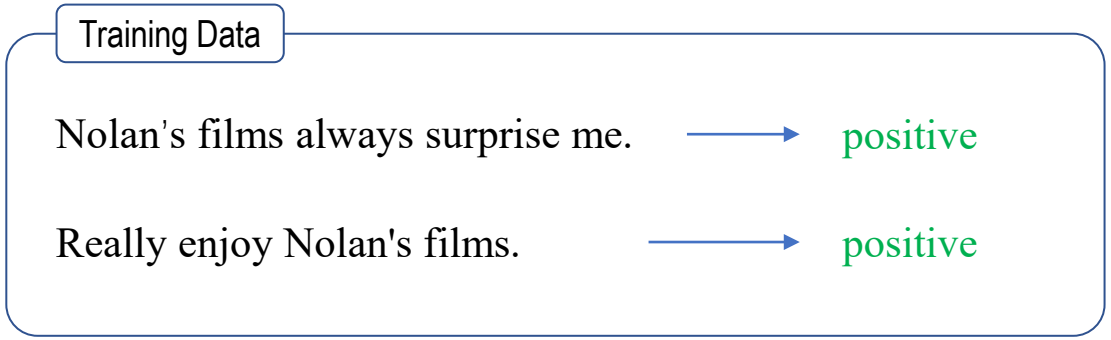
Prompting Large Language Models for Counterfactual Generation: An Empirical Study

Yongqi Li [1 *], Mayi Xu [1 *], Xin Miao [1 *], Shen Zhou [1 *], and Tieyun Qian [1,2 †]
1 School of Computer Science, Wuhan University, Wuhan, China
2 Intellectual Computing Laboratory for Cultural Heritage, Wuhan University, Wuhan, China

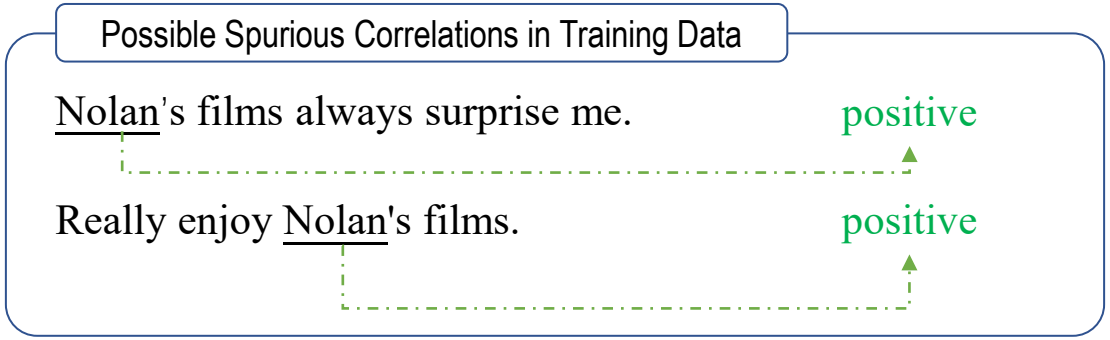
Presenter: Yongqi Li

liyongqi@whu.edu.cn

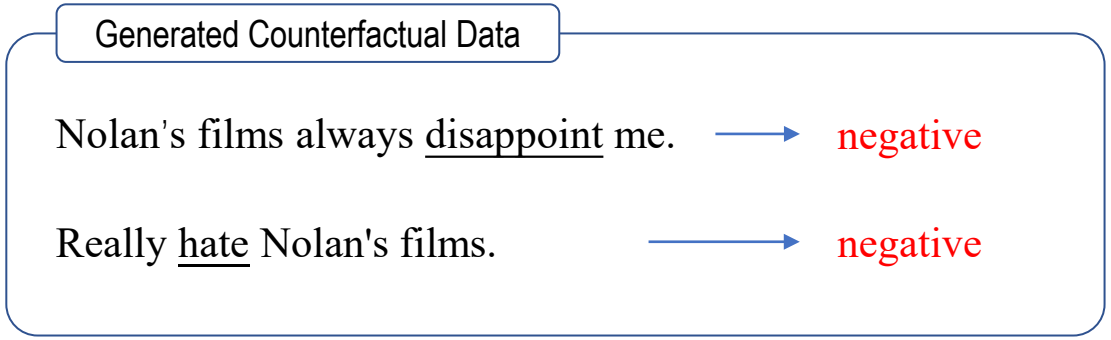
1.1 What is counterfactual generation and what is it used for?



