Evaluation of Knowledge Graph Construction Methods on the Stroke Domain

Elena Atanasovska[1,2][0009−0004−4123−0379], Boshko Koloski[2][0000−0002−7330−0579], and Dragi Kocev[2][0000−0003−0687−0878]

[1] Faculty of Computer and Information Science, University of Ljubljana, Slovenia
ea58493@student.uni-lj.si
[2] Jožef Stefan Institute and Postgraduate School, Ljubljana, Slovenia
{boshko.koloski, dragi.kocev}@ijs.si

Abstract. Knowledge graphs (KGs) provide a powerful framework for representing complex, interrelated data - an essential capability for navigating complex domains such as medicine. This is particularly true in the study of stroke, one of the leading causes of death and long-term disability worldwide. Currently, there is no comprehensive medical KG for the stroke domain, largely because the specialized knowledge embedded in the vast medical literature makes manual construction prohibitively time-consuming. In this work, we explore the steps towards automated construction of a Stroke KG by comparing four relation-extraction methods and proposing novel *LLM-as-a-judge*-based evaluation. Specifically, we benchmark one unsupervised system (OpenIE), two supervised frameworks (REBEL and ReLiK), and one large-language-model approach (Gemma 2 9B). Our contributions are twofold: first, we provide a systematic evaluation of multiple extraction strategies tailored to a domain-specific KG; second, we propose a hybrid evaluation protocol – combining traditional statistical metrics with an LLM-as-a-judge paradigm – to assess graph quality more comprehensively. The results demonstrate that LLM-based methods hold particular promise for generating a robust, high-coverage Stroke KG, a key resource for accelerating both research and clinical decision-making in stroke care.

Keywords: Knowledge Graph Construction · Relation Extraction · Large Language Models · Biomedical Natural Language Processing.

1 Introduction

Stroke represents a severe and escalating global health crisis. Often described as a “silent pandemic” for its pervasive yet under-recognized impact, it is the second leading cause of death worldwide and one of the primary driver of long-term disability and dementia [5,6,8]. The scale of this burden is projected to grow substantially, with annual mortality expected to increase by 50% to 9.7 million and the global economic toll set to reach $2.3 trillion by 2050 [6,7]. This immense human and economic cost creates an urgent imperative for accelerated research to improve prevention, treatment, and rehabilitation.

This burden is distributed inequitably, revealing a stark paradox. While high-income countries have achieved a 42% reduction in stroke incidence over four decades, rates in low and middle-income countries (LMICs) have more than doubled during the same period [9]. Today, LMICs account for 86% of all stroke-related deaths, with strokes occurring on average 15 years earlier than in wealthier nations [6,9]. This escalating crisis, alongside promising new therapeutic breakthroughs [10], has driven an unprecedented surge in scientific literature. *However, this very explosion of knowledge presents a formidable challenge: the volume is now too vast for systematic human analysis.*

This rapid growth of biomedical literature leads to a profound information paradox: while the sheer volume of published data accelerates the potential for discovery, it simultaneously raises insurmountable barriers to its systematic analysis and synthesis. The entirety of PubMed [11], the primary repository for biomedical literature, is expanding at an exponential rate, with tens of thousands of new articles related to stroke published each year [12]. This deluge of information, scattered across hundreds of journals and presented primarily in unstructured natural language, has grown far beyond the capacity of any human researcher, or even a large team, to comprehensively read, process, and connect. This unstructured format represents a fundamental bottleneck, resulting in fragmented knowledge, duplicated research efforts, and, most critically, missed opportunities to uncover the complex patterns essential for the next breakthrough in stroke medicine.

To overcome this challenge of unstructured information and unlock the collective intelligence embedded in the literature, Knowledge Graphs (KGs) have emerged as a powerful and transformative paradigm [13]. Rooted in the principles of the Semantic Web [2], KGs are structured representations of knowledge, typically build using standards like the Resource Description Framework (RDF). Formally, a KG can be defined as a set of factual triplets $\mathcal{T}$, where each triplet $(s, p, o) \in \mathcal{E} \times \mathcal{R} \times \mathcal{E}$ consists of a subject s, a predicate p, and an object o, with $\mathcal{E}$ being the set of entities and $\mathcal{R}$ the set of relations. Entities serve as the nodes of the graph, representing concepts like diseases (e.g., *Ischemic Stroke*) or drugs (e.g., *Alteplase*), while relations (or predicates) are the directed edges that define the connection, such as `treats` or `causes`.

