Summary Of Strengths:

1. This paper contributes a new benchmark for event-level knowledge editing, which can be helpful for the development of the community,
2. This work proposes to test the future trends after knowledge editing (influences), which has been ignored in previous studies.

Summary Of Weaknesses:

1. The motivation is somewhat not convincing enough. Both triplet-level and event-level knowledge editing face inefficiency and incompleteness issues. As shown in Fig. 1, the factual knowledge (...live?; ... work in?) and tendency (... immigration policy?) are also not directly presented in the given event, facing the same problem as the given triplet. It seems that events and triplets are knowledge in different grains. I notice there is a work studying document grain [1], which is also related to this work.
2. I appreciate the efforts of building the benchmark, but more discussions about under-exploited challenges in this direction are needed. It would be helpful for other researchers in the future. Besides, it would be better to provide specific baseline methods built for event-level knowledge editing.

[1] Eva-KELLM: A New Benchmark for Evaluating Knowledge Editing of LLMs

Summary Of Strengths:

- The main advantage of this paper is that it provides a new resource to study event-level knowledge editing. I believe the community will benefit from the research in this paper, and more researchers will further investigate based on this work.
- The paper evaluates multiple baseline methods and various language models, and the experimental section is generally comprehensive. I appreciate the authors' detailed analysis of each experiment in this paper.

Summary Of Weaknesses:

Adding more details on dataset construction and quality evaluation will help improve the overall quality of this dataset. Considering that GPT-3.5 was used to generate various types of events during dataset construction, concerns about the dataset quality are inevitable. It is well known that GPT models have certain degrees of hallucination and flattery, so the lack of secondary verification and comprehensive quality evaluation will raise concerns among many researchers about the quality of the dataset proposed in this paper.

Summary Of Strengths:

1. The paper raises a general awareness over conventional factual triplet-level knowledge editing on LLMs and suggests an enlarged event-level view on KE.
2. While I have some doubts regarding the metrics of the proposed KE benchmark, ELKEN, the motivation of dataset construction is valued and the limitations of ELKEN are clearly stated.

Summary Of Weaknesses:

1. There is an observable gap between event-level KE concept and actual benchmark.
   The authors claim that “updating all implicated facts at once” (Line 073, 154) is one of two major remedies provided by event-level KE in comparison with triplet-level KE, yet the “manually identified impact scope” (Line 220) in the constructed benchmark seem very unlikely to comprehensively measure whether “all implicated facts” are correctly edited. Could the authors provide more details regarding how they formulate the scopes of editing?
   The similar issue resides in the “tendency knowledge” aspect. I find it is hard to be conceived that “6 in-scope and 2 out-of-scope question-answer pairs” (Lines 280-281) generated by GPT-4 exhaust tendency possibilities.
2. The comparison with existing KE benchmarks lacks sufficient details.
   When providing a novel benchmark, a comparison with existing ones is substantial to help readers better recognize the characteristics of the benchmark in terms of scale, question composition, topic domain, etc. Section 5.1 lacks necessary details.
3. The benchmark metrics have a heavy reliance on GPT-4 judgment.
   The comparability of results is crucial for benchmark, yet ELKEN’s metrics “reliability” and “locality” rely heavily on GPT-4 for scoring (Lines 324, 335). Since the GPT-4 model family is subject to constant update, posing a significant worry on the comparability of the proposed benchmark. Additionally, “locality” metric echoes the aforementioned doubts regarding the impact scope of editing. Nonetheless, as the paper reports a positive correlation between GPT and human scoring (Lines 561-564), this undermining may not be as severe as it seems.
4. No mention regarding prior knowledge.
   The paper discusses post-editing “unknown” scenario, but pre-editing “unknown” scenario is equally important. Different language models are likely to have different set of prior knowledge based on their training corpora selection. To my understanding, if we were to examine if a piece of “knowledge” or “memory” in an LLM is successfully edited, we would have to check if there exists such piece of “prior knowledge”. Could the authors explain whether and how “reliability” metric address “prior knowledge”?