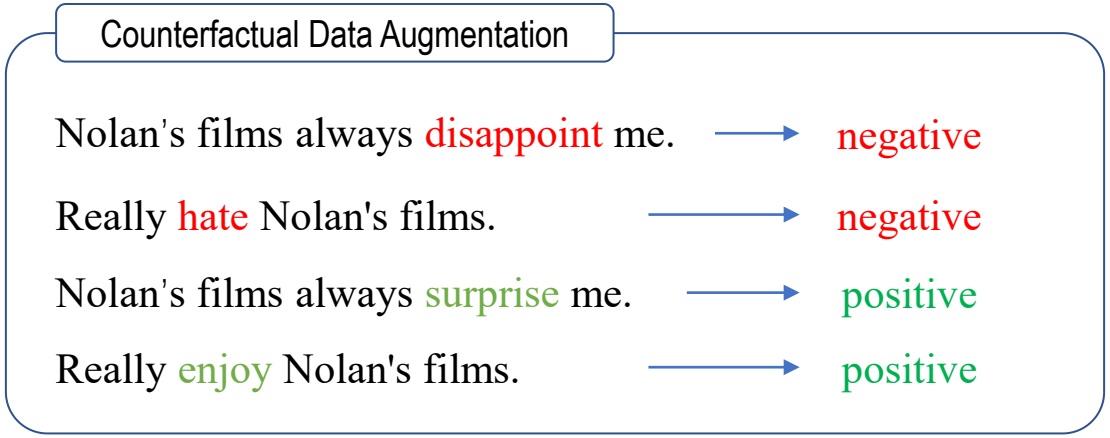
(a) Original training data.



(b) Spurious correlations due to the frequent occurrence of certain sentiment-irrelevant words, which we do not expect.



(c) Generated counterfactual data.



(d) We can mix the counterfactual data with the original data, which can make the model focus on sentiment-relevant words.

1.2 Causal theory foundation

- We can analyze the spurious correlations in data using the Structural Causal Model (SCM).
- The spurious correlations ($X_T \leftarrow C \rightarrow Y$) can be due to the annotation bias. *For example, we only collect the annotated data from Nolan's fans*.
- Counterfactual generation and augmentation are used for the intervention operation in Figure 1(b), which can "cut" the spurious correlations.

