Research Article

Research and Politics
January-March 2024: 1–7
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20531680241231468
journals.sagepub.com/home/rap

Do AIs know what the most important issue is? Using language models to code open-text social survey responses at scale

Jonathan Mellon[1], Jack Bailey[2], Ralph Scott[3], James Breckwoldt[2], Marta Miori[2] and Phillip Schmedeman[1]

Abstract

Can artificial intelligence accurately label open-text survey responses? We compare the accuracy of six large language models (LLMs) using a few-shot approach, three supervised learning algorithms (SVM, DistilRoBERTa, and a neural network trained on BERT embeddings), and a second human coder on the task of categorizing "most important issue" responses from the British Election Study Internet Panel into 50 categories. For the scenario where a researcher lacks existing training data, the accuracy of the highest-performing LLM (Claude-1.3: 93.9%) neared human performance (94.7%) and exceeded the highest-performing supervised approach trained on 1000 randomly sampled cases (neural network: 93.5%). In a scenario where previous data has been labeled but a researcher wants to label novel text, the best LLM's (Claude-1.3: 80.9%) few-shot performance is only slightly behind the human (88.6%) and exceeds the best supervised model trained on 576,000 cases (DistilRoBERTa: 77.8%). PaLM-2, Llama-2, and the SVM all performed substantially worse than the best LLMs and supervised models across all metrics and scenarios. Our results suggest that LLMs may allow for greater use of open-ended survey questions in the future.

Keywords

Large language models, open-text survey questions, text as data, public opinion, most important issue, ChatGPT, GPT-4

Introduction

Open-text questions permit survey respondents to say whatever comes to mind, providing advantages over similar closed items. They avoid prejudging respondents' views (Schuldt and Roh, 2014) and priming preset answers (Ferrario and Stantcheva, 2022).

Still, open-text data have one major drawback: they are hard to use. Often, they need to be coded into closed categories, which presents two main challenges.[1] Firstly, it is time-consuming and expensive, as there are many possible answers to survey questions. Second, this task is often tedious, which may be why scholars often hire others to do it for them.

It would be preferable to spare humans the labor of coding these data by delegating it to computers. Some worry that algorithm-coded data may not be reliable for analysis (He and Schonlau, 2020). Yet artificial intelligence's capability has skyrocketed in recent years (Kaplan et al., 2020). One such tool—the large language model (LLM)—can now perform novel tasks that involve human-readable text (Brown et al., 2020). This raises two interesting questions. Can an LLM perform an open-text survey

[1]Department of Systems Engineering, West Point, West Point, NY, USA
[2]Department of Politics, University of Manchester, Manchester, UK
[3]School of Sociology, Politics and International Studies, University of Bristol, Bristol, UK

Corresponding author:
Jonathan Mellon, Department of Systems Engineering, US Military Academy, West Point, 606 Thayer Rd, West Point, NY 10996, USA.
Email: jonathan.mellon@westpoint.edu.

coding task without tailored training data? And, if so, how does it compare to a trained human coder?

This paper compares LLMs to established machine learning methods and a human at the task of coding responses to the most important issue (MII) question in the British Election Study Internet Panel (BESIP). This is an interesting use case, as MII is a key indicator in research on issue salience (Bevan et al., 2016; Dennison, 2019). In the past, the British Election Study has employed trained human coders. Since 2014, over 657,000 open-text MII responses have been labeled.

We demonstrate modern LLMs' ability to code MII data with near-human accuracy. A human coder initially coded the data, followed by LLMs, several supervised learning methods, and a second human coder. On the most challenging version of the task,[2] the second human coder matched the first's coding 86.6% of the time. State-of-the-art LLMs GPT-4 (80.6%), Claude-1.3 (81.0%), and Claude-2 (79.4%) closely followed. Modern supervised learning methods—a neural network trained on BERT embeddings (74.3%) and DistilRoBERTa (67.7%)—came next, with a traditional support vector machine (SVM) (64.8%) trailing.

Our findings have two implications. First, delegating this work to LLMs could save time and money. Second, doing so could encourage the proliferation of open-ended survey items in the near future. This would be a welcome development. After all, some research questions are difficult to tackle with closed-response survey items. Thus, our results may enhance the measurement of multi-faceted concepts like party and leader images, salient social identities, and issue ownership.

