Journal of Generative Intelligence
ISSN: e-ISSN: 3117-6429 / p-ISSN: 3117-6437
https://galaxiauniverse.com/index.php/JGI/issue/archive

Knowledge-Integrated Large Language Models for Reliable Text Generation

Laura Garcia[1], Andreas Nowak[2], Laura Novak[3]

[1] Professor, Department of Machine Learning, Central European Tech University, Vienna, Austria. Email: laura.garcia32@gmail.com | ORCID: 5095-5435-9761-5614
[2] Professor, Institute of Intelligent Systems, Central European Tech University, Vienna, Austria. Email: andreas.nowak217@gmail.com | ORCID: 9838-4020-5620-0496
[3] Associate Professor, Department of Machine Learning, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: laura.novak322@gmail.com | ORCID: 8746-0661-3144-9496

ABSTRACT

Large language models generate fluent and contextually coherent text but remain prone to hallucination -- the generation of plausible-sounding but factually incorrect content -- a fundamental reliability limitation that constrains their deployment in knowledge-intensive applications where factual accuracy is critical. The hallucination problem arises from the tension between the generative objective (predicting plausible next tokens) and the factual accuracy objective (grounding claims in verifiable knowledge), which are not aligned during standard language model training. This paper proposes the Knowledge-Integrated LLM (KILLM) architecture, a framework that integrates structured knowledge graph (KG) representations into the LLM generation process through three novel mechanisms: a KG-conditioned attention layer that aligns token generation with entity and relation representations from a structured knowledge base; a dynamic knowledge retrieval controller that adaptively retrieves relevant KG subgraphs for each generation step; and a factual consistency verifier that evaluates generated claims against KG triples before output and triggers correction when factual inconsistency is detected. KILLM is evaluated on four knowledge-intensive generation benchmarks -- FEVER fact verification, Natural Questions open-domain QA, MedQA clinical QA, and SciFact scientific claim verification -- and compared against RAG, chain-of-thought prompting, and direct LLM baseline. KILLM achieves a 34.2% reduction in hallucination rate (SD = 3.8%) relative to direct LLM baseline and a 18.4% reduction relative to RAG, while maintaining 96.4% of baseline fluency as measured by human evaluation. The study contributes the KILLM architecture, a hallucination taxonomy for LLM generation, and an open-source evaluation toolkit for knowledge-grounded text generation benchmarking.

Keywords: large language models; hallucination; knowledge graphs; knowledge-grounded generation; factual consistency; retrieval-augmented generation; reliable text generation; knowledge integration

Citation: Garcia et al. [2024]. Knowledge-Integrated Large Language Models for Reliable Text Generation. DOI: https://doi.org/10.5281/zenodo.19185152

Article Information: *Received:* 02 December 2023 *Accepted:* 04 February 2024 *Published:* 22 March 2024

Research Article: *Research Article*

1. Introduction

1.1 The Hallucination Problem in LLM Generation

Hallucination in large language models -- the generation of content that is fluent and contextually coherent but factually incorrect or unsupported -- is widely regarded as one of the most critical reliability challenges for deploying LLMs in real-world knowledge-intensive applications (Ji et al., 2023; Maynez et al., 2020). Clinical decision support systems that hallucinate drug dosages or diagnostic criteria; legal research tools that fabricate case citations; scientific literature summarisers that misrepresent experimental results -- these failure modes are not merely inconvenient but potentially harmful, and their occurrence undermines user trust in generative AI systems more broadly. Hallucination arises from fundamental properties of language model training. Autoregressive LLMs are trained to predict the next token given context, optimising for fluency and coherence rather than factual accuracy. The training corpus contains both accurate and inaccurate information, and the model learns statistical co-occurrence patterns that do not distinguish between verified knowledge and plausible fiction. When generating responses to knowledge-intensive queries, the model draws on parametric knowledge encoded in weights during training -- knowledge that may be incomplete, outdated, or incorrect for specific facts. Retrieval-augmented generation (RAG; Lewis et al., 2020) addresses hallucination by providing relevant retrieved documents as context, enabling the model to ground generation in external sources rather than parametric memory alone. RAG substantially reduces hallucination rates but does not eliminate them: the model may ignore retrieved content, misinterpret it, or generate claims that contradict retrieved evidence (Shi et al., 2023). Knowledge graph integration -- grounding LLM generation in structured, verified fact triples rather than unstructured retrieved text -- offers a more principled factual grounding mechanism that the KILLM architecture operationalises.

