

Researching Responsible and Trustworthy Natural Language Processing

What makes an NLP paper?

Emily Allaway

(With slides adapted from Frank Keller)

28 September 2026

School of Informatics
University of Edinburgh

eallaway@ed.ac.uk

Publishing in *CL

Anatomy of an (NLP) Paper

Publishing in *CL

Anatomy of an (NLP) Paper

Title

Abstract

Introduction

Middle

Conclusion

Extras

What goes in the middle?

- Depends on “type” of paper
 - Method paper
 - Dataset paper
 - Review/survey
 - Position/perspective
- Also depends on paper length
 - *ACL conferences: 4 or 8 pages
- Clear organisation is important
 - Story guides this → Medawar’s fraud

Method papers

Describe a new method, either for a specific task or more generally

2. Related Work

3. Methodology

- The novel methodology
- Abstracted from implementation

4. Experiment details, for example

- Implementation information
- Dataset information
- Evaluation metrics
- Baselines

5. Results

Examples: [Jin et al. \(2025\)](#) (specific task method), [Dai et al. \(2025\)](#) (general method)

Dataset papers

Creates a new dataset either for an existing or a new task

2. Related Work

3. The new dataset

- Details about collection
- Often: human evaluation
- Key information: examples, statistics, etc.
- If a new task: define the task

4. Methods

- Often a simple model is also proposed, OR
- Discuss existing models to be evaluated
- Often includes implementation details

5. Results

Example: [Zhao and Caragea \(2025\)](#) (new task & data)

Why do dataset papers often include a new method too?

Why do dataset papers often include a new method too?

- Construct validity (methodological rigor)
 - does the dataset measure what it is claiming?
- Normative rigor
 - why does the dataset matter

An alternative structure for a dataset paper

3. Dataset collection process
4. Experiment details
 - Details about human & LLM participants
 - Metrics
5. Experiment 1
 - 5.1 Setup
 - 5.2 Results
6. Experiment 2
 - 6.1 Setup
 - 6.2 (and 6.3) Results

From [Chi et al. \(2026\)](#).

Descriptive Headings

Stylistic considerations for section headings:

- Section headings help readers jump around
- avoid cryptic one-word titles also for sections
- use parallelism (all subsections are titled using noun phrases, participles, etc)
- Don't use dangling subsections (1, 2, 2.1, 3, 3.1, 3.2, 4)

Look at the **table of content** to see if your sectioning works.

Put `\tableofcontent` at the beginning of your paper (even if the final version won't have a ToC).

What background is needed to understand & *justify* the work?

- Spend time on this, it takes longer than you expect
- Don't only look at recent works
- Discuss what is *relevant* about prior works
- Make *explicit* what is different about your work
- Read *a lot*

× *Beware:* Citing only recent work

2 Related Work

Stance Detection with External Information. A key line of related work investigates leveraging external information, often from Wikipedia, to enhance stance detection. [He et al. \(2022\)](#) fine-tuned BERT models which take Wikipedia excerpts, in addition to given texts and targets, as inputs and report significantly improved stance detection performance. Subsequent works in the literature either utilized external information in a different formu-

lation of stance detection ([Wen and Hauptmann, 2023](#)) or introduced new knowledge organization and filtering schemes for such information ([Li et al., 2023](#); [Zhu et al., 2022](#)). While these works have primarily focused on fine-tuning smaller, BERT-like models for stance detection, we extend this research to LLMs, which possess emergent reasoning abilities but require significantly more resources for fine-tuning.

Stance Detection with LLMs. Relatedly, another stream of works examines how LLMs can be applied to stance detection. [Weinzierl and Harabagiu \(2024\)](#) and [Lan et al. \(2024\)](#) proposed prompting schemes where reasoning on stance is organized as ensembles or multi-agent discussions. Meanwhile, [Li et al. \(2024\)](#) introduced a calibration network which serves to mitigate internal biases of LLMs. Orthogonal yet complementary to these efforts, our work provides a foundational analysis of how external information influences their decision-making, uncovering unintended effects and offering insights to guide future research.

× *Beware:* Too descriptive

From Jin et al. (2025).

