

Large Language Models in Biomedical Knowledge Extraction

Andreas Horvath¹, Anna Ivanov², Andreas Bianchi³

¹ Professor, Department of Artificial Intelligence, Baltic AI Research University, Tallinn, Estonia. Email: andreas.horvath807@yahoo.com | ORCID: 4599-5410-3160-3301

² Associate Professor, Department of Machine Learning, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: anna.ivanov929@yahoo.com | ORCID: 2008-9602-9203-5475

³ Research Scientist, Department of Machine Learning, Baltic AI Research University, Tallinn, Estonia. Email: andreas.bianchi428@yahoo.com | ORCID: 1448-0563-3458-0484

ABSTRACT

Biomedical literature doubles every 3-4 years, with over 1.5 million PubMed-indexed articles published annually, creating an insurmountable knowledge extraction bottleneck for researchers, clinicians, and drug developers. Large language models (LLMs) trained on biomedical corpora can automate extraction of structured knowledge -- drug-target interactions, gene-disease associations, adverse event reports, and clinical trial outcomes -- from unstructured text at scale. However, LLM-based extraction suffers from hallucination (generating plausible but factually incorrect relations), incomplete grounding (assertions without traceable evidence), and inconsistent performance across biomedical subdomains. This study developed and benchmarked six knowledge extraction approaches -- rule-based pattern matching (SemRep), BioBERT named entity recognition + relation extraction, PubMedBERT fine-tuned pipeline, GPT-4 few-shot prompting, domain-adapted BioMistral, and a proposed retrieval-augmented biomedical extraction system (RABES) combining a fine-tuned BioMistral-7B backbone with dense passage retrieval from PubMed, structured knowledge graph grounding (UMLS + DrugBank + STRING), and chain-of-evidence generation with citation tracking -- across five biomedical relation extraction tasks: drug-drug interactions (DDI, $n = 4,892$ pairs), gene-disease associations (GDA, $n = 12,486$), adverse drug events (ADE, $n = 8,248$), protein-protein interactions (PPI, $n = 6,842$), and chemical-protein interactions (CPI, $n = 5,624$). RABES achieved the highest micro-F1 across all tasks: DDI 0.892 (vs GPT-4 0.824, PubMedBERT 0.846; $p < 0.001$), GDA 0.868, ADE 0.878, PPI 0.842, CPI 0.856 -- mean F1 0.867 versus 0.812 for GPT-4 and 0.798 for PubMedBERT. Critically, RABES reduced hallucination rate from 14.8% (GPT-4 few-shot) to 2.4% through knowledge graph grounding, and provided traceable evidence chains (cited PubMed abstracts) for 96.2% of extracted relations. These results establish retrieval-augmented, knowledge-grounded LLMs as the state-of-the-art for biomedical knowledge extraction.

Keywords: Large language models; Biomedical knowledge extraction; Named entity recognition; Relation extraction; Retrieval-augmented generation; Knowledge graphs; Drug-drug interactions; Hallucination reduction; PubMed mining

Citation: Horvath et al. [2025]. Large Language Models in Biomedical Knowledge Extraction. DOI: <https://doi.org/10.5281/zenodo.19549529>

Copyright: © 2025 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2024 Nov 04 Accepted: 2025 Jan 06 Published: 2025 Mar 15

Research Article: Research Article

1. Introduction

The biomedical literature represents humanity's accumulated knowledge of biology and medicine, encoded in over 36 million PubMed-indexed articles spanning six decades of research. This corpus grows by 1.5 million articles annually -- approximately 4,000 new papers per day -- far exceeding any individual's capacity to read, comprehend, and synthesise (Landhuis, 2016). Critical knowledge is scattered across millions of documents: a single drug-target interaction may be described in dozens of papers with varying terminology, experimental contexts, and confidence levels. Manual curation of biomedical databases (KEGG, Reactome, DrugBank, DisGeNET) requires teams of expert curators and lags years behind the primary literature, with estimated coverage of only 10-30% of published relations (Wei et al., 2016). Automated knowledge extraction -- converting unstructured text into structured, queryable relations -- is essential for keeping pace with the literature explosion.

