mona Diab. 2020. Feqa: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of ACL*.

Greg Durrett and John DeNero. 2013. Supervised learning of complete morphological paradigms. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1185–1195.

Ronen Eldan and Yuanzhi Li. 2023. Tinystories: How small can language models be and still speak coherent english? *arXiv preprint arXiv:2305.07759*.

Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest, and Elena Simperl. 2018. T-rex: A large scale alignment of natural language with knowledge base triples. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Alexander Richard Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. Qafacteval: Improved qa-based factual consistency evaluation for summarization. In *Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 2587–2601.

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.

Eoin Farrell, Yeu-Tong Lau, and Arthur Conmy. 2024. Applying sparse autoencoders to unlearn knowledge in language models. *arXiv preprint arXiv:2410.19278*.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39.

Thomas Fel, Agustin Picard, Louis Bethune, Thibaut Boissin, David Vigouroux, Julien Colin, Rémi Cadène, and Thomas Serre. 2023. Craft: Concept recursive activation factorization for explainability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2711–2721.

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.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, and 1 others. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and 1 others. 2023a. Rarr: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023b. 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. 2023c. Enabling large language models to generate text with citations. *arXiv preprint arXiv:2305.14627*.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. The webnlg challenge: Generating text from rdf data. In *10th International Conference on Natural Language Generation*, pages 124–133. ACL Anthology.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. *arXiv preprint arXiv:2304.14767*.

Mor Geva and 1 others. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. In *ACL*.

Aaron Gokaslan and Vanya Cohen. 2019. Openwebtext corpus. http://Skylion007.github.io/OpenWebTextCorpus.

Tanya Goyal and Greg Durrett. 2020. Evaluating factuality in generation with dependency-level entailment. *arXiv preprint arXiv:2010.05478*.

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

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

Phillip Guo, Aaquib Syed, Abhay Sheshadri, Aidan Ewart, and Gintare Karolina Dziugaite. 2024. Mechanistic unlearning: Robust knowledge unlearning and editing via mechanistic localization. *arXiv preprint arXiv:2410.12949*.

Shailja Gupta, Rajesh Ranjan, and Surya Narayan Singh. 2024. A comprehensive survey of retrieval-augmented generation (rag): Evolution, current landscape and future directions. *Preprint*, arXiv:2410.12837.

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.

Wenjie Hao, Hongfei Xu, Deyi Xiong, Hongying Zan, and Lingling Mu. 2022. Parazh-22m: A large-scale chinese parabank via machine translation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3885–3897.

Chris Hokamp and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. *arXiv preprint arXiv:1704.07138*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. *arXiv preprint arXiv:2104.08202*.

Shlomo Hoory, Amir Feder, Avichai Tendler, Sofia Erell, Alon Peled-Cohen, Itay Laish, Hootan Nakhost, Uri Stemmer, Ayelet Benjamini, Avinatan Hassidim, and 1 others. 2021. Learning and evaluating a differentially private pre-trained language model. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1178–1189.

Yufang Hou, Alessandra Pascale, Javier Carnerero-Cano, Tigran Tchrakian, Radu Marinescu, Elizabeth Daly, Inkit Padhi, and Prasanna Sattigeri. 2024. Wikicontradict: A benchmark for evaluating llms on real-world knowledge conflicts from wikipedia. *Advances in Neural Information Processing Systems*, 37:109701–109747.

Hai Hu, Ziyin Zhang, Weifang Huang, Jackie Yan-Ki Lai, Aini Li, Yina Patterson, Jiahui Huang, Peng Zhang, Chien-Jer Charles Lin, and Rui Wang. 2023. Revisiting acceptability judgements. *arXiv preprint arXiv:2305.14091*.

J Edward Hu, Rachel Rudinger, Matt Post, and Benjamin Van Durme. 2019. Parabank: Monolingual bitext generation and sentential paraphrasing via lexically-constrained neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6521–6528.

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. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Shayekh Bin Islam, Md Asib Rahman, KSM Hossain, Enamul Hoque, Shafiq Joty, and Md Rizwan Parvez. 2024. Open-rag: Enhanced retrieval-augmented reasoning with open-source large language models. *arXiv preprint arXiv:2410.01782*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.