For example, the factual statement "Aspirin is used to prevent recurrent ischemic stroke" can be distilled into the machine-readable triplet: (Aspirin, prevent, Ischemic Stroke). By systematically extracting millions of such triplets from the literature, we can construct a comprehensive graph that maps the landscape of a domain. KGs are a uniquely suitable solution for this problem because they: 1) enable complex, multi-hop queries that are impossible with standard keyword searches; 2) allow for the discovery of implicit, hidden connections through graph-based algorithms like link prediction; and 3) serve as a foundational, structured backbone for a host of downstream AI applications, including clinical decision support systems, hypothesis generation engines, and sophisticated question-answering platforms [29,31].

In this work, we conduct a comprehensive investigation into building a high-quality StrokeKG. We address the core challenges of corpus creation, relation extraction, and meaningful evaluation. We compare four distinct relation extraction methods and introduce a novel LLM-as-a-judge assessment. **Our main contributions are:** (1) We construct a novel corpus of 433k PubMed abstracts focused on stroke, providing an up-to-date dataset for training and evaluating domain-specific models. (2) We perform a systematic comparison of four relation extraction paradigms (OpenIE [25], REBEL [27], ReLiK [26], and Gemma 2 9B [28] LLM) to build distinct StrokeKGs, analyzing their trade-offs in terms of triplet quantity, quality, and diversity. (3) We propose and implement a novel LLM-as-a-judge evaluation framework with 10 clinically-informed criteria to assess triplet quality, moving beyond traditional metrics to measure factual correctness, relevance, and actionability. (4) Our findings provide a clear roadmap for building a large-scale StrokeKG, with LLMs showing the most promise for generating high-quality, clinically relevant knowledge, thereby paving the way for the next generation of data-driven research in stroke.

2 Related work

The construction of KGs for structuring and unlocking insights from vast amounts of text is an active area of research, particularly in specialized scientific domains [1,3,4]. Recent efforts have produced domain-specific KGs for diverse fields such as food and biomedicine, mapping interactions between chemicals and diseases [14], and framework materials in chemistry [15]. The emergence of Large Language Models (LLMs) has significantly accelerated this trend, with automated pipelines being developed to expand KGs in complex fields like cognitive neuroscience [16]. This body of work underscores a consensus: creating high-quality, specialized KGs is a critical step toward data-driven discovery. However, the methods for evaluating the quality and utility of these graphs remain a significant challenge and a subject of ongoing research [17].

KG evaluation methodologies can be broadly grouped into several paradigms. A traditional approach focuses on intrinsic structural and semantic quality. Structural metrics assess the graph's ontological backbone, quantifying how well the defined schema is utilized by the instance data. For example, Seo et al. [19] proposed metrics like Instantiated Class Ratio (ICR) and Instantiated Property Ratio (IPR) to measure the practical usage of a KG's classes and properties. Semantic metrics, on the other hand, evaluate the logical coherence and correctness of the graph's content, often by comparing it to an external knowledge source. However, a key limitation of this approach is that general-purpose semantic metrics struggle in highly specialized domains, as their performance degrades when the KG's concepts are not well-represented in the reference corpus [20]. Other evaluation paradigms include extrinsic, task-based evaluation, which measures a KG's utility by its performance on downstream tasks like classification or regression [21], and user-centric evaluation, which aligns accuracy assessment with the information needs and query patterns of end-users [22]. While valuable, these methods either do not directly assess the fine-grained factual accuracy of individual relations or require significant task-specific setup.

With the rise of LLMs, research has focused on their application to both KG construction and evaluation. For construction, a central debate revolves around the trade-offs between fine-tuning and prompting. Studies have shown that while few-shot prompting is more adaptable to out-of-domain data, fine-tuning a model on a specific task generally yields higher performance and reduces factual hallucinations and omissions [18]. Other advanced methods explore the use of multi-agent systems, where communicative LLM agents collaborate to build a KG by leveraging both parametric knowledge and web search [24]. Critically, this line of work also highlights that standard text-matching metrics (e.g., BLEU, ROUGE) are often insufficient for KG evaluation, as they fail to handle synonymous expressions and thus underestimate true model performance [18].