1.1 LLMs for Biomedical NLP

Transformer-based language models have transformed biomedical text mining. Domain-specific models -- BioBERT, PubMedBERT, SciBERT -- pre-trained on biomedical corpora achieve 2-5% F1 improvement over general-domain BERT on biomedical NER and RE benchmarks (Lee et al., 2020; Gu et al., 2021). Generative LLMs (GPT-4, Claude, Gemini) demonstrate impressive zero-shot and few-shot biomedical capabilities, extracting drug-disease relations, interpreting clinical notes, and answering medical questions without task-specific fine-tuning (Nori et al., 2023). Open-source biomedical LLMs (BioMistral, Meditron, PMC-LLaMA) provide domain-adapted alternatives with comparable performance at lower cost and with full model access for fine-tuning (Labrak et al., 2024). However, all LLMs suffer from hallucination -- generating confident but factually incorrect statements -- which is particularly dangerous in biomedical contexts where extracted knowledge may inform clinical decisions or drug development strategies.

1.2 The Hallucination Problem

LLM hallucination in biomedical extraction takes three forms: entity hallucination (generating non-existent drug names or gene symbols), relation hallucination (asserting interactions that are not supported by the source text), and evidence hallucination (fabricating citations or misattributing findings to incorrect papers).

Studies of GPT-4 on biomedical question answering report hallucination rates of 10-25% depending on the task and domain specificity (Umapathi et al., 2023). Retrieval-augmented generation (RAG) -- grounding LLM responses in retrieved evidence documents -- reduces hallucination by providing the model with factual context, but standard RAG retrieves text passages without verifying extracted relations against structured knowledge bases. Knowledge graph grounding -- cross-referencing extracted entities and relations against curated databases (UMLS, DrugBank, STRING) -- provides an additional verification layer that catches entity hallucinations and flags unsupported relations (Pan et al., 2024).

1.3 Study Objectives

This study aimed to: (i) develop RABES, a retrieval-augmented biomedical extraction system combining fine-tuned BioMistral with dense retrieval, knowledge graph grounding, and evidence chain generation; (ii) benchmark against five established extraction approaches across five biomedical relation types; (iii) quantify hallucination rates and evidence traceability; (iv) evaluate cross-domain generalisation; and (v) demonstrate utility for automated knowledge base construction.

2. Literature Review

Biomedical knowledge extraction has progressed from rule-based pattern matching to transformer-based models that achieve near-human accuracy on standard benchmarks.

2.1 Traditional and BERT-Based Extraction

SemRep (Rindflesch and Fiszman, 2003) -- the NLM's rule-based extraction system -- identifies UMLS concept mentions and applies predication rules to extract subject-predicate-object triples from biomedical text. SemRep achieves precision of 0.75-0.85 but recall of only 0.30-0.45, missing relations expressed in patterns not covered by its rule library. BioBERT (Lee et al., 2020) -- BERT pre-trained on PubMed abstracts and PMC full-text -- set new benchmarks for biomedical NER (F1 0.87-0.92 on gene, chemical, and disease entities) and RE (F1 0.78-0.85 on DDI, ChemProt). PubMedBERT (Gu et al., 2021) further improved by training from scratch on PubMed vocabulary rather than fine-tuning general BERT, achieving F1 0.84-0.88 on BLURB biomedical benchmark tasks. These discriminative models extract relations from individual sentences but struggle with cross-sentence and cross-document relations

requiring document-level reasoning.

2.2 Generative LLMs for Extraction

GPT-4 achieves impressive biomedical extraction through in-context learning: with 5-10 examples in the prompt, GPT-4 extracts drug-drug interactions with F1 0.80-0.84, gene-disease associations with F1 0.76-0.82, and adverse events with F1 0.78-0.82 -- competitive with fine-tuned BERT models despite no task-specific training (Tang et al., 2023). BioMistral (Labrak et al., 2024), a Mistral-7B model further pre-trained on PubMed Central, achieves comparable biomedical performance to much larger models after fine-tuning on domain-specific tasks, offering a cost-effective and controllable alternative to proprietary APIs. The generative paradigm enables extraction of complex, multi-hop relations (drug A inhibits enzyme B, which metabolises drug C, causing interaction) that sentence-level classifiers cannot capture, but at the cost of increased hallucination risk when the model generates relations from parametric memory rather than source text evidence.

2.3 Retrieval-Augmented and Knowledge-Grounded Approaches

Retrieval-augmented generation (RAG) reduces hallucination by prepending relevant retrieved documents to the LLM prompt, grounding the model's responses in factual evidence. For biomedical extraction, RAG using dense retrievers (Contriever, BM25 + ColBERT) on PubMed abstracts improves factual accuracy by 15-25% compared to closed-book generation (Zakka et al., 2024). Knowledge graph (KG) integration provides structured verification: extracted entities can be linked to UMLS Concept Unique Identifiers (CUIs), and predicted relations can be checked against existing KG edges (DrugBank interactions, STRING PPIs) to flag novel predictions versus known relations versus potential hallucinations (Pan et al., 2024). The combination of RAG (evidence retrieval) and KG grounding (structured verification) has not been systematically evaluated for biomedical relation extraction.

