Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 43–45
DOI: 10.15439/2025F4610
ISSN 2300-5963 ACSIS, Vol. 43

Human-centered LLMs for Inclusive Language Technology: The Need to Embrace Variation Holistically in NLP

Barbara Plank
LMU Munich
Email: b.plank@lmu.de

***Abstract*—Large Language Models (LLMs) have advanced rapidly but often still cater primarily to a narrow set of users. This position paper advocates for a human-centered approach to NLP technology—one that embraces linguistic variation, improves reasoning and safety, and better serves diverse communities. We outline key challenges with current LLMs, highlight opportunities in modeling variation in both language and human annotation, and outline a path toward more inclusive and trustworthy language technologies.**

I. Introduction

NATURAL Language Processing (NLP) has entered an era where large pretrained language models (LLMs) dominate. They perform impressively across many benchmarks, yet often fail to properly handle language and are poorly aligned with today's societal values, opinions and attitudes. A central reason is that language technology has historically focused on *standard*, canonical varieties of language in *the input space*, such as English newswire [1]. Moreover, the wide-spread focus on learning from a myopic view of the *output space* has led to ignoring plausible variation in labels—human label variation (HLV) [2]—instead the common practice has focused on learning from a "ground truth," typically only a majority view. This practice has resulted in systems that work well for a narrow population, while excluding dialects, minority languages, diverse user groups and diverging individual perspectives and interpretations. At the same time, LLMs exhibit well-documented trust issues, such as sensitivity to prompt wording, overconfidence under uncertainty, and biases, cf. §II.

In this paper, we advocate for a human-centered perspective on LLMs. Rather than treating language as a homogeneous object with distinct, clear-cut categories (such as between languages), we argue for embracing *variation*—namely, three kinds of variation: in inputs (linguistic diversity), in outputs (human label variation), and in research itself. Doing so will be essential to build inclusive and trustworthy NLP systems.

II. Challenges and Trust Issues with LLMs

Despite their power, LLMs face several well-documented challenges, including (non-exhaustively):

- **Prompt sensitivity:** Performance can vary widely depending on minor linguistic changes in the prompt, undermining robustness (e.g. [3]).
- **Evaluation artifacts and Reasoning Faithfulness:** Widely-used protocols, such as relying on first-token probabilities in multiple-choice QA (e.g. [4], [5]), can distort performance estimates. Moreover, a model's reported accuracy—the correctness of its final conclusion—does not necessarily reflect the validity of its underlying reasoning process [6].
- **Overconfidence and Hallucinations:** Models may provide overly certain answers in ambiguous settings, or hallucinate content, reducing their reliability in sensitive applications (e.g. [7]).

These challenges highlight the need for models that not only generate fluent outputs but also signal uncertainty and acknowledge ambiguity [8].

III. Embracing Variation Holistically

Variation affects all stages of the NLP pipeline: data, modeling, and evaluation. We highlight three key aspects, where the last touches upon the broader research ecosystem.

A. Language Variation

Language is inherently variable, spanning dialects, sociolects, and registers. Traditional NLP systems often ignore this diversity, treating languages as monoliths. Recent work demonstrates the need for multi-dialectal datasets, both for text and speech (e.g., Bavarian Universal Dependencies [15], multi-dialectal German ASR and dialect-to-standard translation [16]). To tackle such issues, robust modeling strategies that can handle non-standard forms are needed. Techniques such as noise injection or subtoken-level modeling provide partial solutions [9], but a broader shift toward valuing non-standard data is required [1].

At the same time, we need systems that not only understand language varieties but also capture the fine-grained nuances of linguistic expressions. For instance, while safety guardrails are important, current LLMs can be overly sensitive, leading to false refusals where harmless requests are blocked. This issue has been systematically studied with resources such as XSTest [13]. To address this, we argue that models must be made safer for the *right reasons*. One promising line of work is steering methods that enable targeted mitigation: for example, Wang et al. propose a surgical and flexible single-vector ablation

approach that reduces false refusals while preserving model performance and inference efficiency [14].

B. Human Label Variation and Pluralistic Alignment

Equally important is variation in human judgments. Annotation is not always reducible to a single "ground truth." Disagreement among annotators often reflects genuine ambiguity or multiple valid perspectives rather than mere noise [2], [10]. Embracing this human label variation (HLV)—through collecting distributions, explanations, or multi-perspective annotations—enables models that better align with human diversity.

HLV also resonates with current discussions of *pluralistic alignment*, which argues that AI systems should not optimize for a single normative view of correctness, but instead represent and respect the spectrum of human perspectives. Sorensen et al. [17] outline a roadmap to pluralistic alignment, emphasizing that different values and interpretations can legitimately coexist. Similarly, Rieser's ACL 2025 keynote [18] highlights the importance of building AI that acknowledges multiple standpoints rather than enforcing one "canonical" answer. By treating variation in human labels as signal rather than noise, we can move toward alignment strategies that capture the richness of human perspectives and support more inclusive language technology.