This gap in refined, scalable, and domain-aware evaluation has led to a promising new direction: using LLMs themselves as automated “judges”. This “LLM-as-a-judge” or “AutoRater” paradigm leverages an LLM’s world knowledge and instruction-following capabilities to assess the quality of another model’s output based on a set of defined criteria [30]. This paradigm offers the potential for scalable, fine-grained feedback that goes beyond simple structural or string-matching metrics. The viability of this concept is actively being explored, with research investigating whether LLMs can serve as reliable “Graph Judgers” for KG construction tasks [23]. Our work builds directly upon this emerging paradigm: We adapt and steer the LLM-as-a-judge concept for the rigorous demands of the stroke domain, designing a novel evaluation framework with 10 clinically-informed criteria to assess the factual correctness, relevance, and utility of triplets extracted for our StrokeKG.

3 Materials and methods

3.1 Corpus collection methodology

To ground our investigation, we constructed a large-scale, domain-specific text corpus focused on stroke research. The corpus was compiled from PubMed [11], the foremost repository of biomedical literature, ensuring comprehensive coverage and high-quality metadata.

The data was collected by querying the NCBI Entrez Programming Utilities (E-utilities) API. First, we retrieved the complete set of PubMed IDs (PMIDs) matching a single, broad search query: `"stroke"`. Second, for each retrieved PMID, we fetched its full metadata record. Records were retained for further processing only if they contained an abstract, which is essential for our relation extraction task. While full-text articles contain more information, focusing on abstracts is a standard practice in large-scale biomedical text mining, as they provide a high-density summary of key findings and ensure computational tractability across a corpus of this magnitude. For each valid record, we stored the following fields: PMID, title, abstract, author list, author affiliations, journal name, publication date, DOI, and Medical Subject Headings (MeSH) terms.

The corpus contains information about abstracts until April 29, 2025. The initial query returned 488,812 unique PMIDs associated with the term "stroke". Of these, 55,689 records were discarded as they lacked an abstract, yielding a preliminary corpus of 433,123 documents. A final deduplication step, based on abstract text, was performed to remove any identical entries. The final corpus consists of *432,356 unique abstracts*. This comprehensive collection originates from 8,940 distinct journals and includes contributions from over one million unique author names, representing a substantial cross-section of the global stroke research community. This dataset, which we term the **Stroke-PubMed Corpus**, serves as the foundation for all subsequent experiments.

3.2 Methodology

In this work, we conduct a comprehensive comparative study of four relation extraction paradigms to construct the StrokeKG. Our methodology is designed to systematically assess the trade-offs between these methods, focusing on the quantity, diversity, and clinical relevance of the extracted knowledge. The overall workflow consists of three main stages (see Fig. 1): (1) building four distinct knowledge graphs from our Stroke-PubMed Corpus using one rule-based, two pre-trained, and one LLM-based method; (2) performing a statistical analysis of the resulting graphs to characterize their structural properties; and (3) implementing a novel LLM-as-a-judge [23] evaluation framework to assess the semantic quality and clinical validity of the extracted triplets.

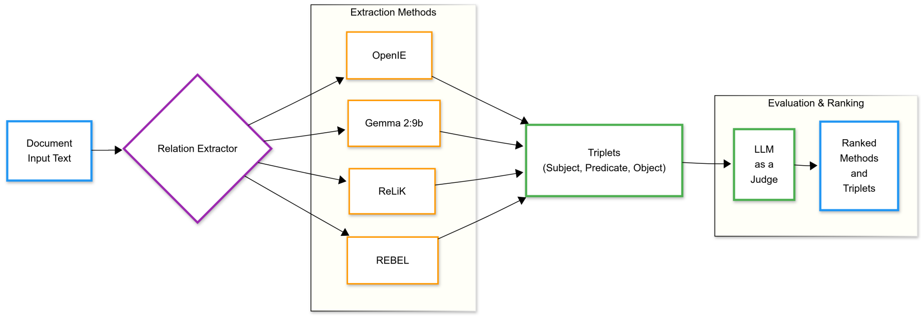

Fig. 1. Proposed methodology: Given a document, we extract its abstract and process it through a set of relation extractors to obtain a collection of triplets, which are then evaluated by a large language model using medically relevant criteria.

Extraction Methods. To ensure a comprehensive evaluation across different paradigms of relation extraction, we selected four representative methods. Our selection was designed to span the spectrum from classic unsupervised techniques to modern generative models. This strategic selection allows for a direct comparison of their inherent strengths and weaknesses in the context of specialized KG construction.