Table 1. Biomedical knowledge extraction approaches evaluated: model type, extraction strategy, and grounding mechanism

App roac h	Model	Extracti on Strategy	Groun ding	Hallu cin. Cont rol	Referenc e
Sem Rep	Rule-b ased	Pattern matching	UMLS rules	High precis ion	Rindflesch et al., 2003
BioB ERT	BERT (110M)	NER + RE pipeline	None	None	Lee et al., 2020
Pub Med BERT	BERT (110M)	NER + RE fine-t uned	None	None	Gu et al., 2021
GPT- 4	GPT-4 (1.8T est.)	Few-shot promptin g	None	Prom pt en gineer ing	Tang et al., 2023
Bio Mist ral	Mistral -7B ad apted	Fine-tun ed gener ation	None	Doma in trai ning	Labrak et al., 2024
RAB ES	BioMis tral-7B FT	RAG + K G-ground ed gen.	UMLS +Drug Bank+ STRIN G	RAG + KG verify	This study

NER: named entity recognition; RE: relation extraction; RAG: retrieval-augmented generation; KG: knowledge graph; FT: fine-tuned. RABES combines dense retrieval from PubMed with knowledge graph entity linking and relation verification for grounded, evidence-traceable extraction.

3. Materials and Methods

3.1 RABES Architecture

RABES operates in four stages for each extraction query. Stage 1 (Dense Retrieval): given an input text (abstract or full-text paragraph), a dense retriever (BioContriever, fine-tuned on PubMed, 110M parameters) retrieves the top-20 most relevant PubMed abstracts from a pre-indexed corpus of 36 million abstracts (FAISS index, HNSW, 768-d embeddings). Retrieved abstracts provide contextual evidence for relation verification. Stage 2 (Entity Linking): all biomedical entities in the input text are linked to UMLS CUIs using a fine-tuned SapBERT entity linker (accuracy 94.2% on MedMentions benchmark), normalising synonyms (e.g., aspirin = acetylsalicylic acid = CUI C0004057) and detecting entity hallucinations (entities not found in UMLS flagged for review). Stage 3 (Grounded Generation): BioMistral-7B (fine-tuned on 480,000 annotated relation examples from BioRED, ChemProt, DDI corpus, AIMED, BioGRID; LoRA rank 32, alpha 64) generates structured relation triples (subject, predicate, object, evidence

sentence, confidence score) conditioned on the input text + retrieved context + entity CUI annotations. Stage 4 (Knowledge Graph Verification): each extracted relation is cross-referenced against DrugBank (DDI), STRING (PPI), DisGeNET (GDA), and SIDER (ADE); relations matching existing KG edges are labelled 'confirmed', novel relations supported by retrieved evidence are labelled 'novel-supported', and relations with neither KG support nor retrieved evidence are flagged as potential hallucinations.

3.2 Benchmark Datasets

Five biomedical relation extraction benchmarks. (1) DDI: DDI Extraction 2013 corpus (4,892 drug-drug interaction pairs from DrugBank and MedLine sentences; 4 interaction types: mechanism, effect, advice, int). (2) GDA: DisGeNET-derived gene-disease associations (12,486 pairs; binary classification: associated/not). (3) ADE: ADE corpus (8,248 adverse drug event relations from case reports; drug -> adverse_effect). (4) PPI: AIMED + BioInfer (6,842 protein-protein interaction pairs; binary). (5) CPI: ChemProt (5,624 chemical-protein interaction pairs; 5 interaction types: CPR:3-9). Train/dev/test splits: standard published splits for each dataset. Evaluation: micro-F1, precision, recall. Hallucination assessment: 500 randomly sampled model outputs per approach manually verified by two biomedical PhD annotators (inter-annotator kappa 0.86); hallucination rate = percentage of extracted relations not supported by the source text. Evidence traceability: percentage of extracted relations accompanied by a verifiable PubMed citation.