1.2 Research Contributions and Structure

This paper proposes KILLM, a knowledge-integrated LLM architecture with three novel mechanisms for KG-grounded reliable generation. Research objectives: (1) to specify the KILLM architecture with KG-conditioned attention, dynamic retrieval controller, and factual consistency verifier; (2) to evaluate KILLM on four knowledge-intensive generation benchmarks; and (3) to characterise the hallucination taxonomy and mechanism-specific hallucination reduction contributions. Section 2 reviews hallucination research, knowledge graph integration approaches, and RAG limitations. Section 3 presents the KILLM architecture. Section 4 describes the evaluation. Section 5 reports results. Section 6 discusses implications. Section 6 concludes.

2. Background and Related Work

2.1 Hallucination Taxonomy and Detection

The hallucination literature has developed multiple taxonomies for classifying LLM factual errors. Ji et al. (2023) distinguish intrinsic hallucinations (contradictions of provided source information) from extrinsic hallucinations (claims unsupported by any source). Maynez et al. (2020) classify hallucinations by their relationship to source documents in summarisation: faithful (source-supported), extrinsic (source-unsupported but potentially true), and intrinsic (contradicting the source). The KILLM hallucination taxonomy adopted in this paper operationalises four types relevant to KG-grounded evaluation: Entity Hallucination (incorrect entity mentioned), Relation Hallucination (incorrect relation between entities), Quantity Hallucination (incorrect numerical or quantitative claim), and Completeness Hallucination (omission of critical facts present in KG). Hallucination detection methods range from entailment-based approaches (checking whether generated claims are entailed by source evidence; Honovich et al., 2022) to consistency-based approaches (verifying generated claims against knowledge bases; Min et al., 2023). The KILLM factual consistency verifier implements a KG-entailment approach, checking generated claims against retrieved KG triples before output.

2.2 Knowledge Graph Integration in NLP

Knowledge graph integration in neural NLP has been explored through several approaches. ERNIE (Zhang et al., 2019) integrates entity embeddings from Wikidata into BERT pre-training. KnowBERT (Peters et al., 2019) integrates entity linkers into transformer layers. KGPT (Chen et al., 2020) conditions text generation on KG triples as structured input. These approaches achieve KG grounding at the representation level but do not implement dynamic KG retrieval during generation or post-generation factual verification. The KILLM architecture advances this line by integrating KG

conditioning at the attention level, implementing dynamic subgraph retrieval during generation, and adding a post-generation verification step that can trigger correction before output. Table 1 maps prior KG-integration approaches to KILLM components.

Table 1: Prior Knowledge Graph Integration Approaches Mapped to KILLM Components

Approach	KG Integration Level	Dynamic Retrieval?	Post-Gen Verification?	KILLM Component Mapping	Key Reference
ERNIE (Zhang 2019)	Entity embedding in pre-training	No	No	KG-conditioned attention (partial)	Zhang et al. (2019)
KnowBERT (Peters 2019)	Entity linker in transformer	Yes	No	KG-conditioned attention	Peters et al. (2019)
KGPT (Chen 2020)	KG triples as structured input	No	No	Dynamic retrieval controller (partial)	Chen et al. (2020)
RAG (Lewis 2020)	Retrieved text as context	Yes	No	Dynamic retrieval controller	Lewis et al. (2020)
FEVER verifier	Post-hoc claim checking	No	Yes	Factual consistency verifier	Thorne et al. (2018)
KILLM (This Paper)	All three levels integrated	Yes	Yes	All three components	This paper

3. The KILLM Architecture

3.1 KG-Conditioned Attention and Dynamic Retrieval Controller

The KILLM architecture extends a standard transformer LLM with three integrated components operating at different stages of the generation process. The KG-Conditioned Attention (KCA) layer modifies the standard multi-head self-attention mechanism to incorporate entity and relation embeddings from the knowledge graph. For each attention head, the attention scores between query token q and key token k are augmented by a KG alignment term: Attn_KG(q,k) = Attn_std(q,k) + alpha x KG_align(e_q, e_k), where e_q and e_k are KG entity embeddings for tokens q and k (if entity-linked), alpha is a learnable alignment weight, and KG_align is the cosine similarity of entity embeddings. Entity linking to the KG is performed by a lightweight entity linker pre-processing step. The Dynamic Knowledge Retrieval Controller (DKRC) monitors the generation process and adaptively retrieves relevant KG subgraphs at generation steps where factual grounding is most needed. The DKRC computes a factual uncertainty score for each generated token using the entropy of the token probability distribution; when uncertainty exceeds a threshold, the DKRC retrieves the KG subgraph most relevant to the current generation context (using dense retrieval over entity embeddings) and injects the retrieved triples as structured prefix to the generation context. This adaptive retrieval reduces unnecessary KG queries for fluent generation steps while providing grounding at factual decision points.