OpenIE [25] represents a traditional, unsupervised, rule-based approach, serving as a baseline that operates without pre-trained domain knowledge. It does not rely on pre-trained data or a fixed schema. Instead, it processes sentences to identify relation triplets based on general linguistic patterns and syntactic dependencies. Its key characteristic is its ability to extract a wide, open-ended set of relations, though often at the cost of precision and semantic consistency.

REBEL [27] and **ReLiK** [26] represent state-of-the-art supervised, pre-trained models. They operate on a fixed schema, allowing us to evaluate the trade-offs of precision and recall in a closed-world assumption. These models are pre-trained on large-scale datasets, designed to perform end-to-end relation extraction and entity linking. Unlike OpenIE, they operate on a predefined, closed set of relation types. REBEL uses a sequence-to-sequence approach to generate triplets directly from text, while ReLiK employs a retrieve-and-link mechanism designed for high accuracy. Their performance is typically characterized by a trade-off between the precision gained from a fixed schema and the recall limited by it.

Gemma 2 9B-parameter model [28] represents the current frontier of few-shot extraction using large language models, allowing us to assess the capabilities of generative AI for this task. It is an LLM that we accessed via a self-hosted Ollama client. The extraction process was designed to be both highly structured and clinically focused, leveraging a combination of detailed prompt engineering and constrained output formatting.
Prompt Engineering. We employed a few-shot prompting strategy to guide the model. The prompt was carefully engineered with four key components:

1. **Role-Playing:** The model was instructed to act as a "stroke medicine expert" to ground its responses in the appropriate clinical context.
2. **Entity and Relation Guidelines:** Clear, explicit rules were provided for the types of entities to extract (e.g., diseases, drugs, biomarkers) and the format of relations (e.g., present tense verbs).
3. **Few-Shot Examples:** A set of high-quality, canonical examples (e.g., `(tissue plasminogen activator, treats, ischemic stroke)`) was included to demonstrate the desired output structure and semantic content.
4. **Quality-Focused Instructions:** The prompt directed the model to prioritize "clinical relevance over completeness" and to only extract factual, evidence-based relationships, thus actively filtering out noise.

Evaluation Framework. Our evaluation framework consists of two sequential stages: a statistical characterization to analyze the structural properties of each method extraction results, followed by a qualitative assessment of the triplets' clinical relevance using an LLM-as-a-judge pipeline.
Statistical Analysis and Redundancy Filtering. First, we characterized the structural properties of each method extraction results, computing metrics for volume (total triplets), diversity (unique entities and relations), and redundancy. This initial analysis revealed that the unsupervised nature of OpenIE resulted in a significantly larger and noisier set of triplets compared to the other methods, with many being syntactically varied but semantically redundant (e.g., `(stroke,`

`causes, disability)` vs. `(disability, is caused by, stroke)`). To ensure a fair comparison, we implemented a similarity-based filtering step exclusively for the OpenIE output before final evaluation.

Clinical Quality Assessment. Following the structural analysis, to evaluate the clinical validity and semantic quality of the extracted facts, we implemented an *LLM-as-a-judge* [23] pipeline using an LLM that was not seen during extraction, specifically *GPT-4o-mini*. The evaluation criteria were developed through a structured meta-prompting process in which a strong LLM itself (*o4-mini-high*) identified key dimensions relevant to assessing clinical utility from the perspective of a stroke specialist. While this approach does not replace expert clinical validation, it offers a scalable proxy for estimating alignment with established medical knowledge. The final framework included 10 distinct criteria, described below, with each triplet rated on a 1 (poor) to 5 (excellent) scale:

LLM-as-a-Judge Evaluation Criteria

1. **Clinical relevance ($w_1 = 3$):** Does the statement address important clinical aspects of stroke diagnosis, treatment, prognosis, or prevention?
2. **Evidence strength ($w_2 = 2.5$):** Is the statement supported by strong clinical evidence, such as randomized controlled trials, meta-analyses, or large cohort studies?
3. **Specificity ($w_3 = 1$):** How specific is the statement to stroke subtype (ischemic, hemorrhagic, SAH) or patient population (age, gender, comorbidities)?
4. **Guideline concordance ($w_4 = 2.5$):** Is the statement consistent with current stroke guidelines (e.g., AHA/ASA, ESO)?
5. **Pathophysiological accuracy ($w_5 = 1$):** Does the statement accurately describe stroke pathophysiology or mechanisms (e.g., thrombotic occlusion, vasospasm)?
6. **Diagnostic utility ($w_6 = 1.5$):** Does the statement provide useful information regarding stroke diagnosis or differentiation from mimics?
7. **Therapeutic implications ($w_7 = 3$):** Does it mention interventions or management strategies with proven impact on outcomes?
8. **Prognostic value ($w_8 = 1.5$):** Does it discuss factors that meaningfully affect stroke prognosis or recovery?
9. **Population impact ($w_9 = 1$):** How broadly applicable is the statement across populations or settings?
10. **Potential for harm (inverted, $w_{10} = 3$):** Does the statement avoid misleading or harmful implications that could affect patient care?