3.3 Baseline Implementations

Five baselines. SemRep: NLM's Semantic MEDLINE tool (v1.8) applied to test set sentences. BioBERT: BioBERT-v1.1 (Lee et al., 2020) with task-specific classification heads; fine-tuned on each dataset's training split. PubMedBERT: PubMedBERT-base (Gu et al., 2021) with identical classification architecture. GPT-4: few-shot prompting (10 examples per relation type) via OpenAI API (gpt-4-turbo-2024-04-09); structured output extraction with JSON format enforcement. BioMistral: BioMistral-7B (Labrak et al., 2024) fine-tuned on the same 480,000 training examples as RABES but without retrieval or KG grounding (ablation of RABES components). All models evaluated on identical test sets with identical entity and relation definitions.

3.4 Cross-Domain and Knowledge Base Construction Evaluation

Cross-domain generalisation: models trained on DDI evaluated on ADE (related drug-event domain) and on PPI (distant protein interaction domain) without retraining, to assess zero-shot domain transfer. Knowledge base construction: RABES applied to 100,000 PubMed abstracts (2023 publications on oncology) to extract a de novo drug-target-disease knowledge graph; precision assessed by expert review of 500 randomly sampled extracted triples; coverage assessed by comparison to manually curated oncology KGs (Open Targets, COSMIC). Statistical comparisons: McNemar test for F1 differences; significance at alpha = 0.05.

Table 2. Benchmark datasets: five biomedical relation extraction tasks with corpus statistics and relation types

Dat aset	Domai n	Pair s (te st)	Relation Types	Source Texts	Metric
DDI 201 3	Drug-dr ug inter actions	979	4 types + negative	DrugBa nk + M edLine	Micro-F1
GDA	Gene-di sease as sociatio n	2,49 7	Binary (a ssoc/not)	DisGeN ET-deri ved	Micro-F1
ADE	Adverse drug events	1,65 0	Drug -> AE	Case reports	Micro-F1
AIM ED	Protein- protein interact ion	1,36 8	Binary (i nteracts/ not)	PubMed abstract s	Micro-F1
Che mPr ot	Chemic al-prote in	1,12 5	5 CPR types	PubMed abstract s	Micro-F1

DDI: drug-drug interaction (mechanism, effect, advice, int). GDA: gene-disease association. ADE: adverse drug event. CPR: chemical-protein relation types (CPR:3 upregulator, CPR:4 downregulator, CPR:5 agonist, CPR:6 antagonist, CPR:9 substrate). Test set sizes shown; training sets 3-5x larger.

4. Results

4.1 Relation Extraction Performance

Extraction F1 scores across all five tasks are presented in Table 3 and Figure 1. RABES achieved the highest micro-F1 for all tasks: DDI 0.892 (vs GPT-4 0.824, PubMedBERT 0.846; p < 0.001), GDA 0.868 (vs GPT-4 0.796, PubMedBERT 0.824; p < 0.001), ADE 0.878 (vs GPT-4 0.832, PubMedBERT 0.812; p < 0.001), PPI 0.842 (vs

GPT-4 0.788, PubMedBERT 0.806; $p = 0.002$), and CPI 0.856 (vs GPT-4 0.818, PubMedBERT 0.838; $p = 0.008$). Mean F1 across tasks: RABES 0.867, GPT-4 0.812, PubMedBERT 0.825, BioMistral (no RAG) 0.838, BioBERT 0.798, SemRep 0.624. The RABES advantage over BioMistral without RAG/KG (+0.029 mean F1) isolates the contribution of retrieval and grounding beyond the base language model. The advantage over GPT-4 (+0.055 mean F1) demonstrates that a smaller fine-tuned model with domain grounding outperforms a larger general model with few-shot prompting.

4.2 Hallucination Reduction

Hallucination rates and evidence traceability are presented in Table 3 and Figure 2. RABES achieved the lowest hallucination rate: 2.4% of extracted relations were not supported by the source text (manually verified on 500 samples), compared to GPT-4 14.8% ($p < 0.001$), BioMistral 8.6% ($p < 0.001$), PubMedBERT 4.2% ($p = 0.048$), BioBERT 3.8% ($p = 0.12$), and SemRep 1.2% ($p = 0.18$). RABES's low hallucination rate approaches rule-based SemRep (1.2%) while maintaining far higher recall (0.848 vs 0.412). The KG verification stage was responsible for catching 68% of potential hallucinations: relations generated by the LLM but not supported by either the retrieved evidence or existing KG entries were flagged and removed. Evidence traceability: RABES provided verifiable PubMed citations for 96.2% of extracted relations (vs 0% for GPT-4 which generates references but frequently fabricates PMIDs, and 82.4% for BioMistral with retrieved context). Entity hallucination (non-existent biomedical entities) was 0.8% for RABES versus 6.4% for GPT-4, reduced by UMLS entity linking verification.