C. Research diversity and interdisciplinarity

Finally, we need to address variation not only in model inputs and outputs, but also in our research design and the broader research ecosystem. Embracing variation in research means fostering interdisciplinary collaboration while centering human needs and agency. This requires continuously asking whether the AI technologies we develop are those that people truly need and want [19], [20]. Only by integrating diverse perspectives throughout the development process and engaging with insights from multiple disciplines can we build research that captures the richness and diversity of human language—and, by extension, human communication.

IV. Toward Trustworthy Human-Centered NLP

Building inclusive language technology requires a shift in perspective toward **human-centered NLP**. Human-centered NLP emphasizes that systems must serve diverse communities, respect linguistic variation, and ultimately support human needs and agency—to design NLP systems with humans in mind from the ground up. The reason this shift is necessary is simple yet profound: **variation is inherent to language**. Human language is characterized by productivity, ambiguity, and a rich spectrum of linguistic variation across speakers, dialects, registers, and contexts. This variation, while natural and inevitable, exposes the limitations of NLP systems typically trained on narrow and standardized datasets.

Variation inevitably gives rise to **uncertainty**. Ambiguity in meaning, multiple valid interpretations, or disagreements in human judgments all contribute to uncertainty in language processing. Yet, today's models are not good at handling uncertainty: they often fail to recognize when they do not know, and instead provide overly confident answers even in ambiguous or adversarial cases, and may hallucinate. Such behavior undermines user trust and risks excluding communities whose language varieties fall outside the models' dominant (post-)training distributions.

Understanding and explicitly modeling uncertainty, however, is not only key to technical robustness but also central to human-centered NLP. A human-centered perspective requires systems that acknowledge and communicate their limitations transparently. If models can identify when they may be wrong, or signal when several plausible interpretations exist, users are better equipped to interpret outputs, retain agency, and make informed decisions.

This brings us back to the central question: *what is trust?* As early as 1979, David G. Hays offered a concise and influential definition:

> *Trust arises from knowledge of origin as well as from knowledge of functional capacity.* [21]

Decades later, this observation remains highly relevant. Today, the topic of trustworthiness in NLP and AI is an ongoing discussion deserving special attention [22], [23]. To establish trust in a human-centered way, it is time to rethink how we design tasks and their evaluation protocols. Why now? It is increasingly difficult to predict a priori when models trained on web-scale data will work well. In a hypothetical world with complete knowledge of both origin and functional capacity, each task instance could be routed to the most suitable model (or agent, nowadays), enabling not only the full use of LLM capabilities but also trust in their predictions. The absence of such knowledge today is directly tied to our lack of confidence in deploying models in real-world scenarios.

This challenge is compounded by a dramatic shift in the field: with the advent of large generative language models, the traditional compartmentalized notion of tasks is breaking down. General-purpose, task-agnostic approaches demand that we move toward a more holistic view of language, placing trustworthiness—and thus human-centered NLP—at the core [12]. This requires rethinking what constitutes a task and developing multi-faceted evaluation protocols that go beyond narrow benchmarks, toward assessing both the origins and the functional capacities of our models. For example, it is an opportunity to rethink the classical NLP tasks we designed, go beyond classification labels (e.g., Natural Language Inference labels) but more towards the nuanced human understanding of languages (e.g., explanations, or more broadly, to uncover more about the reasons and processes beyond an outcome).

V. Conclusion

LLMs offer tremendous potential, but their inclusiveness and trustworthiness remain limited. By embracing variation in language and annotation, and by grounding our models in human-centered principles, we can move toward a *Trust LLM Ecosystem*: NLP systems based on modular, agentic AI that are uncertainy-aware, robust and inclusive and can thereby

better serve diverse communities [11] and support trustworthy interaction with language technology.

ACKNOWLEDGMENT

This position paper is an extended version of an earlier keynote talk I gave at the Annual Meeting of the Association for Computational Linguistics (ACL) 2024 in Bangkok, Thailand. I thank the ACL 2024 organizers, particularly Andre Martins, Vivek Srikumar and Lun-Wei Ku for inviting me. I thank all FedCSIS 2025 organizers, particularly Marcin Paprzyck, for inviting me to Krakow and the opportunity to write this position paper, and present a keynote at FedCSIS. Thanks to all MaiNLP and CIS members for the feedback on earlier versions of the talk. Finally, I would like to acknowledge the European Research Council for funding DIALECT 101043235 that has in large parts enabled research to support my vision of the *variety space* [1] – toward technology that in future is able to capture the full spectrum of variation in natural language, with all its rich, beautiful yet challenging spectrum of diversity.