To create a composite quality score (CQS) for each triplet, we computed a weighted sum of the 10 criteria scores (s_i). The weights (w_i) were determined by a strong LLM acting as a domain expert (*gpt-o1-mini-high*) to reflect relative clinical importance. The score for "Potential for Harm" (s_{10}) was inverted to match the direction of the other criteria. The final CQS ranges from 20 to 100 and is calculated as: CQS $= \left(\sum_{i=1}^{9} w_i \cdot s_i\right) + w_{10} \cdot (6 - s_{10})$. This aggregation enables a nuanced comparison of the methods based on their ability to generate clinically valid and useful knowledge.

4 Results

The results presented in this paper are based on a representative 10,000-abstract subset of our full corpus, a scale necessitated by current computational constraints. To ensure the validity of our findings, this subset was strategically sampled to mirror the word-count distribution of the complete 433k-abstract dataset, thereby preserving a comparable average abstract length and variability.

Our evaluation of this subset reveals a clear performance hierarchy and distinct trade-offs across the four extraction paradigms. The aggregate results, summarized in Table 1, show that while the rule-based OpenIE method is the most prolific in terms of triplet volume, the LLM-based Gemma 2 9B model achieves a substantially higher clinical quality score. A deeper analysis, broken down by our 10 clinical criteria, highlights the specific strengths and weaknesses of each approach.

Table 1. Comparison of Relation Extraction Methods on Stroke Abstracts

Method	**Total Triplets**	**Unique Triplets**	**Selectivity Rate (%)**	**Unique Relations**	**Unique Entities**	**Dup. Rate**	**Avg. LLM Score**
OpenIE	327,404	313,898	0.0	52,174	186,053	4.1%	42.77
REBEL	142,760	125,738	0.0	169	75,807	11.9%	43.21
ReLiK	51,426	39,056	6.6	193	23,700	24.1%	50.92
Gemma 2 9B	72,591	70,516	13.9	13,095	79,021	2.9%	57.39

**Duplication Rate = (Total Triplets − Unique Triplets)/Total Triplets[%].*
***Average LLM Score is the mean CQS across all triplets from a method.*
****Average LLM Score for OpenIE is calculated on the filtered set to ensure fair comparison; all other metrics for all methods are on the raw outputs.*
*****Note: The Avg. LLM Score is calculated using the full-precision source data. The scores in Table 2 are rounded for presentation, which may cause minor disagreements if used for recalculation.*

The Gemma 2 9B model stands out as the top performer, achieving the highest average Composite Quality Score (CQS) of 57.39. As shown in Table 1, it also had the highest selectivity rate at 13.9%. This high rate is not a weakness but a key strength of the prompting-based approach: the prompt explicitly instructed the model to extract only medically relevant information. Therefore, these "omissions" represent an active, intelligent filtering of irrelevant text at the point of extraction, a capability the other methods lack. The detailed criteria breakdown in Table 2 further reveals why the extracted triplets are superior: Gemma consistently scores highest on the most heavily weighted criteria, including Clinical Relevance (3.60), Pathophysiological Accuracy (3.36), and Potential for Harm (Inverted) (3.60). This indicates that the LLM is demonstrating a nuanced understanding of the clinical context, prioritizing quality and relevance over sheer volume.

At the other end of the spectrum, OpenIE exemplifies a classic quantity-over-quality trade-off. While it generated a massive volume of triplets (over 327,000) with zero selectivity rate, its average CQS was the lowest (42.77). Table 2 shows a systemic failure across all criteria, with particularly low scores in Evidence Strength (1.72) and Guideline Concordance (1.62). This suggests that its purely syntactic, pattern-based approach is insufficient for a medical domain, producing a high volume of factually shallow or clinically irrelevant statements that would add significant noise to a knowledge graph.

