Responsible NLP Checklist

Paper title: *ATOM: AdapTive and OptiMized dynamic temporal knowledge graph construction using LLMs*
Authors: *Yassir Lairgi, Ludovic Moncla, Khalid Benabdeslem, Rmy Cazabet, Pierre Clau*

> How to read the checklist symbols:
>
> ☑ the authors responded ‘yes’
>
> ☒ the authors responded ‘no’
>
> N/A the authors indicated that the question does not apply to their work
>
> □ the authors did not respond to the checkbox question
>
> For background on the checklist and guidance provided to the authors, see the Responsible NLP Checklist page at ACL Rolling Review.

☑ A. Questions mandatory for all submissions.

☑ A1. Did you describe the limitations of your work?
This paper has a Limitations section.

☒ A2. Did you discuss any potential risks of your work?
technical limitations like hallucinations and error propagation are discussed in the dedicated Limitations section (Section 6), potential broader societal risks or ethical harms were not discussed as they were not deemed applicable to this specific contribution.

☑ B. Did you use or create scientific artifacts? (e.g. code, datasets, models)

☑ B1. Did you cite the creators of artifacts you used?
Section 4.2 cites the creator of the NYT News dynamic and temporal dataset (Singh, 2023). The paper also cites baseline methods and tools such as iText2KG, Graphiti, and DocRED throughout Section 2 and Section 4.

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
Table T.1 caption discusses the licensing restrictions of the original NYT articles, explaining that only lead paragraphs are used/released to comply with these terms. Section 1 mentions the code and dataset are available at a GitHub repository.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 4.2 discusses the suitability of the NYT News dataset for analyzing temporal dynamics, evolving situations, and real-world scenarios, which is consistent with the nature of a news archive and the intended research goals.

☒ B4. Did you discuss the steps taken to check whether the data that was collected/used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect/anonymize it?
The dataset consists of public news articles (NYT) focusing on public events and figures (specifically COVID-19 dynamics in 2020), rather than private individual data requiring anonymization steps.

The Responsible NLP Checklist used at ACL Rolling Review is adopted from NAACL 2022, with the addition of ACL 2023 question on AI writing assistance and further refinements based on ARR practice.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Table T.1 provides comprehensive statistics for the created 2020-COVID-NYT dataset. Section 4.2 and Appendix E describe the dataset construction process, the temporal range (2020), and the annotation methodology.

☑ B6. Did you report relevant statistics like the number of examples, details of train/test/dev splits, etc. for the data that you used/created?
Table T.1 reports detailed statistics including the total number of articles, grouped articles by publication date, total atomic facts, the number of 5-tuples, and token counts.

☑ **C. Did you run computational experiments?**

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 4.3 and Section 4.6 list the specific LLM versions used (e.g., gpt-4.1-2025-04-14, claude-sonnet-4-2025-01-31). Table T.2 details the total input/output token usage and the estimated cost in USD for the experiments.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4.2 specifies the temperature setting (0) to minimize stochasticity. Appendix C details the estimation and specific values used for merging thresholds (theta values) and similarity weights.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Tables 1 and 2 report results using mean values and standard deviations (e.g., 0.552 +/- 0.124) to represent the stability and quality across multiple experimental runs.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation, such as NLTK, SpaCy, ROUGE, etc.), did you report the implementation, model, and parameter settings used?
Section 4.2 and Appendix C specify the use of specific models via API, including text-embedding-3-large for embeddings and specific versions of GPT and Claude models.

☑ **D. Did you use human annotators (e.g., crowdworkers) or research with human subjects?**

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix E (Sections E.1 and E.2) reports the specific guidelines and rules provided to the human annotators for verifying atomic fact decomposition and 5-tuple extraction (e.g., rules on Atomicity, Decontextualization, and Temporal Normalization).

☒ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Annotation was performed by the research team members as part of dataset curation.

☒ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating (e.g., did your instructions explain how the data would be used)?
Dataset is coming from NYT dataset publicly available and curation has been done by research team members.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
approval was not required as: (1) we used publicly available news articles, not human subjects data, (2) annotation was performed by research team members, not vulnerable populations, (3) no sensitive or personal information beyond what is already public was collected

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No, we did not. No bias is reconducted in the annotation

☑ **E. Did you use AI assistants (e.g., ChatGPT, Copilot) in your research, coding, or writing?**

☑ E1. If you used AI assistants, did you include information about their use?
Section7 illustrate the use of AI assistants in the paper. Moreover, appendix E explicitly states that the ground truth for the dataset was constructed through a semi-manual annotation process utilizing the Claude Sonnet 4 chat model.