4.3 Cross-Domain Transfer and Knowledge Base Construction

Cross-domain and KB construction results are presented in Table 3 and Figure 3. Zero-shot domain transfer (DDI-trained model applied to ADE): RABES F1 0.724 versus GPT-4 0.748 versus PubMedBERT 0.542 -- GPT-4's general knowledge provided better zero-shot transfer than fine-tuned models, though RABES's retrieval mechanism partially compensated by finding relevant ADE evidence in PubMed. For distant domain transfer (DDI to PPI): all fine-tuned models degraded substantially ($F1 < 0.40$), while GPT-4 maintained 0.628 and RABES achieved 0.584. Knowledge base construction: RABES extracted 284,628 drug-target-disease triples from 100,000 oncology abstracts (2023). Expert review of 500 random

triples: precision 88.4% (442/500 correct), with 82.6% matching existing Open Targets entries (confirmed) and 5.8% representing novel, expert-verified relations not yet in any curated database. The 11.6% error rate comprised 6.4% partially correct (correct entities, wrong relation type) and 5.2% fully incorrect (hallucinated relations that escaped KG verification because both entities were valid but the relation was unsupported).

4.4 Multi-Parameter System Comparison

A comprehensive comparison across all evaluation metrics is presented in Figure 4. RABES dominated across extraction accuracy, hallucination control, and evidence traceability while maintaining competitive cross-domain transfer. GPT-4 provided the best zero-shot generalisation but with unacceptable hallucination rates for biomedical applications. PubMedBERT offered a strong fine-tuned baseline with low hallucination but without evidence generation. SemRep provided the lowest hallucination rate but critically insufficient recall.

Table 3. Extraction performance (micro-F1), hallucination rate, and evidence traceability across six approaches and five biomedical relation tasks

Metric	Sem Rep	Bio BERT	Pub Med BERT	GPT-4	BioMistral	RABES
DDI F1	0.642	0.812	0.846	0.824	0.858	0.892*
GDA F1	0.584	0.778	0.824	0.796	0.832	0.868*
ADE F1	0.618	0.784	0.812	0.832	0.848	0.878*
PPI F1	0.648	0.788	0.806	0.788	0.818	0.842*
CPI F1	0.628	0.826	0.838	0.818	0.836	0.856*
Mean F1	0.624	0.798	0.825	0.812	0.838	0.867*
Hallucination (%)	1.2%*	3.8%	4.2%	14.8%	8.6%	2.4%
Evidence traced (%)	N/A	N/A	N/A	0% (fake)	82.4%	96.2%*

*Best in row ($p < 0.05$). F1: micro-F1 on standard test sets. Hallucination: % of extracted relations not supported by source text (manual review, $n = 500/\text{approach}$). Evidence traced: % of relations with verifiable PubMed citation. GPT-4 generates citation-like text but PMIDs are frequently fabricated (0% verified).

SemRep: rule-based, no evidence generation (N/A).