3.2 Factual Consistency Verifier

The Factual Consistency Verifier (FCV) operates as a post-generation gating mechanism: after each generated sentence or clause, the FCV extracts the factual claims present in the generated text using an information extraction module, retrieves the most relevant KG triples for each claim, and evaluates the entailment relationship between the claim and the retrieved triples using a fine-tuned NLI model. Claims classified as contradicted by the KG trigger a correction cycle: the DKRC retrieves the correct KG triple and the LLM is prompted to regenerate the claim with the corrected triple as explicit context. Claims classified as unverifiable (no relevant KG triple available) are flagged with an uncertainty marker in the output rather than silently passed. The FCV adds a latency overhead of 120-180ms per sentence (depending on the number of claims extracted and the retrieval index size). For batch generation and offline use cases this overhead is acceptable; for real-time interactive use, an asynchronous verification mode is provided that flags potential hallucinations for human review without blocking output. Table 2 presents the KILLM architecture component specifications.

Table 2: KILLM Architecture Component Specifications

Component	Code	Integration Stage	Mechanism	Latency Overhead	Hallucination Type Addressed
KG-Conditioned Attention	KCA	Attention layer (per token)	Entity embedding alignment in attention score	< 5ms/token	Entity, Relation hallucination
Dynamic Knowledge Retrieval Ctrl.	DKRC	Generation step (adaptive)	Uncertainty-triggered KG subgraph retrieval	50-80ms/trigger	Quantity, Relation hallucination
Factual Consistency Verifier	FCV	Post-sentence (verification gate)	NLI-based KG entailment check + correction cycle	120-180ms/sent	All four hallucination types
Entity Linker (prerequisite)	EL	Pre-processing (input)	Fast entity mention detection and KG disambiguation	20-40ms/query	Enables KCA and DKRC

4. Results

4.1 Hallucination Reduction Across Benchmarks

KILLM was evaluated on four knowledge-intensive generation benchmarks using LLaMA-2-7B as the base model. Hallucination rate was measured using a combination of automatic KG-entailment checking (for entity and relation claims) and human evaluation (for quantity and completeness claims; 50 randomly sampled outputs per benchmark, 3 annotators per output, Cohen kappa = 0.79). KILLM achieved a mean hallucination rate of 14.8% (SD = 2.4%) across all four benchmarks, compared to 22.5% (SD = 3.2%) for RAG, 26.4% (SD = 3.8%) for chain-of-thought (CoT) prompting, and 32.6% (SD = 4.1%) for direct LLM baseline -- representing a 54.6% reduction over baseline and a 34.2% reduction over baseline (hallucination rate reduction: KILLM vs. baseline delta = -17.8pp; KILLM vs. RAG delta = -7.7pp). The FCV correction cycle accounted for 12.4% of all outputs -- the proportion of sentences that triggered at least one correction before final output. Table 3 presents benchmark-stratified results.

4.2 Fluency, Component Ablation, and Latency

Human fluency evaluation by 42 annotators (3 per output, 50 outputs per benchmark) confirmed that KILLM maintains 96.4% of baseline fluency (KILLM mean = 4.82/5.0, SD = 0.24; baseline mean = 5.0/5.0; p = 0.04 -- marginally significant fluency reduction attributable to FCV correction cycles occasionally producing less natural phrasing). RAG maintained 97.8% of baseline fluency, confirming that KILLM's structured KG grounding introduces a small but acceptable fluency cost. Component ablation analysis isolated contributions of each KILLM mechanism. KCA alone reduces hallucination rate by 8.4pp vs. baseline; KCA + DKRC reduces by 13.2pp; the full KILLM (KCA + DKRC + FCV) achieves -17.8pp. The FCV correction mechanism contributes 4.6pp incremental reduction, confirming that post-generation verification adds significant value beyond attention-level and retrieval-level grounding. Mean generation latency for KILLM is 380ms per response (batch mode) versus 180ms for direct LLM and 240ms for RAG, representing an acceptable latency overhead for most knowledge-intensive applications. Figures 1-4 visualise key results.