Matias Jentoft and David Samuel. 2023. Nocola: The norwegian corpus of linguistic acceptability. *arXiv preprint arXiv:2306.07790*.

Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2023a. Llmlingua: Compressing prompts for accelerated inference of large language models. *arXiv preprint arXiv:2310.05736*.

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. 2020. X-factr: Multilingual factual knowledge retrieval from pretrained language models. *arXiv preprint arXiv:2010.06189*.

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. *arXiv preprint arXiv:2102.00894*.

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.

Jordan Kodner, Salam Khalifa, Khuyagbaatar Batsuren, Hossep Dolatian, Ryan Cotterell, Faruk Akkus, Antonios Anastasopoulos, Taras Andrushko, Aryaman Arora, Nona Atanalov, Gábor Bella, Elena Budianskaya, Yustinus Ghanggo Ate, Omer Goldman, David Guriel, Simon Guriel, Silvia Guriel-Agiashvili, Witold Kieraś, Andrew Krizhanovsky, and 11 others. 2022. SIGMORPHON–UniMorph 2022 shared task 0: Generalization and typologically diverse morphological inflection. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 176–203, Seattle, Washington. Association for Computational Linguistics.

Piotr Komorowski, Elena Golimblevskaia, Reduan Achtibat, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. 2025. Attribution-guided decoding. *arXiv preprint arXiv:2509.26307*.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 9332–9346.

Mikko Kurimo, Sami Virpioja, Ville Turunen, and Krista Lagus. 2010. Morpho challenge 2005-2010: Evaluations and results. In *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*, pages 87–95.

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.

Philippe Laban, Tobias Schnabel, Paul N Bennett, and Marti A Hearst. 2022. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive science*, 41(5):1202–1241.

Patrick Leask, Neel Nanda, and Noura Al Moubayed. 2025. Inference-time decomposition of activations (itda): A scalable approach to interpreting large language models. *arXiv preprint arXiv:2505.17769*.

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

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

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Yitzhak Gadre, Hritik Bansal, Etash Guha, Sedrick Scott Keh, Kushal Arora, and 1 others. 2024a. Datacomp-lm: In search of the next generation of training sets for language models. *Advances in Neural Information Processing Systems*, 37:14200–14282.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 110–119.

Mufei Li, Siqi Miao, and Pan Li. 2024b. Simple is effective: The roles of graphs and large language models in knowledge-graph-based retrieval-augmented generation. *arXiv preprint arXiv:2410.20724*.

Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. 2021. Large language models can be strong differentially private learners. *arXiv preprint arXiv:2110.05679*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 3214–3252.

Chris Yuhao Liu, Liang Zeng, Yuzhen Xiao, Jujie He, Jiacai Liu, Chaojie Wang, Rui Yan, Wei Shen, Fuxiang Zhang, Jiacheng Xu, and 1 others. 2025. Skywork-reward-v2: Scaling preference data curation via human-ai synergy. *arXiv preprint arXiv:2507.01352*.

Yikang Liu, Yeting Shen, Hongao Zhu, Lilong Xu, Zhiheng Qian, Siyuan Song, Kejia Zhang, Jialong Tang, Pei Zhang, Baosong Yang, and 1 others. 2024. Zhoblimp: a systematic assessment of language models with linguistic minimal pairs in chinese. *arXiv preprint arXiv:2411.06096*.

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.

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.

Mohamed Maamouri, Ann Bies, Tim Buckwalter, and Wigdan Mekki. 2004. The penn arabic treebank: Building a large-scale annotated arabic corpus. In *NEMLAR conference on Arabic language resources and tools*, volume 27, pages 466–467. Cairo.

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.

Mitch Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330.

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

Zaiqiao Meng, Fangyu Liu, Ehsan Shareghi, Yixuan Su, Charlotte Collins, and Nigel Collier. 2022b. Rewire-then-probe: A contrastive recipe for probing biomedical knowledge of pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4798–4810.