LLMs and social scientific research

Large language models (LLMs) predict the next word in a text based on a large quantity of text training data, and (sometimes) human feedback. Training LLMs is computationally expensive, as modern LLMs use billions or trillions of unique parameters. This structure allows LLMs to interpret natural language instructions and apply abstraction to complete the task. This is true even for novel tasks that they have never been trained to perform (LeCun et al., 2015). For instance, GPT-3.5, untrained in poetry-writing, can complete the prompt:

Write a 3 line poem about Margaret Atwood eating a canary

With a response that never appeared in its training data:

Margaret Atwood dines,

With a canary on her plate,

A fleeting, feathered feast.

The ability to follow novel instructions raises the question of how these models can assist social scientists. Social scientists have tested LLMs with mixed results. LLMs matched or exceeded non-expert human performance in labeling tweet sentiment, tone in political adverts, manifesto ideology, virtues mentioned in political speeches (Ornstein et al., 2022), and in explaining hate speech classifications (Huang et al., 2023). However, they underperformed modern supervised methods in predicting personality and suicidal tendencies (Amin et al., 2023). This inconsistent performance means LLMs' ability to categorize open-text survey responses is uncertain.

Research design and methods

To evaluate LLMs' open-text coding, we must compare them to competing approaches: a research assistant and supervised machine learning methods (a neural network trained on BERT embeddings, DistilRoBERTa, and an SVM).

Our test/training data come from the BESIP (Fieldhouse et al., 2022) respondents' answer to the question "As far as you're concerned, what is the SINGLE MOST important issue facing the country at the present time?" in waves 1–23 of BESIP (2014–2022).

We evaluate LLMs, supervised methods, and the second human coder against the original human coder. Our main measure is accuracy, calculated as the percentage of responses that match the original human coder's category. There are a large number of categories, the distributions of which are uneven. As such, we also report four other metrics in appendix I: Cohen's kappa, F1, AUC ROC, and the Pedersen index. Unless otherwise specified, our conclusions about coder performance are consistent across accuracy and the other metrics we examine.

Novel question scenario

We test the LLMs, human coder, and supervised approaches across two scenarios. In the "novel question" scenario, we investigate how well LLMs fare when labeling open-text data from a novel survey item. Here, existing data is scarce. As such, traditional supervised approaches require a human to label training data before the model can be used. In this scenario, we train the supervised models on 1,000 randomly sampled open-text responses and human-assigned labels from BESIP waves 21–23. The LLMs employ a few-shot prompt.

We test two difficulty levels for the "novel question" scenario: first, performance on the *overall test set* of MII responses from waves 21–23 (81,266 cases), which represents typical performance, and second, performance on the *unique test set* of unique MII responses in those waves

(14,923 cases), providing a stricter test by over-representing rare and idiosyncratic responses.

Existing question scenario

In the "existing question" scenario, we investigate how well LLMs compare where existing data is plentiful. This might be the case when labeling new responses to an existing question that humans have previously labeled. In this case, supervised methods would enjoy much more training data. In our case, we train the supervised models on the full set of around 576,000 coded open-text responses to the MII question from waves 1–20 of the BESIP. The LLMs still use a few-shot prompt in this scenario.

For this "existing question" scenario, we test each coder on the *new test set* of responses from waves 21–23 of the BESIP that had not appeared in previous waves of the data (13,965 cases). We focus on new responses because performance on exact matches to previous responses is not that informative since a researcher could simply copy those labels rather than coding them afresh.

Human classification

Even human coders do not agree 100% of the time. As a result, comparing each computer algorithm to the original labels does not provide a fair test of their capability relative to a trained human coder. Instead, the relevant benchmark is their performance relative to an *independent* human coder.

To ensure a fair comparison, we employed another human coder (a co-author on this paper) to code a random sample of 1,000 open-text responses from each of the three test sets (*overall*, *unique*, and *new* wave 21–23 responses).[3] They received 1 hour of one-to-one training from another team member with experience coding these responses.

LLM classification

Unlike supervised machine learning approaches, LLMs do not require training on a researcher's existing open-text data. Rather, they use the existing background knowledge that they accumulated after being trained on other text data to complete the task. As such, we needed only to construct an appropriate natural language prompt for the LLMs that would lay out what the task was and how to complete it.

The full prompt is shown in appendix C. It starts:

> Here are some open-ended responses from the British Election Study to the question "what is the most important issue facing the country?". Please assign one of the following categories to each open ended text response, returning the original response and the most relevant label.