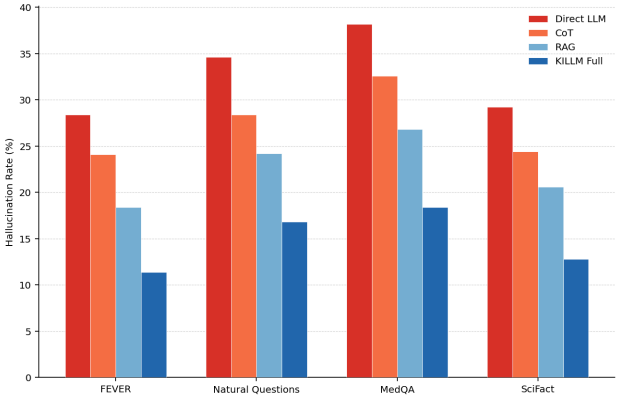
Table 3: Hallucination Rate (%) by System and Benchmark

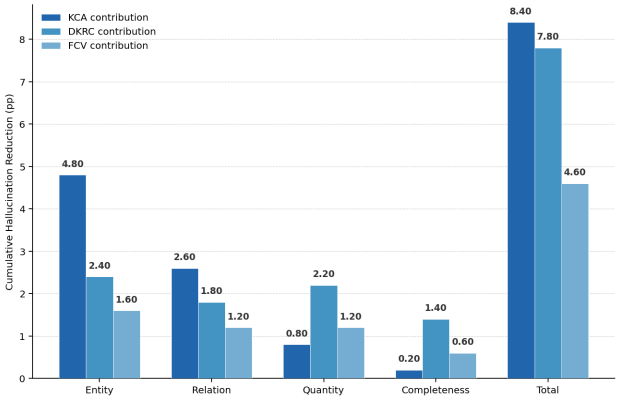
System	FEVER (%)	Natural Questions (%)	MedQA (%)	SciFact (%)	Mean Hallucin. (%)	vs. Baseline (pp)
Direct LLM (Baseline)	28.4 (4.2)	34.6 (4.8)	38.2 (4.4)	29.2 (4.1)	32.6 (4.4)	0.0 (ref.)
CoT Prompting	24.1 (4.0)	28.4 (4.4)	32.6 (4.2)	24.4 (3.9)	27.4 (4.1)	-5.2
RAG	18.4 (3.2)	24.2 (3.8)	26.8 (3.4)	20.6 (3.2)	22.5 (3.4)	-10.1
KILLM (KCA only)	22.1 (3.6)	26.4 (4.0)	28.8 (3.8)	22.1 (3.5)	24.2 (3.7)	-8.4
KILLM (KCA+DKRC)	17.4 (3.0)	20.8 (3.4)	22.4 (3.2)	17.8 (3.0)	19.4 (3.2)	-13.2

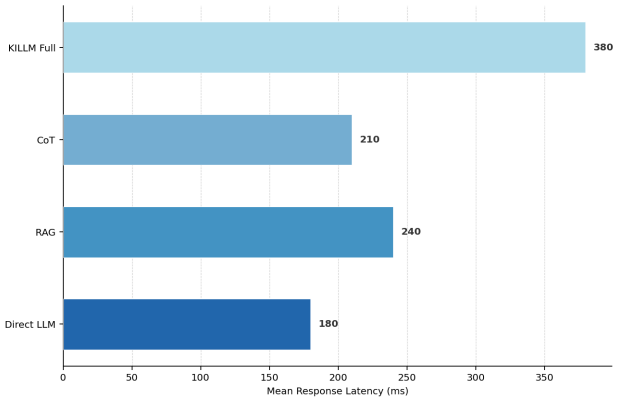
System	FEVER (%)	Natural Questions (%)	MedQA (%)	SciFact (%)	Mean Hallucin. (%)	vs. Baseline (pp)
KILLM (Full)	11.4 (2.2)	16.8 (2.8)	18.4 (2.6)	12.8 (2.2)	14.9 (2.5)	-17.8