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*.

Vladislav Mikhailov, Tatiana Shamardina, Max Ryabinin, Alena Pestova, Ivan Smurov, and Ekaterina

Artemova. 2022. Rucola: Russian corpus of linguistic acceptability. *arXiv preprint arXiv:2210.12814*.

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

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.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, and 1 others. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, and 1 others. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2017. The e2e dataset: New challenges for end-to-end generation. *arXiv preprint arXiv:1706.09254*.

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*.

Bo Pang, Erik Nijkamp, Wenjuan Han, Linqi Zhou, Yixian Liu, and Kewei Tu. 2020. Towards holistic and automatic evaluation of open-domain dialogue generation. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 3619–3629.

Guilherme Penedo, Hynek Kydlícek, Loubna Ben Allal, and Thomas Wolf. 2024. Fineweb: decanting the web for the finest text data at scale. *HuggingFace. Accessed: Jul*, 12.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Boci Peng, Yun Zhu, Yongchao Liu, Xiaohe Bo, Haizhou Shi, Chuntao Hong, Yan Zhang, and Siliang Tang. 2024. Graph retrieval-augmented generation: A survey. *ACM Transactions on Information Systems*.

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, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. Mad-x: An adapter-based framework for multi-task cross-lingual transfer. *arXiv preprint arXiv:2005.00052*.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers. *Advances in Neural Information Processing Systems*, 34:4816–4828.

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.

Changle Qu, Sunhao Dai, Xiaochi Wei, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, Jun Xu, and Ji-Rong Wen. 2025. Tool learning with large language models: A survey. *Frontiers of Computer Science*, 19(8):198343.

Zackary Rackauckas. 2024. Rag-fusion: a new take on retrieval-augmented generation. *arXiv preprint arXiv:2402.03367*.

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*.

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

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. *arXiv preprint arXiv:2004.07667*.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522.

David Samuel, Lucas Georges Gabriel Charpentier, and Sondre Wold. 2024. More room for language: Investigating the effect of retrieval on language models. *arXiv preprint arXiv:2404.10939*.

Diego Sanmartin. 2024. Kg-rag: Bridging the gap between knowledge and creativity. *arXiv preprint arXiv:2405.12035*.

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.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*.

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*.

Yves Scherrer. 2020. Tapaco: A corpus of sentential paraphrases for 73 languages. In *Language Resources and Evaluation Conference*, pages 6868–6873. European Language Resources Association (ELRA).

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36:68539–68551.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. Questeval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 6594–6604.

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.

Dong Shu, Xuansheng Wu, Haiyan Zhao, Daking Rai, Ziyu Yao, Ninghao Liu, and Mengnan Du. 2025. A survey on sparse autoencoders: Interpreting the internal mechanisms of large language models. *arXiv preprint arXiv:2503.05613*.

Toby Simonds, Kemal Kurniawan, and Jey Han Lau. 2024. Modem: Mixture of domain expert models. *arXiv preprint arXiv:2410.07490*.

Eetu Sjöblom. Annotation of subtitle paraphrases using a new web tool.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, and 1 others. 2024. Dolma: An open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 15725–15788.

Taiga Someya and Yohei Oseki. 2023. Jblimp: Japanese benchmark of linguistic minimal pairs. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1581–1594.

Taiga Someya, Yushi Sugimoto, and Yohei Oseki. 2024. Jcola: Japanese corpus of linguistic acceptability. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9477–9488.

Michelle Suijkerbuijk, Zoë Prins, Marianne de Heer Kloots, Willem Zuidema, and Stefan L Frank. 2025. Blimp-nl: A corpus of dutch minimal pairs and acceptability judgments for language model evaluation. *Computational Linguistics*, pages 1–35.

Mujeen Sung, Jinhyuk Lee, Sean Yi, Minji Jeon, Sungdong Kim, and Jaewoo Kang. 2021. Can language models be biomedical knowledge bases? *arXiv preprint arXiv:2109.07154*.

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.