2.1 Argument Quality Assessment

The definition of argument quality is a complex problem, and many studies have conducted exploration on this question (Swanson et al., 2015; Wachsmuth et al., 2016; Joshi et al., 2023; Fromm et al., 2023). Building on a comprehensive summary of previous work (Hamblin, 1970; Johnson and Blair, 1977; Aristotle and Kennedy, 1991; Eemeren and Grootendorst, 2003), Wachsmuth et al. (2017) proposed a taxonomy of argumentation quality with three major dimensions: logic, rhetoric, and dialectic. According to their definitions, logic focuses on whether the argument is built on acceptable and relevant premises that are sufficient to support the conclusion, while rhetoric assesses the argument's ability to persuade the intended audience of the author's stance, and dialectic examines whether the argument contributes meaningfully and acceptably to resolving the issue for the target audience. Based on this taxonomy, Lauscher et al. (2020) constructed an argument quality corpus and explored interactions between different dimensions. Toledo et al. (2019) presented an argument quality annotation method that can transform binary judgments made by multiple annotators for a given argument into a reliable overall argument quality score. Based on this method, they also constructed an argument quality dataset containing approximately 5.3k arguments annotated with the overall quality scores. Gretz et al. (2020) released a larger argument quality dataset with around 30k arguments, following the same annotation protocol.

Following these data collection efforts, the computational methods for argument quality assessment have evolved significantly. Marro et al. (2022) used argument structure information derived from graph embeddings to enhance the performance of argument quality assessment. Wang et al. (2023b) leveraged contrastive learning to distinguish arguments of different quality more effectively. Bao

✓ Do: Explicitly contrast with prior work

Language Models Pretrained language models (PLMs) such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) have been extensively used in stance detection (Glandt et al., 2021; Allaway and McKeown, 2020; Li et al., 2021). More recently, large language models (LLMs) have been developed, offering the advantage of handling downstream tasks directly through prompting techniques (Le Scao et al., 2023; Touvron et al., 2023; Team et al., 2023; Naveed et al., 2023). While some research on stance detection has utilized LLMs (Gatto et al., 2023; Li et al., 2023a; Fraile-Hernandez and Peñas, 2024), these applications are limited to monolingual scenarios. Existing multilingual stance detection works solely employ PLMs (Vamvas and Sennrich, 2020). In contrast, our study leverages both PLMs and LLMs in the multilingual context, which enables us to explore ZSSD in a more comprehensive setting, enhancing our understanding and capabilities in this area.

From Zhao and Caragea (2025).

Publishing in *CL

Anatomy of an (NLP) Paper

Title

Abstract

Introduction

Middle

Conclusion

Extras

Conclusion

Summarizes the middle and provides a future perspective:

- provide an analysis of the most important results
- analyze the results overall, not individually (do that in the middle, in a Discussion section)
- do not present new evidence or new results here
- provide a future perspective:
 - recommendations that derive from your work
 - future direction of your work
 - re-iterate the scope and limitations of your work (already in the intro)

Conclusion ties together loose ends, provides closure.

Publishing in *CL

Anatomy of an (NLP) Paper

Title

Abstract

Introduction

Middle

Conclusion

Extras

- Limitations → required by *ACL conferences
- Ethics statement
- Appendices

× *Don't*: Surface level

Limitations

Currently, most assessments of CoT distillation focus primarily on accuracy ([Magister et al., 2023](#); [Ho et al., 2023](#); [Shridhar et al., 2023](#); [Wang et al., 2023c](#)), which is insufficient because safe LLMs rely heavily on trustworthy CoTs. We hope the community to develop standards for evaluating the quality of CoTs, rather than relying solely on automatic assessments by GPT-4.

From [Dai et al. \(2025\)](#).

Limitations

× *Don't*: Focus on future work

Limitations

Our study focuses on three key domains where LLMs explicitly struggle—Math, Intention, and Time—building on insights from existing literature. However, LLMs also face challenges in areas such as long-tail knowledge and domain-specific expertise, where external resources are essential. Expanding SMART-ER to these domains could further refine model self-awareness and improve calibration in knowledge boundary, complementing the strong OOD performance that SMARTAgent has already demonstrated. Additionally, while we evaluate our approach on two major model families, extending our analysis to a broader range of architectures, including Qwen, DeepSeek, and varying model sizes, could further validate and enhance the generalizability of our findings.

- ✓ Do: Describe scope of study (future work OK in this context)

6 Limitations

Hyperparameter Search The current EXPO adopts the simplest form of uniform extrapolation and requires manual hyperparameter search for α . Future work could explore how to determine the optimal α automatically and adaptively (i.e., using different α values for different model modules). For example, the information from optimizer states and parameter gradients during the later phase of alignment training could be useful for this purpose.

From [Zheng et al. \(2025\)](#).