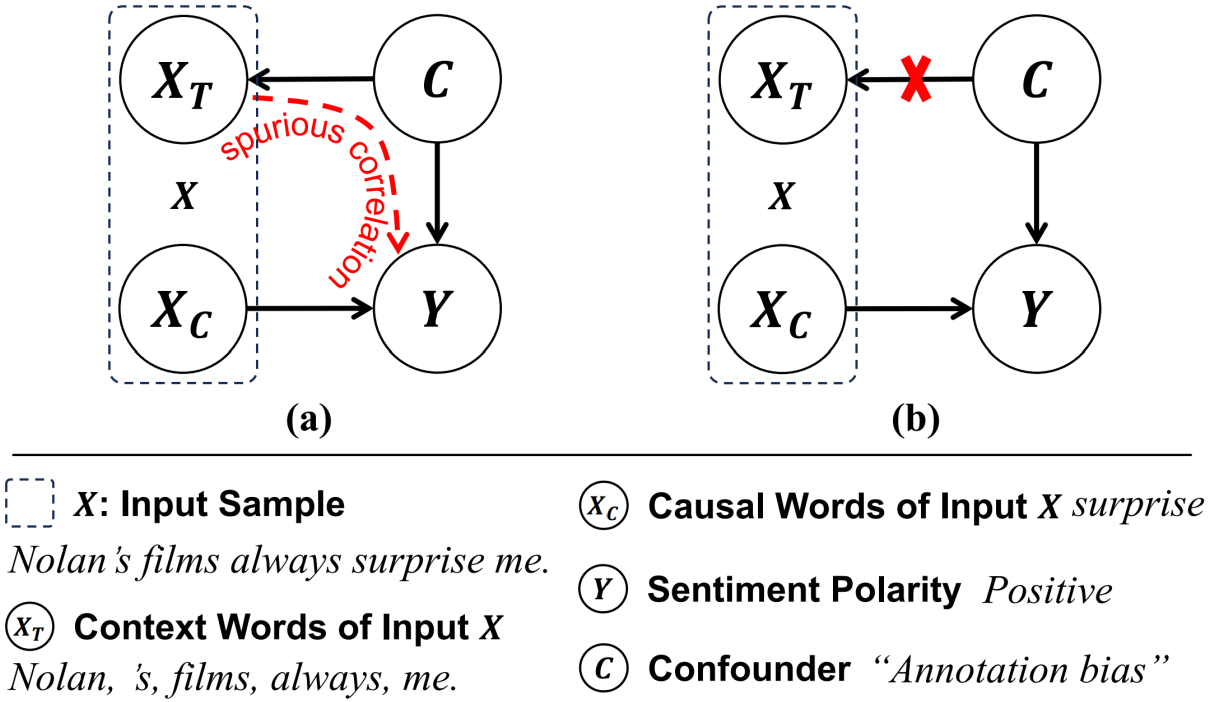


Figure 1: (a) Structural causal model of the Sentiment Analysis (SA) task, (b) Counterfactual generation is used for the intervention operation to X_T to eliminate spurious correlations in the data.