Ekaterina Taktasheva, Maxim Bazhukov, Kirill Koncha, Alena Fenogenova, Ekaterina Artemova, and Vladislav Mikhailov. 2024. Rublimp: Russian benchmark of linguistic minimal pairs. *arXiv preprint arXiv:2406.19232*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. 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 4149–4158.

Alex Tamkin, Mohammad Taufeeque, and Noah D Goodman. 2023. Codebook features: Sparse and discrete interpretability for neural networks. *arXiv preprint arXiv:2310.17230*.

Daniela Trotta, Raffaele Guarasci, Elisa Leonardelli, and Sara Tonelli. 2021. Monolingual and cross-lingual acceptability judgments with the italian cola corpus. *arXiv preprint arXiv:2109.12053*.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Activation addition: Steering language models without optimization. *arXiv e-prints*, pages arXiv–2308.

Thomas Vakili, Aron Henriksson, and Hercules Dalianis. 2023. End-to-end pseudonymization of fine-tuned clinical bert models.

Romeo Valentin, Sydney M Katz, Vincent Vanhoucke, and Mykel J Kochenderfer. 2025. Db-ksvd: Scalable alternating optimization for disentangling high-dimensional embedding spaces. *arXiv preprint arXiv:2505.18441*.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. *arXiv preprint arXiv:2004.04228*.

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. *arXiv preprint arXiv:2504.13079*.

Liangyu Wang, Junxiao Wang, Jie Ren, Zihang Xiang, David E Keyes, and Di Wang. 2025c. Flashdp: Private training large language models with efficient dp-sgd. *arXiv preprint arXiv:2507.01154*.

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.

Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li. 2024. Knowledge editing for large language models: A survey. *ACM Computing Surveys*, 57(3):1–37.

Xu Wang, Zihao Li, Benyou Wang, Yan Hu, and Difan Zou. 2025d. Model unlearning via sparse autoencoder subspace guided projections. *arXiv preprint arXiv:2505.24428*.

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. Cola: The corpus of linguistic acceptability (with added annotations).

Maurice Weber, Dan Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, and 1 others. 2024. Redpajama: an open dataset for training large language models. *Advances in neural information processing systems*, 37:116462–116492.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2019. Neural text generation with unlikelihood training. *arXiv preprint arXiv:1908.04319*.

John Wieting and Kevin Gimpel. 2018. Paranmt-50m: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462.

Guangzhi Xiong, Qiao Jin, Xiao Wang, Yin Fang, Haolin Liu, Yifan Yang, Fangyuan Chen, Zhixing Song, Dengyu Wang, Minjia Zhang, and 1 others. 2025. Rag-gym: Optimizing reasoning and search agents with process supervision. *arXiv preprint arXiv:2502.13957*.

Ziyao Xu, Shaohang Wei, Zhuoheng Han, Jing Jin, Zhe Yang, Xiaoguang Li, Haochen Tan, Zhijiang Guo, and Houfeng Wang. 2025. Citecheck: Towards accurate citation faithfulness detection. *arXiv preprint arXiv:2502.10881*.

Naiwen Xue, Fei Xia, Fu-Dong Chiou, and Marta Palmer. 2005. The penn chinese treebank: Phrase structure annotation of a large corpus. *Natural language engineering*, 11(2):207–238.

Kevin Yang and Dan Klein. 2021. Fudge: Controlled text generation with future discriminators. *arXiv preprint arXiv:2104.05218*.

Shuo Yang, Yuqin Dai, Guoqing Wang, Xinran Zheng, Jinfeng Xu, Jinze Li, Zhenzhe Ying, Weiqiang Wang, and Edith CH Ngai. 2025. Realfactbench: A benchmark for evaluating large language models in real-world fact-checking. In *Proceedings of the 33rd*

ACM International Conference on Multimedia, pages 13435–13441.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. Paws-x: A cross-lingual adversarial dataset for paraphrase identification. *arXiv preprint arXiv:1908.11828*.

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*.

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.

Qinan Yu, Jack Merullo, and Ellie Pavlick. 2023. Characterizing mechanisms for factual recall in language models. *arXiv preprint arXiv:2310.15910*.

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.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.