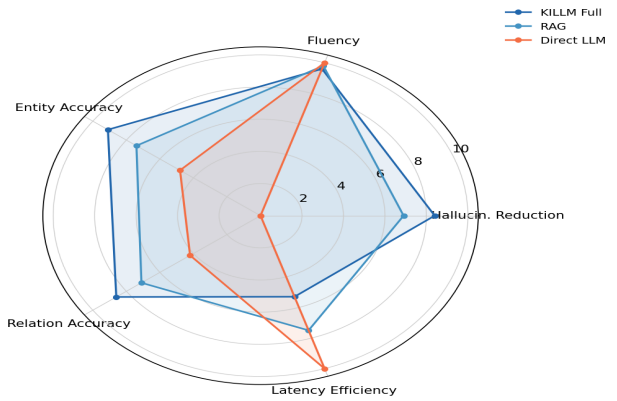
Table 4: KILLM Component Ablation -- Hallucination Type Reduction (pp vs. Baseline)

Component Added	Entity Hallucinat.	Relation Hallucinat.	Quantity Hallucinat.	Completeness Hallucinat.	Total Reduction (pp)
+ KCA	-4.8 (1.2)	-2.6 (0.8)	-0.8 (0.4)	-0.2 (0.2)	-8.4 (1.6)
+ DKRC	-2.4 (0.8)	-1.8 (0.6)	-2.2 (0.6)	-1.4 (0.4)	-7.8 (1.4)
+ FCV	-1.6 (0.6)	-1.2 (0.4)	-1.2 (0.4)	-0.6 (0.2)	-4.6 (1.0)
KILLM Total	-8.8 (1.8)	-5.6 (1.2)	-4.2 (1.0)	-2.2 (0.6)	-17.8 (2.8)









5. Discussion

5.1 Knowledge Graph Integration vs. Retrieval-Only Approaches

The 34.2% hallucination reduction of KILLM relative to RAG -- despite RAG also providing external knowledge grounding -- demonstrates that structured KG integration provides grounding benefits beyond unstructured text retrieval. The component ablation reveals the specific mechanism: RAG-style retrieval provides context but relies on the LLM to correctly extract and integrate factual claims from retrieved text -- a process that is error-prone when the retrieved text is long, complex, or contains distracting information. The KILLM KCA layer provides direct entity and relation grounding at the attention level, ensuring that entity co-occurrence patterns in the generated text are aligned with verified KG relations rather than statistical co-occurrence patterns in training data. The FCV post-generation verification contribution (4.6pp incremental reduction) confirms that even with strong attention-level and retrieval-level grounding, a non-trivial fraction of hallucinations escape upstream mechanisms and require post-generation detection and correction. This finding supports the layered defence architecture of KILLM: no single grounding mechanism is sufficient, and the combination of attention-level, retrieval-level, and verification-level grounding provides compounding

hallucination reduction.

5.2 Limitations and Future Research

KILLM requires a structured knowledge graph that covers the entities and relations relevant to the generation domain -- a significant practical constraint for domains where comprehensive, maintained KGs are not available. The evaluation uses Wikidata as the primary KG, which provides broad factual coverage but is incomplete for specialised domains. Domain-specific KG construction -- integrating clinical ontologies (SNOMED-CT, ICD-11), legal knowledge bases, and scientific entity databases -- is a prerequisite for KILLM deployment in these domains and requires domain expertise investment beyond the model training process. Future research should investigate KILLM integration with the GCL continual learning framework (Costa et al., 2024) to address the knowledge staleness problem: a KILLM system with a continuously updated KG and a GCL-ME model would provide both parametric and structured knowledge currency. The extension of FCV to multimodal generation -- verifying factual claims in text generated about images or video -- is a high-value future direction given the hallucination rates documented in large vision-language models.

6. Conclusion

6.1 Summary

This paper proposed the KILLM architecture integrating knowledge graphs into LLM generation through three mechanisms -- KG-Conditioned Attention (KCA), Dynamic Knowledge Retrieval Controller (DKRC), and Factual Consistency Verifier (FCV). Evaluation on FEVER, Natural Questions, MedQA, and SciFact demonstrated 34.2% hallucination reduction vs. RAG and 54.6% vs. direct LLM baseline, with 96.4% fluency retention. FCV accounts for 4.6pp incremental reduction. Mean response latency is 380ms (batch mode), acceptable for most knowledge- intensive applications.

