Third Year Paper Proposal

Large Language Models make human coding in the social sciences obsolete

Benjamin Lira

GAP in knowledge

Few of the available papers have outcomes, so they are only evaluated on how well they match the humans

Humans are turkers, easy to beat

Constructs are not complex psychological phenomena (e.g., personality, emotions, etc. as opposed to sentiment)

No data on demographics, so what about bias

What about other languages

Existing tests on publicly available benchmarks may be affected by contamination, that is the tests might be included in the training data for these models[1] .

Research Questions

Quality of ratings

Ratings and Demographics

Predictiveness of Ratings

- Does few shot performance produce better results than zero shot performance?

Ancilliary Questions

- Does few shot performance produce better results than zero shot performance?
- How does GPT-4, GPT-3.5, and open source models compare
- What effect does temperature have on rating quality
- What effect does aggregating multiple ratings have on quality

Methods

Reading List

In the past weeks, multiple papers have emerged, discussing the appropriateness of GPT for automated labelling and qualitative coding. Below is a reading list, with some comments for each.

[1] seems to be the only article suggesting caution. They suggest that performance is not uniform, depends on prompt quality, idiosincracies of the text data, and the complexity of the constructs. Thus, they recommend always validating against human ratings. They used 27 tasks in 11 datasets.
They also introduce the idea of a *consistency score* obtained by repeatedly applying ratings and evaluating distance to the modal response. `temp = .6`

Reiss[2] found negative results.

[3]

[4]

[5] provides a thorough investigation of LLMs for computational social science.

[6]

Rathje et al[7] has probably written the paper closest to the paper I wanted to write. They focus not just on sentiment but also on discrete emotions. They compare the accuracy of LLMs (.66 - .75) to that of LIWC (.20 - .30). It seems like they obtain good results mostly because they are rating simple constructs (sentiment, discrete emotions, and offensiveness) in short texts (mostly tweets)

[8]

[9]

[10]

[11]

[2]

[12]

[13]

Some notes of caution

I wonder if it this is a project worth pursuing. Below is a list of potential problems.

- While the project seemed like a good idea when it came about, the space has quickly saturated and

what seemed like a big contribution has now been covered in parts by the multitude of papers cited here. Thus, the size of the contribution gets smaller by the day.
- This is not an **evergreen** research question. Soon GPT-5 will be out and the answers will be outdated. Thus, the relevance of the contribution gets smaller by the day.
- It is difficult to find enough datasets that match inclusion criteria.

Extra notes to self

- Making every classification binary can be helpful to standardize performance. E.g., don't do multiclass or multilabel classification.

Large language models (LLMs) are neural networks that are trained on massive amounts of text data and can generate natural language in response to various inputs. LLMs have shown remarkable capabilities in natural language understanding and generation tasks, such as question answering, summarization, and text classification. However, most LLMs are trained on data that is predominantly in English, which may limit their ability to handle other languages and domains.

Few-shot learning is a paradigm that aims to leverage the prior knowledge of LLMs to perform new tasks with minimal supervision. In few-shot learning, the LLM is given a few labeled examples of a new task, along with a natural language prompt that describes the task. The LLM then uses its generative power to produce an answer for the task. Few-shot learning has been applied to various natural language processing tasks, such as sentiment analysis, relation extraction, and named entity recognition.

However, few-shot learning with LLMs has not been extensively explored for the task of classifying psychological constructs from open-ended text. Psychological constructs are abstract concepts that are used to describe and measure human behavior and mental processes, such as personality traits, emotions, attitudes, and motivations. Classifying psychological constructs from text is a challenging task that requires a deep understanding of the meaning and context of the text, as well as the theoretical and empirical foundations of the constructs.

Classifying psychological constructs from text has many potential applications in psychology and related fields, such as education, health, and social sciences. For example, classifying text responses to personality questionnaires can help assess individual differences and predict outcomes. Classifying text responses to surveys or interviews can help measure attitudes and opinions on various topics. Classifying text responses to prompts or scenarios can help elicit emotions and motivations.

However, classifying psychological constructs from text also poses several challenges and limitations. First, there is no consensus on the definition and measurement of many psychological constructs, which may lead to ambiguity and inconsistency in the labels. Second, there is often a lack of large-scale labeled data for many psychological constructs, which may limit the performance of supervised learning methods. Third, there may be ethical and social implications of using LLMs to classify psychological constructs from text, such as privacy, bias, and fairness.

Therefore, in this paper, we aim to investigate the following research questions:

- How do LLMs perform in few-shot classification of psychological constructs from open-ended text?
- How do LLMs compare to human annotators in terms of accuracy and reliability?
- How do LLMs relate to outcomes such as behavior, performance, or well-being?
- How do LLMs relate to demographics such as age, gender, or ethnicity? Do LLMs exhibit any bias or discrimination in their classifications?

To answer these questions, we use a set of articles that cover various psychological constructs and domains. We use a standard few-shot learning framework with natural language prompts to classify the text responses into predefined categories. We evaluate the performance of LLMs against human annotators and baseline methods. We also analyze the correlations between LLM classifications and outcomes and demographics.

We hope that this paper will contribute to the literature on few-shot learning with LLMs and provide insights into the potential and limitations of using LLMs for classifying psychological constructs from text.

References

1. Pangakis, N., Wolken, S., & Fasching, N. (n.d.). *Automated Annotation with Generative AI Requires Validation.*
2. Reiss, M. V. (n.d.). *Testing the reliability of ChatGPT for text annotation and classification: A cautionary remark.*
3. Kim, J., & Lee, B. (n.d.). *AI-augmented surveys: Leveraging large language models for opinion prediction in nationally representative surveys.*
4. Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants? *Trends in Cognitive Sciences*, *27*(7), 597–600. https://doi.org/10.1016/j.tics.2023.04.008
5. Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (n.d.). *Can Large Language Models Transform Computational Social Science?*

6. Sahu, G., Rodriguez, P., Laradji, I. H., Atighehchian, P., Vazquez, D., & Bahdanau, D. (n.d.). *Data augmentation for intent classification with off-the-shelf large language models.*
7. Rathje, S., Mirea, D.-M., Sucholutsky, I., Marjieh, R., Robertson, C., & Bavel, J. J. V. (n.d.). *GPT is an effective tool for multilingual psychological text analysis.* https://doi.org/10.31234/osf.io/sekf5
8. Ding, B., Qin, C., Liu, L., Chia, Y. K., Joty, S., Li, B., & Bing, L. (n.d.). *Is GPT-3 a good data annotator?*
9. Liu, D. L. (n.d.). *Professor Bryony Hoskins University of Roehampton, London, U.K.* 108.
10. Wang, S., Liu, Y., Xu, Y., Zhu, C., & Zeng, M. (2021). Want to reduce labeling cost? GPT-3 can help. *Findings of the Association for Computational Linguistics: EMNLP 2021*, 41954205. https://doi.org/10.18653/v1/2021.findings-emnlp.354
11. Gilardi, F., Alizadeh, M., & Kubli, M. (n.d.). *ChatGPT outperforms crowd-workers for text-annotation tasks.*
12. He, X., Lin, Z., Gong, Y., Jin, A.-L., Zhang, H., Lin, C., Jiao, J., Yiu, S. M., Duan, N., & Chen, W. (n.d.). *AnnoLLM: Making large language models to be better crowdsourced annotators.*
13. Zhu, Y., Zhang, P., Haq, E.-U., Hui, P., & Tyson, G. (n.d.). *Can ChatGPT reproduce human-generated labels? A study of social computing tasks.*