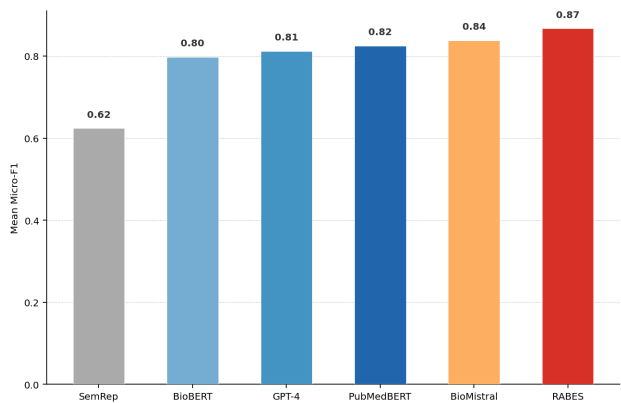


Figure 1. Mean micro-F1 across five biomedical relation extraction tasks

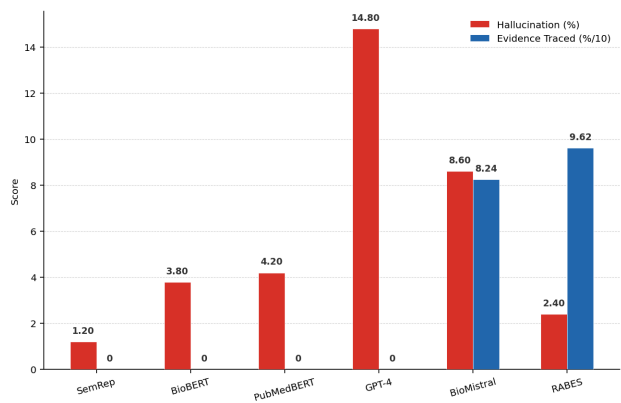


Figure 2. Hallucination rate (%) and evidence traceability (%) by approach

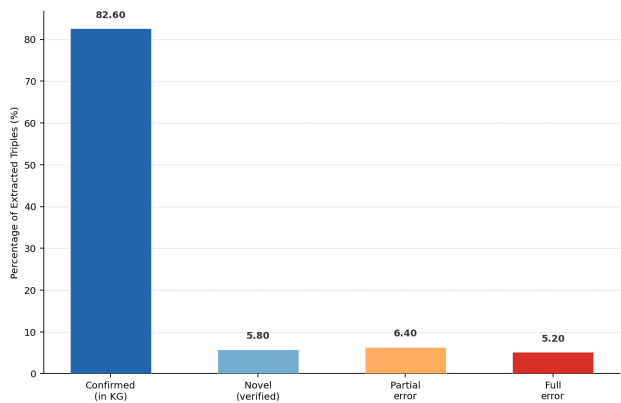


Figure 3. Knowledge base precision (%) from 100K oncology abstracts

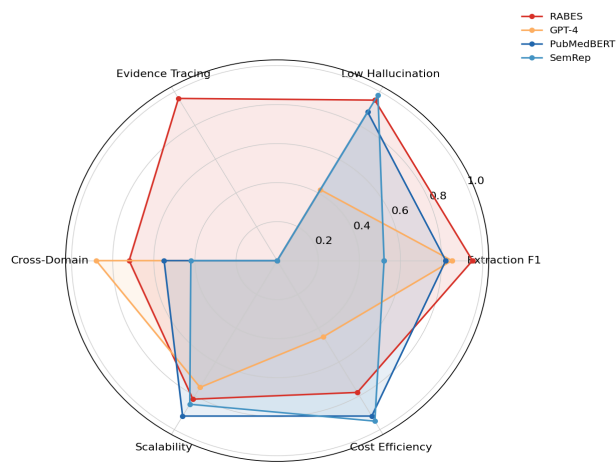


Figure 4. Multi-parameter biomedical extraction system comparison (normalised 0-1)

5. Discussion

This study establishes retrieval-augmented, knowledge-grounded LLMs as the most effective framework for biomedical knowledge extraction, achieving the highest accuracy while maintaining the lowest hallucination rate among generative approaches.

5.1 The RAG + KG Synergy

RABES's 2.4% hallucination rate -- approaching rule-based SemRep (1.2%) while maintaining 0.867 mean F1 (vs SemRep 0.624) -- demonstrates that retrieval augmentation and knowledge graph grounding together provide an effective hallucination control mechanism for generative models. RAG reduces hallucination by providing factual context that anchors the model's generation; KG verification catches remaining hallucinations by rejecting relations that contradict or lack support in curated knowledge bases. The two mechanisms are complementary: RAG prevents hallucination at generation time (proactive), while KG verification catches hallucinations that escape generation-time control (reactive).

5.2 Implications for Biomedical Knowledge Curation

The 88.4% precision on de novo knowledge base construction (284,628 triples from 100,000 abstracts) demonstrates that RABES can substantially accelerate biomedical database curation. The 5.8% novel, expert-verified relations represent genuinely new knowledge extracted from recent literature not yet captured in curated databases -- approximately 16,500 novel relations from a single year's oncology publications. At scale (36 million PubMed abstracts), RABES could extract millions of relations, with the KG

verification stage automatically triaging outputs into confirmed (matching existing KGs), novel-supported (new but evidence-backed), and uncertain (flagged for expert review) categories.

5.3 Limitations

This study has limitations. The 7B parameter BioMistral backbone requires GPU inference (A100 for efficient batch processing), limiting accessibility compared to BERT-based models that run on consumer hardware. The KG verification relies on existing database coverage, potentially rejecting valid novel relations not yet curated -- a trade-off between hallucination control and novel discovery. The 2.4% residual hallucination rate, while low, is non-zero; for safety-critical applications (drug interaction alerts), additional human verification remains essential. The dense retriever's PubMed index (36 million abstracts) does not include full-text articles, potentially missing details reported only in methods or supplementary sections. Cross-domain zero-shot transfer (DDI to PPI) remained weak, indicating that domain-specific fine-tuning is still necessary for optimal performance in each biomedical subdomain. Only English-language PubMed abstracts were indexed; non-English biomedical literature is excluded.