Limitations

✓ Do: Discuss assumptions

Limitation

Dependency on Backbone Models The effectiveness of our GLPN-LLM framework is closely tied to the performance of the underlying backbone models, namely FCN. While these models provide strong feature representations, any limitations in their ability to capture comprehensive semantic relationships can directly impact the label propagation process. Consequently, the overall detection accuracy is highly dependent on the quality and robustness of these backbone models. Additionally, the reliance on specific backbones may limit the adaptability of our framework to other feature extraction architectures that might offer different advantages.

Reliance on High-Confidence Pseudo Labels

Our approach relies on the generation of high-confidence pseudo labels by the LLM to enhance label propagation. However, the accuracy of these pseudo labels is contingent upon the LLM's ability to produce reliable predictions. Inaccurate or biased pseudo labels can introduce noise into the label propagation process, potentially degrading the model's performance. Ensuring the reliability of pseudo labels is crucial, and future work may explore more robust methods for pseudo label verification and refinement to mitigate this limitation.

From [Hu et al. \(2025\)](#).

Exercise 2: Evaluating limitations

Look at the limitations on the next page. Discuss the following questions:

1. Can you identify the scope of the study?
2. Is it clear in which situations the results may not be applicable??
3. Are there points you would expand on?
4. Given the abstract, are there any potential limitations not mentioned you notice?

These limitations are from papers that appeared at ACL 2025.

Exercise 2: Evaluating limitations

Limitations

Sensitive to Prompts. As with other LLM prompting studies (Zhang et al., 2024b; Huang et al., 2024b; Zhang et al., 2024a; Chen et al., 2024), our results may be sensitive to prompt. While our prompts underwent rigorous review and testing, and our main experiments report averages across over 8,000 problems, optimizing prompts for this specific task remains a significant challenge and area for future research.

Generalizability to Other Programming Tasks.

In accordance with scientific rigor, this study defines its scope as Human-LLM collaboration within competitive programming, a domain chosen to examine the capabilities and limitations of both LLMs and human performance. While acknowledging the potential relevance to broader programming tasks, we limit our evaluations and analyses to this specific context and defer extending the representativeness of our results to general software development or other programming domains. Despite this focus, elements of our work offer valuable insights applicable to diverse programming scenarios. The problem-solving process shares fundamental similarities across programming contexts, and our proposed human feedback taxonomy and methods for improving problem comprehension in LLMs may readily translate. Developers, for example, can leverage clear and detailed feedback on specifications, as demonstrated in our benchmark, to guide LLMs towards a better understanding of software requirements. We believe this highlights pathways for broader applicability and welcome further discussion.

Exercise 2: Evaluating limitations

Abstract

While recent research increasingly emphasizes the value of human-LLM collaboration in competitive programming and proposes numerous empirical methods, a comprehensive understanding remains elusive due to the fragmented nature of existing studies and their use of diverse, application-specific human feedback. Thus, our work serves a three-fold purpose: First, we present the first taxonomy of human feedback consolidating the entire programming process, which promotes fine-grained

evaluation. Second, we introduce ELABORATIONSET, a novel programming dataset specifically designed for human-LLM collaboration, meticulously annotated to enable large-scale simulated human feedback and facilitate cost-effective real human interaction studies. Third, we introduce ELABORATION, a novel benchmark to facilitate a thorough assessment of human-LLM competitive programming. With ELABORATION, we pinpoint strengths and weaknesses of existing methods, thereby setting the foundation for future improvement.

Exercise 2: Evaluating limitations

Limitations

Although this work demonstrates the potential of RLKGF, several issues need to be addressed. The quality of feedback derived from knowledge graphs depends heavily on the completeness and accuracy of the graph itself, particularly in open domains. Our experiments are limited to disease diagnosis tasks without exploring RLKGF's generalization to other tasks and domains. Additionally, due to data limitations, we do not conduct experiments across a broader medical framework.

The current task format is single-turn Q&A, and future work should explore multi-turn dialogues to better leverage the potential advantages of knowledge graph structure and semantics in multi-step reasoning. Moreover, RLKGF currently focuses primarily on entity-level feedback for model responses, with limited focus on overall response

fluency. Furthermore, experimental comparisons show that although RLKGF improves consistency between model responses and knowledge, there is still significant room for enhancement. Designing appropriate reward ranges and investigating the impact of different methods on model parameter adjustments are crucial for continuous knowledge learning.

Appendices

Bonus content - main paper should be self-contained

✓ DOs

- Put examples
- Hyperparameter details
- Extra analyses
- Annotation instructions
- Follow formatting guidelines

✗ DON'Ts

- Put *all* your examples
- All experiment details
- Main results
- All annotation information
- Include content that spills into margins

Often used for details in the *Reproducibility Checklist* (we'll discuss in a few weeks)

Exercise 3: Project pitches

Discuss your pitch

- Do you have a compelling example?
- Is the importance clear?
- How rigorous is your proposed work?