2.1 Evaluation Framework

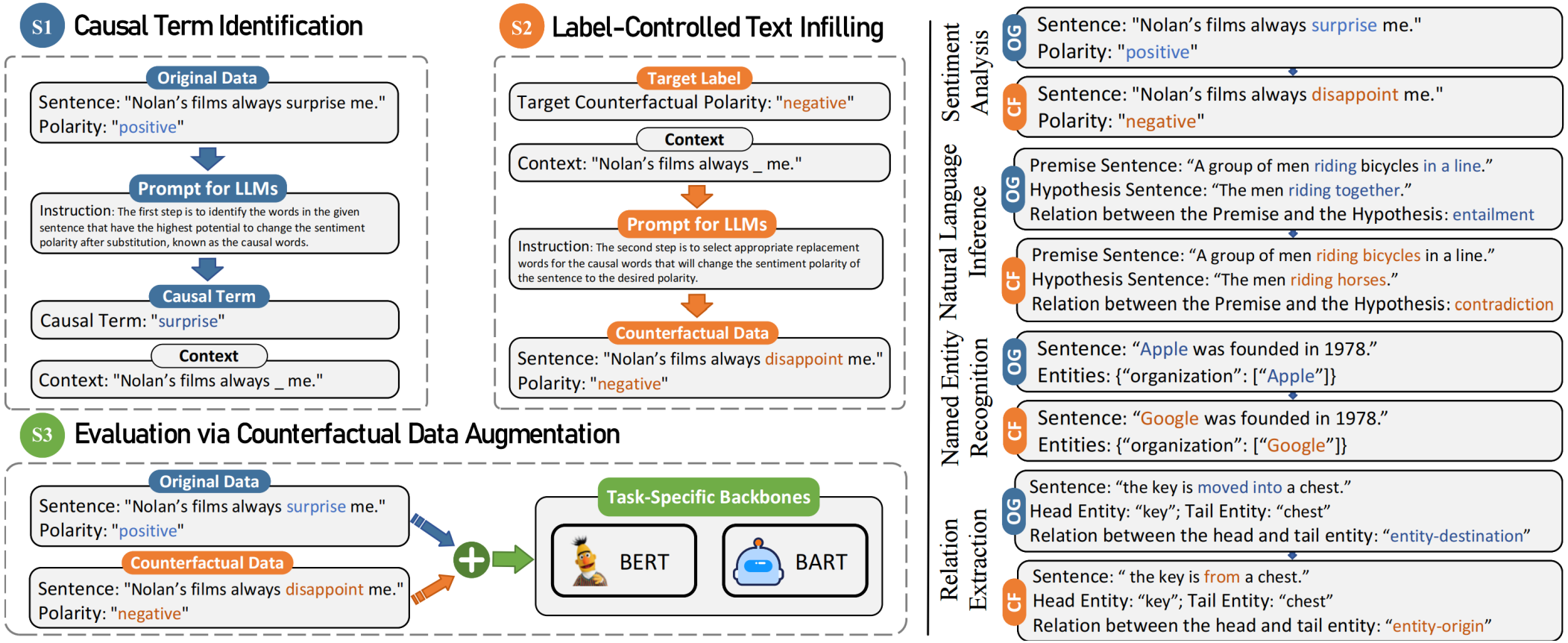


Figure 2: Left: The proposed framework for evaluating counterfactuals generated by LLMs (SA task). Right: Original (OG) samples and generated counterfactual (CF) samples on **SA, NLI, NER and RE** tasks.

3.1 Evaluation Settings

Datasets

- SST-2 and IDMB for the SA task
- SNLI and MNLI for the NLI task
- CoNLL2003 and OntoNotesV5 for the NER task
- SemEval2010 and TACRED for the RE task

Few-shot Settings

- Spurious correlations are particularly prevalent in few-shot settings.
- We conduct experiments using randomly sampled {5,10,20,50}shot training set on each dataset.

Compared Methods

- LLMs: The performance of LLMs themselves.
- SLMs (Original): The original few-shot performance of SLMs via the BERT-based or BART-based fine-tuning methods.
- SLMs (Internal knowledge augmented): The augmented SLMs with counterfactual data generated by internal knowledge tailored methods.
- SLMs (LLMs augmented): The augmented SLMs with counterfactual data generated by LLMs.

3.2 Strengths of LLMs for counterfactual generation

Observations

- The counterfactual samples generated by LLM on the **SA, NLI, and NER** tasks all improve the performance of SLMs.

- In most cases, LLM-generated counterfactuals can improve the performance of SLMs more **than well-designed internal counterfactual generation methods**.