REFERENCES

[1] B. Plank, "What to do about non-standard (or non-canonical) language in NLP." In KONVENS 2016.
[2] B. Plank, "The 'Problem' of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation." In *Proc. EMNLP* 2022.
[3] A. Leidinger, R. van Rooij, and E. Shutova, "The Language of Prompting: What Linguistic Properties Make a Prompt Successful?," in *Findings of the Association for Computational Linguistics: EMNLP 2023*.
[4] X. Wang, B. Ma, C. Hu, L. Weber-Genzel, P. Röttger, F. Kreuter, D. Hovy, and B. Plank. "My Answer is C: First-Token Probabilities Do Not Match Text Answers in Instruction-Tuned Language Models." In *Findings of the Association for Computational Linguistics: ACL 2024*.
[5] R. Dominguez-Olmedo, M. Hardt, C. Mendler-Dünner. "Questioning the Survey Responses of Large Language Models." In NeurIPS 2024.
[6] P. Mondorf and B. Plank, "Comparing Inferential Strategies of Humans and Large Language Models in Deductive Reasoning," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9370–9402, Bangkok, Thailand, 2024.
[7] A. Testoni, B. Plank, and R. Fernández, "RACQUET: Unveiling the Dangers of Overlooked Referential Ambiguity in Visual LLMs," in *Proc. EMNLP* 2025.
[8] J. Baan, N. Daheim, E. Ilia, D. Ulmer, H.-S. Li, R. Fernández, B. Plank, R. Sennrich, C. Zerva, and W. Aziz, "Uncertainty in Natural Language Generation: From Theory to Applications" arXiv:2307.15703.
[9] V. Blaschke, H. Schütze, and B. Plank. "Does Manipulating Tokenization Aid Cross-Lingual Transfer? A Study on POS Tagging for Non-Standardized Languages." In Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023).
[10] L. Aroyo. "The Many Faces of Responsible AI." Keynote at the Conference on Neural Information Processing Systems (NeurIPS) 2023.
[11] D.Yang, D. Hovy, D. Jurgens, and B. Plank, "Socially Aware Language Technologies: Perspectives and Practices," *Computational Linguistics*, vol. 51, no. 2, June 2025.
[12] R. Litschko, M. Müller-Eberstein, R. van der Goot, L. Weber-Genzel, and B. Plank. 2023. "Establishing Trustworthiness: Rethinking Tasks and Model Evaluation." In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing.
[13] P. Röttger, H. Kirk, B. Vidgen, G. Attanasio, F. Bianchi, and D. Hovy. "XSTest: A Test Suite for Identifying Exaggerated Safety Behaviours in Large Language Models." In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers).
[14] X. Wang, C. Hu, P. Röttger, B. Plank. "Surgical, Cheap, and Flexible: Mitigating False Refusal in Language Models via Single Vector Ablation." In ICLR 2025.
[15] V. Blaschke, B. Kovačić, S. Peng, H. Schütze, and B. Plank, "MaiBaam: A Multi-Dialectal Bavarian Universal Dependency Treebank," in *Proceedings of LREC-COLING*, 2024.
[16] V. Blaschke, M. Winkler, C. Förster, G. Wenger-Glemser, and B. Plank, "A Multi-Dialectal Dataset for German Dialect ASR and Dialect-to-Standard Speech Translation," in *Proceedings of Interspeech*, 2025.
[17] T. Sorensen, J. Moore, J. Fisher, M. Gordon, N. Mireshghallah, C. M. Rytting, A. Ye, L. Jiang, X. Lu, N. Dziri, T. Althoff, and Y. Choi, "A Roadmap to Pluralistic Alignment," arXiv preprint arXiv:2402.05070, 2024.
[18] V. Rieser, "Whose Gold? Re-imagining Alignment for Truly Beneficial AI" Keynote at the Annual Meeting of the Association for Computational Linguistics (ACL), 2025.
[19] S. Bird. 2020. "Decolonising Speech and Language Technology." In Proceedings of the 28th International Conference on Computational Linguistics.
[20] V. Blaschke, C. Purschke, H. Schuetze, and B. Plank. 2024. "What Do Dialect Speakers Want? A Survey of Attitudes Towards Language Technology for German Dialects." In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)
[21] D. G. Hays, "Applications". In Meeting of the Association for Computational Linguistics, 1979.
[22] K. Baum, M. A. Köhl, and E. Schmidt. "Two challenges for CI trustworthiness and how to address them." In Proceedings of the 1st Workshop on Explainable Computational Intelligence (XCI 2017). Association for Computational Linguistics.
[23] J. Eisenstein. "Informativeness and invariance: Two perspectives on spurious correlations in natural language." In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.