The pre-trained, closed-set methods, ReLiK and REBEL, occupy a middle ground and illustrate a distinct precision-recall trade-off. ReLiK achieved a respectable CQS of 50.92, outperforming REBEL and OpenIE. Its strength lies in its higher precision, scoring better across nearly all criteria compared to REBEL. This comes at the cost of a 6.6% selectivity rate, though this filtering appears less "intelligent" than the LLM's, but a lot faster than it, likely resulting from an inability to match its fixed schema rather than an assessment of relevance. Conversely, REBEL achieved 100% coverage but with a much lower CQS (43.21), comparable to that of OpenIE. Both models are severely limited by their small, predefined sets of relation types (193 for ReLiK, 169 for REBEL), preventing them from capturing the rich diversity of information present in the literature.

Table 2. Average LLM Scores by Criterion per Method.

Criterion	**Gemma 2 9B**	**OpenIE**	**ReLiK**	**REBEL**	**Weight**
Clinical Relevance (s_1)	3.60	2.53	2.87	2.23	3.0
Evidence Strength (s_2)	3.05	1.72	2.47	1.82	2.5
Specificity (s_3)	2.79	1.76	2.32	1.78	1.0
Guideline Concordance (s_4)	2.79	1.62	2.43	1.81	2.5
Pathophysiological Accuracy (s_5)	3.36	1.81	2.65	2.01	1.0
Diagnostic Utility (s_6)	2.48	1.63	2.18	1.71	1.5
Therapeutic Implications (s_7)	2.47	1.56	2.15	1.72	3.0
Prognostic Value (s_8)	3.00	1.83	2.37	1.87	1.5
Population Impact (s_9)	3.02	1.87	2.47	1.89	1.0
Potential for Harm (s_{10}, Inverted)	3.60	2.16	2.89	2.25	3.0
Average Score	**3.02**	**1.85**	**2.48**	**1.91**	–

The distinct clinical utility and methodological trade-offs of each extraction approach are underscored by the representative triplets shown in Table 3. The pre-trained models, ReLiK and REBEL, excel at extracting granular, factual knowledge. They are shown to capture established clinical relationships with high precision, accurately identifying specific interventions for their indications or linking guideline-recommended therapies to the conditions they treat. This demonstrates their strength in mapping information to a fixed schema.

In contrast, OpenIE's output illustrates its classic quantity-over-quality trade-off. As seen in the table, it can capture both fundamental, high-value clinical facts as well as more generalized statements that, while correct, lack the specificity of the schema-based models. Finally, the LLM-based Gemma model showcases a higher level of clinical synthesis. It moves beyond simple facts to extract nuanced, high-level insights about risk factors and preventative strategies, reflecting a deeper contextual understanding. Collectively, these examples highlight how different methods can surface complementary aspects of stroke knowledge, from foundational facts to actionable clinical insights.

Table 3. Representative triplets extracted by each method.

Subject	**Predicate**	**Object**	**Method**	**Score**
Mechanical thrombectomy	has use	large vessel occlusion	ReLiK	100.0
Clopidogrel	medical condition treated	acute ischaemic stroke	REBEL	98.5
ischemic stroke	subclass of	stroke	REBEL	100.0
Reperfusion therapy	is treatment for	acute ischemic stroke	OpenIE	100.0
drug	was administered within	3 hours	OpenIE	100.0
Oral anticoagulation	reduces the risk of	stroke	Gemma	100.0
hypertension	increases	stroke risk	Gemma	98.0
Atrial fibrillation	causes	cardioembolic stroke	Gemma	100.0

5 Conclusion

In this work, we addressed the critical challenge of constructing a specialized, high-quality Knowledge Graph for the stroke domain from the vast and unstructured biomedical literature. We conducted a systematic comparison of four distinct relation extraction paradigms—rule-based, pre-trained, and large language model-based—and introduced a novel, clinically-informed LLM-as-a-judge framework to move beyond traditional evaluation metrics.

Our findings provide a clear performance hierarchy among the tested methods. The prompting-based LLM approach (Gemma 2 9B) proved to be clearly superior, not just in generating the highest quality triplets (with average CQS: 57.39), but also in demonstrating an ability to intelligently filter for medical relevance. In stark contrast, the rule-based OpenIE system, despite its massive output, produced predominantly low-quality, noisy information unsuitable for a clinical KG. The pre-trained models, ReLiK and REBEL, occupied a middle ground, highlighting a classic precision-recall trade-off but were ultimately constrained by their fixed, limited schemas. This demonstrates that for complex, specialized domains, the contextual understanding and flexible, open-schema extraction capabilities of LLMs are paramount.