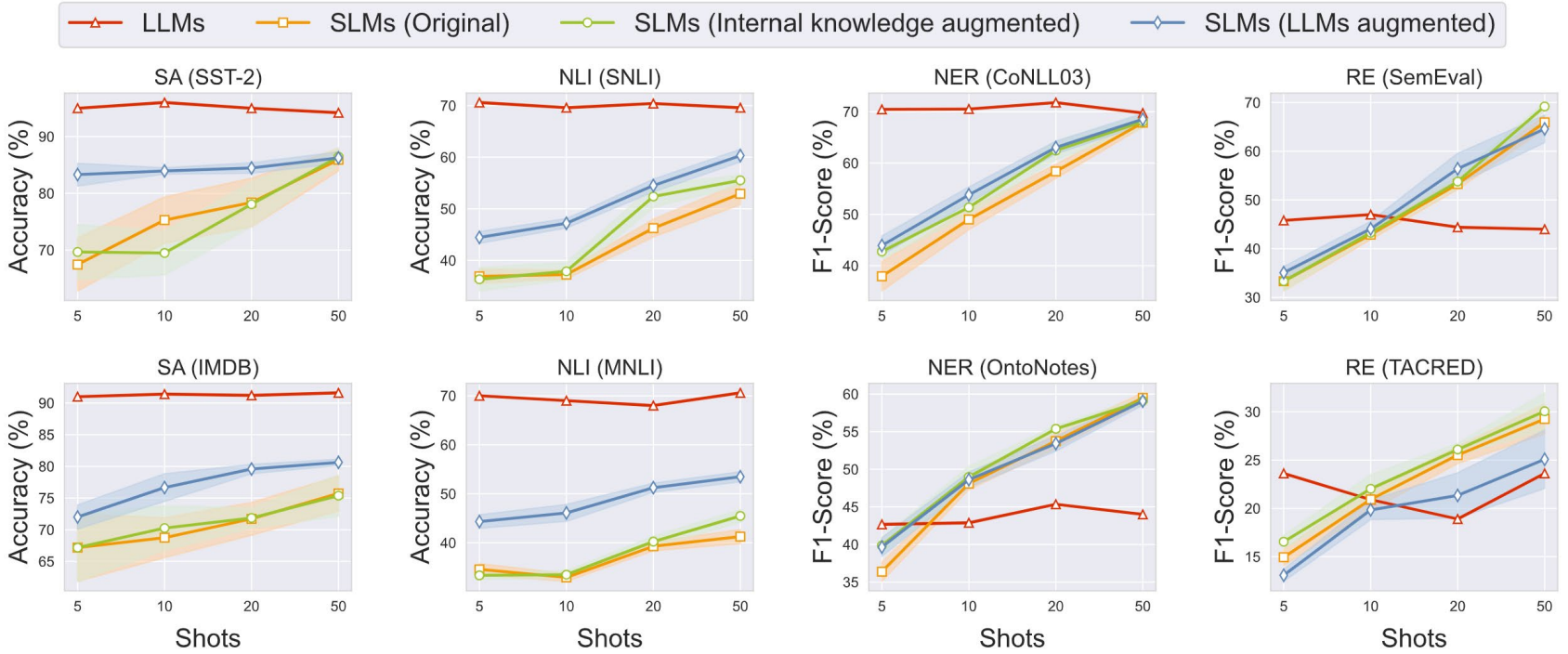


Figure 3: Performance comparison under few-shot settings. The LLMs refer to GPT-3.5. The results of SLMs are obtained by averaging the performance of BERT-based and BART-based fine-tuned models.