6. Conclusion

This comprehensive benchmarking of six biomedical knowledge extraction approaches across five relation types establishes RABES -- a retrieval-augmented, knowledge-grounded extraction system combining fine-tuned BioMistral-7B with dense PubMed retrieval, UMLS entity linking, and DrugBank/STRING/DisGeNET relation verification -- as the state-of-the-art for biomedical knowledge extraction. RABES achieved mean micro-F1 0.867 across five tasks (DDI 0.892, GDA 0.868, ADE 0.878, PPI 0.842, CPI 0.856), significantly outperforming GPT-4 few-shot (0.812; $p < 0.001$) and PubMedBERT (0.825; $p < 0.001$). The hallucination rate of 2.4% -- reduced from 14.8% for GPT-4 through combined RAG and KG grounding -- approaches rule-based reliability while maintaining neural model recall. Evidence traceability of 96.2% (verifiable PubMed citations for extracted relations) provides the auditability essential for biomedical applications. De novo knowledge base construction from 100,000 oncology abstracts achieved 88.4% precision with 5.8% novel expert-verified discoveries. RABES code, BioMistral fine-tuning scripts, PubMed FAISS index, and evaluation datasets are available

at <https://github.com/BAIRU-BioNLP/RABES-2025> under MIT licence. Future work should extend to full-text article extraction, develop multilingual biomedical retrieval for non-English literature, integrate real-time PubMed monitoring for continuous knowledge base updating, and deploy RABES as a clinical decision support tool for drug interaction checking and adverse event signal detection.

References

- Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N. and Liu, X. (2021) Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1), pp. 1-23.
- Labrak, Y., Bazoge, A., Morin, E., Gourraud, P.A., Rouvier, M. and Dufour, R. (2024) BioMistral: a collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*.
- Landhuis, E. (2016) Scientific literature: information overload. *Nature*, 535(7612), pp. 457-458.
- Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S. and So, C.H. (2020) BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4), pp. 1234-1240.
- Nori, H., King, N., McKinney, S.M., Carignan, D. and Horvitz, E. (2023) Capabilities of GPT-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.
- Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J. and Wu, X. (2024) Unifying large language models and knowledge graphs: a roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36(7), pp. 3313-3330.
- Rindflesch, T.C. and Fiszman, M. (2003) The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *Journal of Biomedical Informatics*, 36(6), pp. 462-477.
- Tang, L., Sun, Z., Idnay, B., Nestor, J.G., Soroush, A. and Elhadad, N. (2023) Evaluating large language models on medical evidence summarization. *NPJ Digital Medicine*, 6(1), p. 158.
- Umapathi, L.K., Pal, A., Malaviya, S., Rani, D., Maheshwari, H. and Jyothi, P. (2023) Med-HALT: medical domain hallucination test for large language models. In: *Proceedings of the 27th Conference on Computational Natural Language Learning*, pp. 314-326.
- Wei, C.H., Peng, Y., Leaman, R., Davis, A.P., Mattingly, C.J. and Li, J. (2016) Assessing the state of the art in biomedical relation extraction: overview of the BioCreative V chemical-disease relation (CDR) task. *Database*, 2016, p. baw032.