The primary limitation of our current study is the use of a 10,000-abstract subset, necessitated by computational constraints. While this subset was strategically sampled to be representative, a full-scale analysis is a crucial next step. Furthermore, while our LLM-as-a-judge framework offers a powerful, scalable proxy for evaluation, it cannot replace validation by human clinical experts. In addition, the computational effort required to evaluate triplets grows linearly with the number of extracted triplets and can become substantial when exhaustive extraction methods are used.

Future work will proceed along three main avenues. First, we will scale our extraction pipelines to cover the entire 433K–document Stroke-PubMed corpus. Second, instead of relying on a single "best" method, we will develop a hybrid-ensemble strategy to build the first large-scale *StrokeKG*. This will leverage each paradigm's strengths—using the high-precision triplets from Gemma and ReLiK as a reliable core, while cautiously integrating OpenIE's broader (but noisier) outputs after rigorous automated quality filtering. Third, to avoid the expense of manual evaluation, we will train surrogate models on the 10k examples already evaluated, then use these fast, low-cost scorers to annotate the remaining documents.

Acknowledgements. This work was supported by HE TRUSTroke project funded by the EU HORIZON-HLTH-2022-STAYHLTH-01-two-stage under grant agreement No 101080564, as well as by the Projects P2-0103, GC-0001 and GC-002 by the Slovene Research and Innovation Agency, and a Young Researcher Grant PR-12394 (BK).

References

1. Abu-Salih, B. (2021). Domain-specific knowledge graphs: A survey. arXiv. https://arxiv.org/abs/2011.00235
2. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. *Scientific American* **284**(5), 34–43 (2001).
3. Zhong, L., Wu, J., Li, Q., Peng, H., & Wu, X. (2024). A comprehensive survey on automatic knowledge graph construction. ACM Computing Surveys, 56(4), Article 94, 1–62. https://doi.org/10.1145/3618295
4. Nicholson, D. N., & Greene, C. S. (2020). Constructing knowledge graphs and their biomedical applications. Computational and Structural Biotechnology Journal, 18, 1414–1428.
5. Owolabi, M.O., Thrift, A.G., Martins, S., Johnson, W., Pandian, J., Abd-Allah, F., Varghese, C., et al.: The state of stroke services across the globe: Report of World Stroke Organization–World Health Organization surveys. *International Journal of Stroke* **16**(8), 889–901 (2021).
6. Owolabi, M.O., et al.: Primary stroke prevention worldwide: translating evidence into action. *The Lancet Public Health* **7**(1), e74–e85 (2022).
7. Stroke deaths could jump 50% by 2050, study warns. *Axios*, 10 October 2023. https://www.axios.com/2023/10/10/stroke-deaths-could-jump-50-by-2050-study-warns
8. Collins, L.: Stroke predicted to kill 10 million a year by 2050. *Deseret News*, 12 October 2023. https://www.deseret.com/2023/10/12/23914309/stroke-kills-10-million-a-year-by-2050-study-healthy-science