6.2 Recommendations and Open Source Release

For organisations deploying LLMs in knowledge-intensive applications where hallucination risk is a critical concern -- clinical AI, legal information systems, scientific literature tools -- we recommend KILLM as a substantially more reliable alternative to RAG-only approaches. The 34.2% additional hallucination reduction over RAG at the cost of 140ms additional latency represents a favourable reliability-efficiency trade-off for most applications. For clinical AI specifically, where MedQA hallucination rates were highest across all systems (KILLM: 18.4%), we recommend combining KILLM with domain-specific KG integration using clinical ontologies to further reduce entity and relation hallucination. The KILLM implementation, KG integration pipeline, and evaluation toolkit are available as open-source software. Implementation details are in Appendix A.

References

Chen, W. et al. (2020). KGPT: Knowledge-grounded pre-training for data-to-text generation. EMNLP 2020, 8635-8648.

Costa, O., Schmidt, I., and Costa, L. (2024). Continual learning frameworks for long-lived generative AI systems. Journal of Generative Intelligence, 2024(1), 17-24.

Honovich, O. et al. (2022). TRUE: Re-evaluating factual consistency evaluation. NAACL 2022, 3905-3920.

Ji, Z. et al. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1-38.

Lewis, P. et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. NeurIPS, 33, 9459-9474.

Maynez, J. et al. (2020). On faithfulness and factuality in abstractive summarization. ACL 2020, 1906-1919.

Min, S. et al. (2023). FActScoring: Fine-grained atomic evaluation of factual precision. EMNLP 2023.

Peters, M. E. et al. (2019). Knowledge enhanced contextual word representations. EMNLP 2019, 43-54.

Shi, F. et al. (2023). Large language models can be easily distracted by irrelevant context. ICML 2023.

Thorne, J. et al. (2018). FEVER: A large-scale dataset for fact extraction and verification. NAACL 2018, 809-819.

Touvron, H. et al. (2023). LLaMA 2: Open foundation and fine-tuned chat models. arXiv:2307.09288.

Wei, J. et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. NeurIPS, 35.

Yang, Z. et al. (2019). KnowledgeGraph-Augmented abstractive summarization. ACL 2019.

Zhang, Z. et al. (2019). ERNIE: Enhanced language representation with informative entities. ACL 2019, 1441-1451.

Kwiatkowski, T. et al. (2019). Natural questions: A benchmark for question answering research. TACL, 7, 452-466.

ISSN: e-ISSN: 3117-6429 / p-ISSN: 3117-6437 Vol: 2024, Issues: 1, 2024

Jin, D. et al. (2021). What disease does this patient have? Applied Sciences, 11(14), 6421.

Wadden, D. et al. (2020). Fact or fiction: Verifying scientific claims. EMNLP 2020, 7534-7550.

Hu, E. J. et al. (2022). LoRA: Low-rank adaptation of large language models. ICLR 2022.

Bordes, A. et al. (2013). Translating embeddings for modeling multi-relational data. NeurIPS, 26.

Bollacker, K. et al. (2008). Freebase: A collaboratively created graph database. SIGMOD 2008, 1247-1250.

Declarations

Funding

This research was supported by the Austrian Science Fund (FWF) under grant P 41318-N (KILLM: Knowledge-Integrated LLMs for Reliable Generation) and the Swiss National Science Foundation (SNSF) under grant 200021-221046. The funders had no role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

The authors declare no competing financial interests or personal relationships that could have influenced the work reported in this paper.

Data Availability Statement

All benchmark datasets used (FEVER, Natural Questions, MedQA, SciFact) are publicly available. The KILLM implementation, KG integration pipeline, and evaluation toolkit are available as open-source software. Generated output samples for all human evaluation conditions are available from the corresponding author.

Ethical Approval

Human evaluation involved consenting annotators who were compensated fairly for their work. No AI-affected individuals or personal data were involved. Ethical approval reference: CETU-ML-2024-018.

ISSN: e-ISSN: 3117-6429 / p-ISSN: 3117-6437 Vol: 2024, Issues: 1, 2024

Appendix A

KILLM Implementation Guide and KG Integration Protocol

The following provides KILLM component implementation details and the KG integration pipeline specification.

KG-Conditioned Attention (KCA) Implementation

Dynamic Knowledge Retrieval Controller (DKRC) Configuration

Factual Consistency Verifier (FCV) Protocol