3.3 Weaknesses analysis of LLMs for counterfactual generation

3.3.1 The quality of generated counterfactuals is bounded by LLMs' task-specific performance

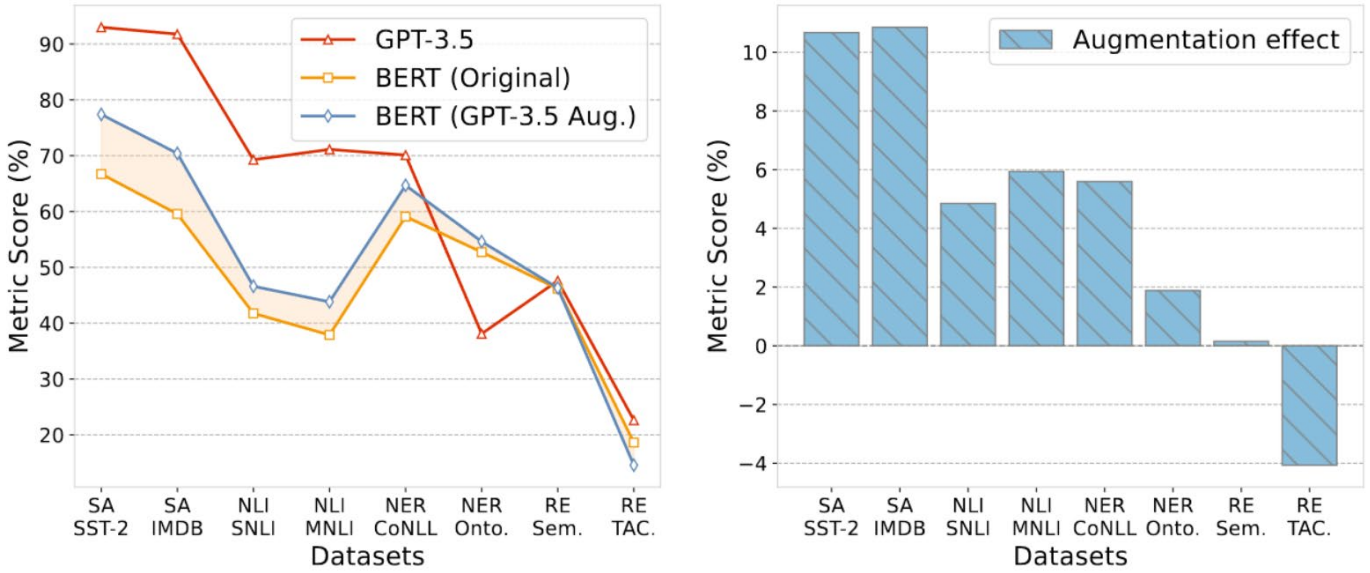


Figure 4: Task-specific performance (left) of LLMs and augmentation effects on SLMs (right).

3.3 Weaknesses analysis of LLMs for counterfactual generation

3.3.2 LLMs fail to fully consider entity constraints when generating counterfactuals for RE

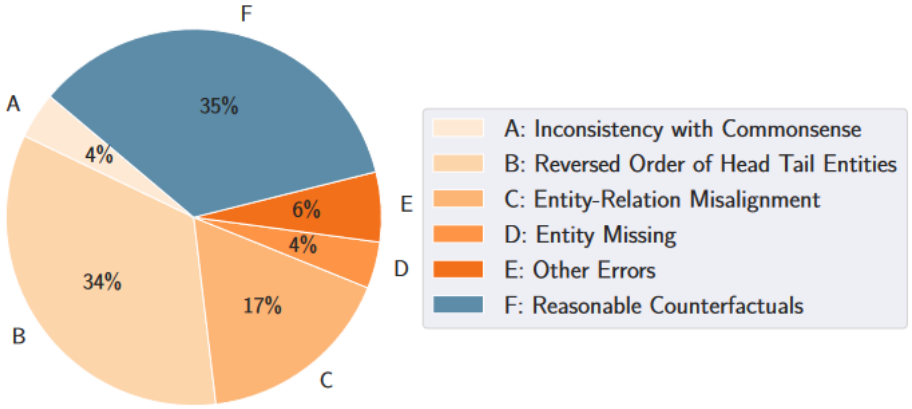


Figure 5: Reasons that lead to unreasonable counterfactuals and corresponding proportions.

Case of Type B	The flight departs from an airport on territory of a member state to which the Treaty applies.	Entity-Origin
Counterfactual	The flight arrives at an airport on the territory of a member state to which the Treaty applies.	Destination-Entity
Case of Type C	The woods that furnish the best charcoal for painters are the beech and vine.	Instrument-Agency
Counterfactual	The beech and vine are the origin of the best charcoal for painters .	Entity-Origin

Table 1: Cases corresponding to the left figure.

3.3 Weaknesses analysis of LLMs for counterfactual generation

3.3.3 Selection bias in LLMs undermines counterfactual generation for the RE task

- LLMs prefer to choose more **abstract** relation types as the target counterfactual ones such as "Entity-Origin".
- Such selection bias leads to a **serious imbalance** of labels in the generated counterfactual sample set.