9. Feigin, V.L., Roth, G.A., Naghavi, M., Parmar, P., Krishnamurthi, R., Chugh, S., Mensah, G.A., Norrving, B., Shiue, I., Ng, M., Estep, K., Cercy, K., Murray, C.J.L., Forouzanfar, M.H.; Global Burden of Diseases, Injuries and Risk Factors Study 2013 and Stroke Experts Writing Group: Global burden of stroke and risk factors in 188 countries, during 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet Neurol.* **15**(9), 913–924 (2016).
10. UCLA Health: UCLA discovers first stroke rehabilitation drug to repair brain damage in mice. UCLA Health News, March 18, 2025. `https://www.uclahealth.org/news/release/ucla-discovers-first-stroke-rehabilitation-drug-repair-brain`
11. National Library of Medicine. *PubMed*. National Institutes of Health (NIH). Available at: `https://pubmed.ncbi.nlm.nih.gov`
12. National Center for Biotechnology Information: PubMed Search for "stroke." U.S. National Library of Medicine, National Institutes of Health. `https://pubmed.ncbi.nlm.nih.gov/?term=stroke`
13. Hogan, A., Blomqvist, E., Cochez, M., D'amato, C., De Melo, G., Gutierrez, C., Kirrane, S., Labra Gayo, J.E., Navigli, R., Neumaier, S., Ngonga Ngomo, A.-C., Polleres, A., Rashid, S.M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., and Zimmermann, A. (2021). Knowledge Graphs. ACM Comput. Surv. 54, 4, Article 71 (June 2021), 37 pages.
14. Cenikj, G., Strojnik, L., Angelski, R., Ogrinc, N., Koroušić Seljak, B., Eftimov, T.: From language models to large-scale food and biomedical knowledge graphs. *Scientific Reports* **13**(1), 7815 (2023).
15. Langer, S., Neuhaus, F., Nürnberger, A.: CEAR: Automatic construction of a knowledge graph of chemical entities and roles from scientific literature. arXiv preprint arXiv:2407.21708 (2024).
16. Sarabadani, A., Rahsepar Fard, K., Dalvand, H.: ExKG-LLM: Leveraging Large Language Models for Automated Expansion of Cognitive Neuroscience Knowledge Graphs. arXiv preprint arXiv:2503.06479 (2025).
17. Choi, S., Jung, Y.: Knowledge Graph Construction: Extraction, Learning, and Evaluation. Applied Sciences **15**(7), 3727 (2025).
18. Ghanem, H., Cruz, C.: Fine-Tuning vs. Prompting: Evaluating the Knowledge Graph Construction with LLMs. In: Plexousakis, D., Troncy, R., Ristoski, P., et al. (eds.) Proceedings of the Workshops at the Extended Semantic Web Conference 2024. CEUR Workshop Proceedings, vol. 3747. CEUR-WS.org (2024).
19. Seo, S., Cheon, H., Kim, H., Hyun, D.: Structural Quality Metrics to Evaluate Knowledge Graphs. arXiv preprint arXiv:2211.10011 (2022)
20. Fernández, M.-F., Gómez-Pérez, A., Stama, J.: Dynamic knowledge graph evaluation. Semantic Web. 12(6), 1055–1072 (2021).
21. Heist, N., Hertling, S., Paulheim, H.: KGrEaT: A Framework to Evaluate Knowledge Graphs via Downstream Tasks. arXiv preprint arXiv:2308.10537 (2023).
22. Zhang, Y., Xiao, G.: A novel customizing knowledge graph evaluation method for incorporating user needs. Scientific Reports **14**(1), 9594 (2024).
23. Huang, H., Chen, C., Sheng, Z., Li, Y., Zhang, W.: Can LLMs be Good Graph Judge for Knowledge Graph Construction?. arXiv preprint arXiv:2411.17388 (2024)
24. Zhu, Y., Wang, X., Chen, J., Qiao, S., Ou, Y., Yao, Y., Deng, S., Chen, H., Zhang, N.: LLMs for Knowledge Graph Construction and Reasoning: Recent Capabilities and Future Opportunities. arXiv preprint arXiv:2305.13168 (2024).
25. O. Etzioni, M. Banko, S. Soderland, and D. Weld. Open information extraction from the web. *Communications of the ACM*, 51(12):68–74, 2008.

26. R. Orlando, P.-L. Huguet Cabot, E. Barba, and R. Navigli. ReLiK: Retrieve and LinK, Fast and Accurate Entity Linking and Relation Extraction on an Academic Budget. *arXiv preprint arXiv:2408.00103*, 2024.
27. P.-L. Huguet Cabot and R. Navigli. REBEL: Relation Extraction By End-to-end Language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2370–2381, Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics.
28. Gemma Team: M. Riviere, S. Pathak, P.G. Sessa, C. Hardin, S. Bhupatiraju, L. Hussenot, T. Mesnard, et al. Gemma 2: Improving Open Language Models at a Practical Size. *arXiv preprint arXiv:2408.00118*, 2024.
29. X. Zou, "A Survey on Application of Knowledge Graph," *Journal of Physics: Conference Series*, vol. 1487, no. 1, p. 012016, Mar. 2020. doi: 10.1088/1742-6596/1487/1/012016.
30. Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. From Generation to Judgment: Opportunities and Challenges of LLM-as-a-judge. *arXiv preprint arXiv:2411.16594*, 2025. URL: `https://arxiv.org/abs/2411.16594`.
31. Blaž Škrlj, Boshko Koloski, Senja Pollak, and Nada Lavrač. From Symbolic to Neural and Back: Exploring Knowledge Graph-Large Language Model Synergies. *arXiv preprint arXiv:2506.09566*, 2025. `https://arxiv.org/abs/2506.09566`