- Emily Allaway and Kathleen McKeown. 2020. [Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8913–8931, Online. Association for Computational Linguistics.
- Michael Alley. 2018. *The Craft of Scientific Writing*, 4 edition. Springer, New York, NY.
- Alison Chi, Sanghyeok Choi, Hannah Rohde, and Emily Allaway. 2026. Hey, guess what: Cues lead to inferences about typicality. *In submission*.
- Chengwei Dai, Kun Li, Wei Zhou, and Songlin Hu. 2025. [Capture the key in reasoning to enhance CoT distillation generalization](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 441–465, Vienna, Austria. Association for Computational Linguistics.
- Boyi Deng, Yu Wan, Baosong Yang, Yidan Zhang, and Fuli Feng. 2025. [Unveiling language-specific features in large language models via sparse autoencoders](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4563–4608, Vienna, Austria. Association for Computational Linguistics.

- Chenxin Diao, Nataliya Stepanova, and Emily Allaway. 2026. How broad is that claim? mapping generalization in nlp research. *Accepted to EMNLP*.
- Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. [A fine-grained comparison of pragmatic language understanding in humans and language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4194–4213, Toronto, Canada. Association for Computational Linguistics.
- Shuguo Hu, Jun Hu, and Huaiwen Zhang. 2025. [Synergizing LLMs with global label propagation for multimodal fake news detection](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1426–1440, Vienna, Austria. Association for Computational Linguistics.
- Bojun Jin, Jianzhu Bao, Yufang Hou, Yang Sun, Yice Zhang, Huajie Wang, Bin Liang, and Ruifeng Xu. 2025. [A multi-persona framework for argument quality assessment](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12148–12170, Vienna, Austria. Association for Computational Linguistics.

- Benjamin Newman, Yoonjoo Lee, Aakanksha Naik, Pao Siangliulue, Raymond Fok, Juho Kim, Daniel S Weld, Joseph Chee Chang, and Kyle Lo. 2024. Arxivdigestables: Synthesizing scientific literature into tables using language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9612–9631.
- Quang Minh Nguyen and Taegyeon Kim. 2025. [Is external information useful for stance detection with LLMs?](#) In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14798–14807, Vienna, Austria. Association for Computational Linguistics.
- Sukannya Purkayastha, Zhuang Li, Anne Lauscher, Lizhen Qu, and Iryna Gurevych. 2025. [LazyReview: A dataset for uncovering lazy thinking in NLP peer reviews](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3280–3308, Vienna, Austria. Association for Computational Linguistics.
- Cheng Qian, Emre Can Acikgoz, Hongru Wang, Xiusi Chen, Avirup Sil, Dilek Hakkani-Tür, Gokhan Tur, and Heng Ji. 2025. [SMART: Self-aware agent for tool overuse mitigation](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4604–4621, Vienna, Austria. Association for Computational Linguistics.

- Abhilasha Ravichander, Shruti Ghela, David Wadden, and Yejin Choi. 2025. [HALoGEN: Fantastic LLM hallucinations and where to find them](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1402–1425, Vienna, Austria. Association for Computational Linguistics.
- Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3027–3035.
- Keith Tyser, Ben Segev, Gaston Longhitano, Xin-Yu Zhang, Zachary Meeks, Jason Lee, Uday Garg, Nicholas Belsten, Avi Shporer, Madeleine Udell, et al. 2024. Ai-driven review systems: evaluating llms in scalable and bias-aware academic reviews. *arXiv preprint arXiv:2408.10365*.
- Yating Wu, Ritika Mangla, Greg Durrett, and Junyi Jessy Li. 2023. Qudeval: The evaluation of questions under discussion discourse parsing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5344–5363.

- Lian Yan, Chen Tang, Yi Guan, Haotian Wang, Songyuan Wang, Haifeng Liu, Yang Yang, and Jingchi Jiang. 2025. [RLKGF: Reinforcement learning from knowledge graph feedback without human annotations](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6619–6633, Vienna, Austria. Association for Computational Linguistics.
- Chenye Zhao and Cornelia Caragea. 2025. [Bilingual zero-shot stance detection](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29900–29919, Vienna, Austria. Association for Computational Linguistics.
- Chujie Zheng, Ziqi Wang, Heng Ji, Minlie Huang, and Nanyun Peng. 2025. [Model extrapolation expedites alignment](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1025–1041, Vienna, Austria. Association for Computational Linguistics.
- Justin Zobel. 1997. *Writing for computer science*. Springer.