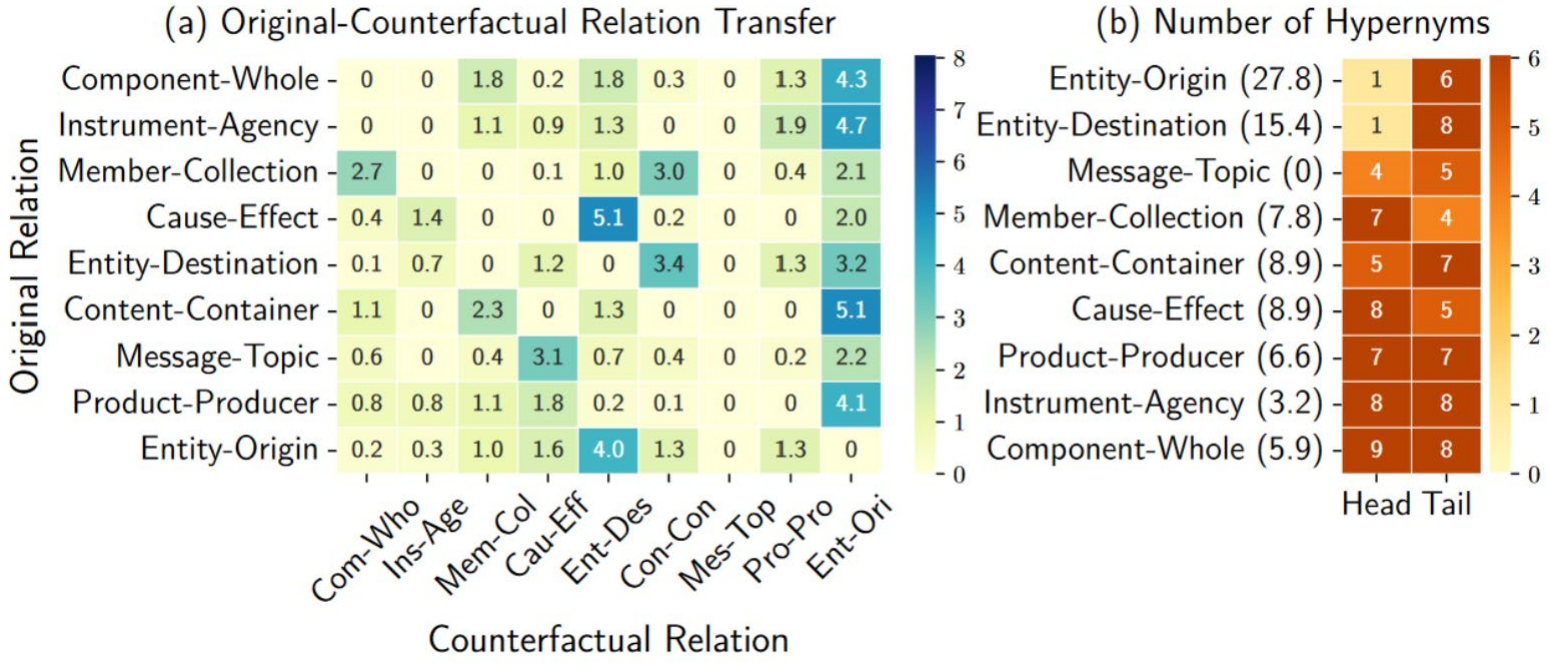


Figure 6: (a) Visualization of original-counterfactual relation transfer frequency. The number represents the frequency of the corresponding transition every 100 samples. (b) Visualization of the number of hypernyms for each head and tail concept. The number in () represents the average frequency of being the target counterfactual relation for every 100 samples.

4.1 Intrinsic properties of LLMs

1) Increasing parameter size **cannot** improve counterfactual generation of LLMs

2) Alignment techniques (such as reinforcement learning from human feedback) may **help** improving counterfactual generation of LLMs.