- Zakka, C., Chaurasia, A., Duber, H.C., Lungren, M.P. and Shah, N. (2024) Almanac: retrieval-augmented language models for clinical medicine. *NEJM AI*, 1(2), pp. 1-12.
- Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S. and Poon, H. (2022) BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6), p. bbac409.
- Peng, Y., Yan, S. and Lu, Z. (2019) Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. In: *Proceedings of the 18th BioNLP Workshop*, pp. 58-65.
- Li, F., Jin, Y., Liu, W., Rawat, B.P.S., Cai, P. and Yu, H. (2020) Fine-tuning bidirectional encoder representations from transformers (BERT)-based models on large-scale electronic health record notes. *JMIR Medical Informatics*, 7(3), p. e14830.
- Herrero-Zazo, M., Segura-Bedmar, I., Martinez, P. and Declerck, T. (2013) The DDI corpus: an annotated corpus with pharmacological substances and drug-drug interactions. *Journal of Biomedical Informatics*, 46(5), pp. 914-920.
- Bravo, A., Pinero, J., Queralt-Rosinach, N., Rautschka, M. and Furlong, L.I. (2015) Extraction of relations between genes and diseases from text and large-scale data analysis. *BMC Bioinformatics*, 16(1), p. 55.
- Gurulingappa, H., Rajput, A.M., Roberts, A., Fluck, J., Hofmann-Apitius, M. and Toldo, L. (2012) Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of Biomedical Informatics*, 45(5), pp. 885-892.
- Pyysalo, S., Ginter, F., Moen, H., Salakoski, T. and Ananiadou, S. (2013) Distributional semantics resources for biomedical text processing. In: *Proceedings of Languages in Biology and Medicine*, pp. 39-44.
- Sung, C., Lin, C.Y. and Ho, C.H. (2024) Biomedical entity linking with retrieval-augmented generation. In: *Proceedings of the 2024 Conference on Empirical Methods in NLP*, pp. 4824-4836.
- Agrawal, M., Hagselmann, S., Lang, H., Kim, Y. and Sontag, D. (2022) Large language models are few-shot clinical information extractors. In: *Proceedings of the 2022 Conference on EMNLP*, pp. 1998-2022.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V. and Goyal, N. (2020) Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, pp. 9459-9474.

Declarations

Funding

This study was supported by the Estonian Research Council grant PRG2348 (A. Horvath), the Swiss National Science Foundation (SNSF) grant 205321_218642 (A. Ivanov), and the Estonian Centre of Excellence in IT (EXCITE, TAR16013) (A. Bianchi). Computational resources provided by the University of Tartu HPC Centre and Swiss National Supercomputing Centre (CSCS). The funding bodies had no role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

A. Horvath is a co-inventor on a pending patent application (EP25182456.4) covering the RABES retrieval-augmented extraction architecture with knowledge graph grounding. A. Ivanov and A. Bianchi declare no competing interests.

Data Availability Statement

RABES code (BioMistral fine-tuning, BioContriever retrieval, KG verification pipeline), pre-trained model weights, PubMed FAISS index construction scripts, and evaluation datasets are available at <https://github.com/BAIRU-BioNLP/RABES-2025> under MIT licence. DDI, ChemProt, and AIMED datasets are publicly available from their respective sources.

Ethical Approval

This computational study used publicly available, de-identified biomedical text corpora and knowledge bases. No human subjects, animal experiments, or personal data were involved. No ethics approval was required.

Appendix A

RABES Architecture and Knowledge Graph Verification Protocol

This appendix provides the complete RABES pipeline specification, BioMistral fine-tuning configuration, dense retrieval index construction, and KG verification logic.

A.1 Pipeline Components

- 1a. Dense retriever: BioContriever (110M params);
FAISS HNSW index over 36M PubMed abstracts;
768-d embeddings; top-20 retrieval per query
- 1b. Entity linker: SapBERT (110M params) fine-tuned
on UMLS 2024AA; links mentions to CUIs with 94.2%
accuracy on MedMentions
- 2a. Generator: BioMistral-7B; LoRA fine-tuning (rank
32, alpha 64) on 480K relation examples; prompt:
input text + retrieved context + CUI annotations
- 2b. Output format: JSON triples {subject_CUI,
predicate, object_CUI, evidence_sentence,
confidence, PMID}
- 3a. KG verification: DrugBank (334K DDI edges),
STRING (11.8M PPI edges), DisGeNET (624K GDA
edges), SIDER (139K ADE edges)
- 3b. Triage logic: match_KG -> 'confirmed'; no_match
+ evidence_PMIID_valid -> 'novel-supported';
no_match + no_evidence -> 'potential_hallucination'

A.2 Fine-Tuning and Inference Configuration

- 4a. Training: 480K examples (BioRED 120K +
ChemProt 80K + DDI 40K + AIMED 30K + GDA 120K
+ ADE 90K); 3 epochs; AdamW lr 2e-5; A100 x4, 18h
- 4b. LoRA: rank 32, alpha 64, dropout 0.05; applied to
q_proj, k_proj, v_proj, o_proj; trainable params 42M /
total 7B (0.6%)
- 5a. Inference: input text -> retrieve 20 PubMed
abstracts (BioContriever, 200ms) -> entity link
(SapBERT, 50ms) -> generate (BioMistral, 800ms) ->
verify (KG lookup, 100ms)
- 5b. Total latency: 1.15 s per abstract; throughput:
3,100 abstracts/hour on single A100; 100K abstracts
processed in 32 hours