Fred Zhang and Neel Nanda. 2023. Towards best practices of activation patching in language models: Metrics and methods. *arXiv preprint arXiv:2309.16042*.

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

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. Paws: Paraphrase adversaries from word scrambling. *arXiv preprint arXiv:1904.01130*.

Ziyin Zhang, Yikang Liu, Weifang Huang, Junyu Mao, Rui Wang, and Hai Hu. 2024b. Mela: Multilingual evaluation of linguistic acceptability. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2658–2674.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1097–1100.

A Comparative Analysis of Existing LKS Methods

Figure 1 illustrates three strategies for separating knowledge from language: isolation, coverage, and attenuation. Isolate separation (left) aims to localize and isolate knowledge-bearing units or features inside the model (e.g., neurons, heads, subspaces, or modules), enabling targeted, auditable interventions on K while preserving stable L. Override separation (middle) shifts factual authority to an updatable external source $K_{external}$: the model forms a query, retrieves evidence, and conditions generation on the retrieved knowledge to override or suppress conflicting parametric knowledge. Attenuate separation (right) reduces the model's reliance on parametric factual memory by weakening or discouraging K during training (e.g., regularization, unlearning, defactualized objectives), making the model behave more like a language-and-reasoning engine and lowering errors driven by outdated or hallucinated internal facts.

B RAG Optimization

The performance of these retrieval approaches is constrained by three bottlenecks: how well the retrieval is triggered, how well the evidence is selected, and how effectively the evidence is integrated into the 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. (2023b); Ma et al. (2023) refine the query by rewriting, mitigating the mismatch between natural user questions and retrievable evidence to enhance 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. (2024a); 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. (2024); Xiong et al. (2025) focus on process supervision, enforcing intermediate evidence selection before final generation.

C Faithfulness VS Truthfulness

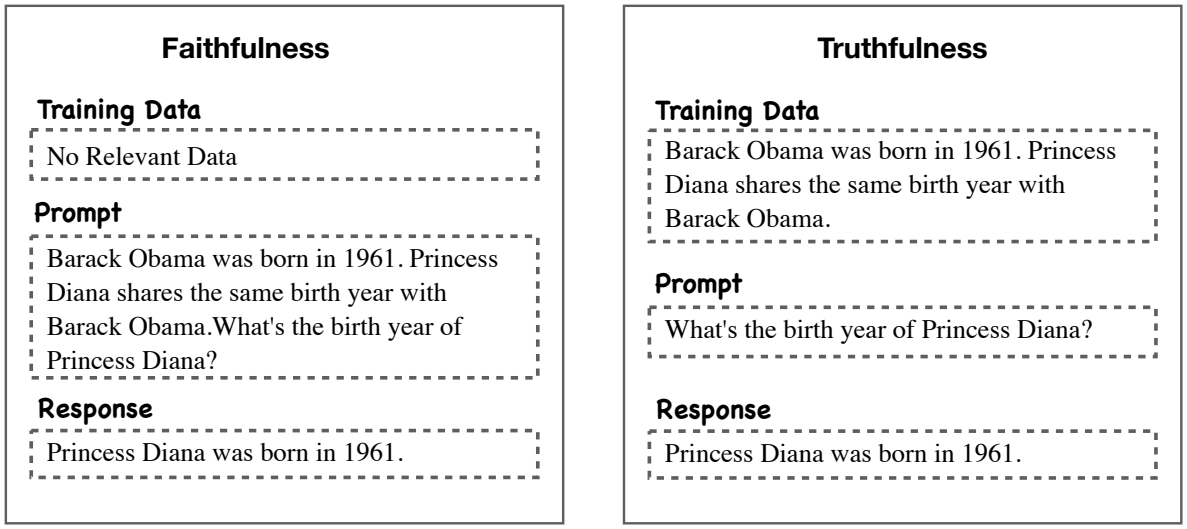

Figure 2: Two kinds of knowledge-side evaluation. Faithfulness measures whether a response can be supported by the given evidence, whereas truthfulness measures whether the model's internal (parametric) knowledge agrees with real-world facts.