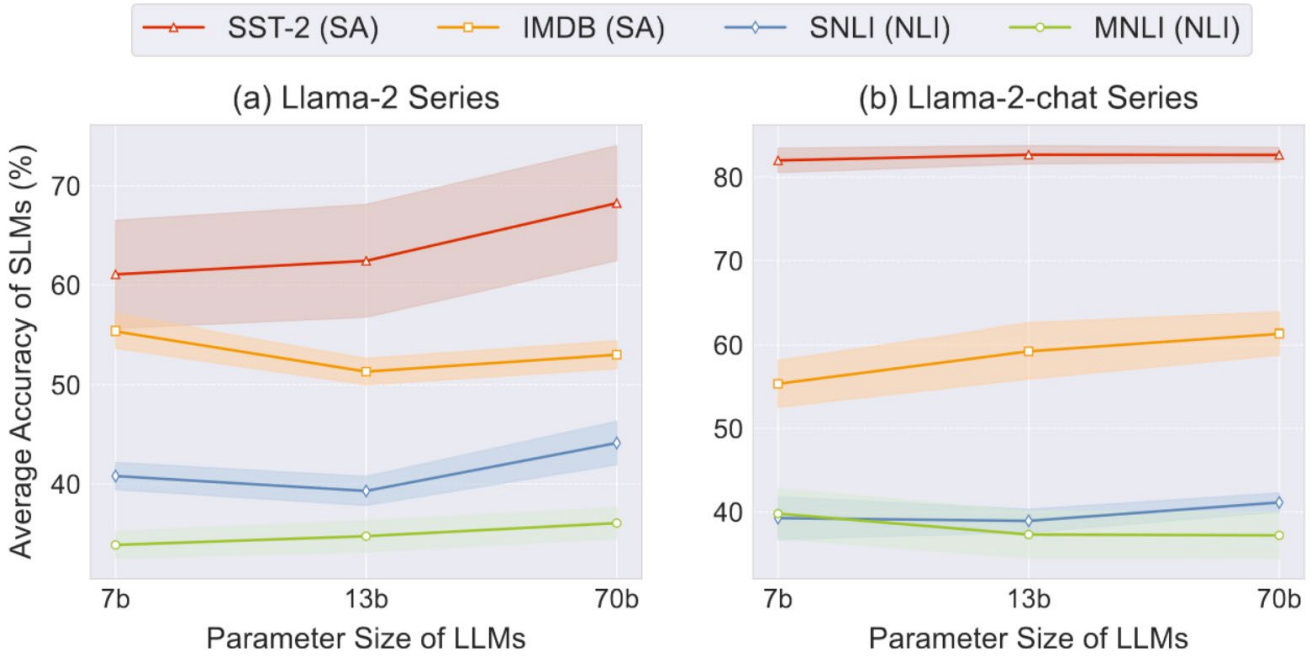


Figure 7: Performance comparison of counterfactually augmented SLMs. The counterfactuals are generated by Llama-2 (left) and Llama-2-chat (right, the **aligned ones**) series with different parameter sizes.

4.2 Impact of prompt designing

1) Task guidelines are critical for counterfactual generation

2) Chain-of-thought does not always help

3) Even unreasonable demonstration can yield reasonable counterfactuals

The audience can find details in the paper.

Conclusion

- This paper presents the first evaluation framework and a systematical empirical study on the capability of LLMs in generating counterfactuals.
- Experimental results on four typical NLU tasks including SA,NLI, NER, and RE demonstrate that LLMs can generate satisfactory counterfactuals in most cases.
- However, LLMs also have their weaknesses when dealing with complex tasks like RE due to the ignorance of entity constraints and inherent selection bias. (**Future directions**)

Future Work

- Exploring how to incorporate human-defined principles to generate higher-quality counterfactuals, including in the LLM alignment phase or in the prompt design phase.
- Exploring broader applications of LLM-generated counterfactuals beyond data augmentation.
- Exploring the intrinsic mechanism of LLM in performing counterfactual generation or counterfactual inference.

Thanks for Watching!

Presenter: Yongqi Li

liyongqi@whu.edu.cn