We listed the categories and pre-empted errors by adding detail to the category names. We mentioned Russia's invasion of Ukraine (which occurred after most LLMs' training data) in the prompt and included an instruction to only provide one label for each response. We provided three examples of the response format.

With only three examples, the LLM relies on the *task description* rather than on *learning by example* (Brown et al., 2020).

We sought to include as many LLMs in our study as possible. In total, we considered 13 open-source and 16 closed-source LLMs. We first tested whether each model could complete the task at all. Appendix F shows each model's response for 15 randomly selected open-text responses. Only one open-source LLM was able to complete the classification task in this form (Meta's Llama-2). Closed-source LLMs were not universally capable of completing the task. Older GPT-3 models from OpenAI could not either. Neither could LLMs from Aleph and Cohere or Google's PaLM-2 text-bison-001.

We could only systematically test OpenAI's GPT-3.5-turbo, and GPT-4, Google's PaLM-2 chat-bison-001, Anthropic's Claude-1.3 and Claude-2, and Meta's Llama-2-Chat LLMs. We do not claim there is *no* prompt that could enable other models to complete the task, but it is notable that only recent LLMs are able to do so with our simple task description.

We queried OpenAI, Google, and Anthropic models using their APIs. For Meta's Llama-2 model, we used replicate.com since we lacked the computing resources to run it locally. We processed 25 open-text responses at a time, as larger batches degraded some models' performance. In general, the LLMs did a good job of using the categories we provided. But some manual edits were necessary. For example, we corrected instances where LLMs labeled responses as "covid" instead of "coronavirus" or where they capitalized the "europe" label. Despite instructions to return a single label, LLMs sometimes returned multiple labels, in which case we used only the first label. LLMs sometimes failed to return a label. Consequently, we retried failed responses up to two times, substantially reducing missing labels for some LLMs. Lastly, we cleaned returned text to aid merging back into the original dataset.

Supervised classification

If human coders represent the traditional approach and LLMs the state-of-the-art, supervised methods represent the status quo. It was therefore appropriate to include them in our analysis. In this case, we use a standard SVM and two more sophisticated supervised models fine-tuned on labeled training data.

To test the SVM approach, we use the RTextTools package in R (Jurka et al., 2013). To represent a typical use case, we apply minimal customizations. We stem the words in the open-text responses and eliminate numbers but otherwise keep the default settings.

The more sophisticated supervised models that we included were DistilRoBERTa and a neural network fit to BERT embeddings. We discuss each in turn and present additional details in appendix B. DistilRoBERTa is a distilled version of the RoBERTa-base model (Liu et al., 2019) that includes around 82 million parameters (Sanh et al., 2020). We fine-tune it using the two training datasets that we discuss above. Others have used similarly fine-tuned versions of DistilRoBERTa to classify tax law (Gu et al., 2022) and detect the sentiment of tweets (Ramos and Chang, 2023).

We also fit a neural network to BERT embeddings of the same training datasets used for fine-tuning DistilRoBERTa and training the SVM. Text embedding places a response in a high-dimensional semantic space based on the prior BERT training. This allows the neural network to classify a new text response correctly, even if it uses words not present in the fine-tuning data, provided the model was trained on text with a similar *meaning* (and thus close in semantic space). The neural network is then used to predict the MII categories for each test set.

Results

We first consider the “novel question” scenario. Figure 1 shows the accuracy of the LLMs, supervised approaches trained on 1,000 randomly sampled cases, and the second human coder. We present two sets of results on each plot: one for the 50-category classification task that we had each algorithm perform and one after recoding these results into 13 higher-order categories.[4]

When considering the 13-category accuracy for the *overall test set*, the best-performing LLM (Claude-1.3: 97.5% accuracy) matches the human’s performance (97.0%). By contrast, the best-performing supervised approach achieved performance slightly below human level (95.3%).

For the *unique test set*, even the best LLM performs less well than the human coder (though only by 1.6 percentage points in the case of Claude-1.3). Supervised approaches again performed less well, with the most accurate (the neural network) performing only about as well as the least accurate LLMs and 10.7 percentage points lower than the human coder. The small neural network’s accuracy and performance may, thus, be vulnerable to distributional changes. Researchers who use these models should label further training data for each new round of data collection.

Across both test sets, the patterns are similar when looking at accuracy for 50 categories except for PaLM-2 and Llama-2 which see precipitous falls in performance.

There is substantial variation in how well LLMs perform under the “novel question” scenario. GPT-4 (90.1%), Claude-1.3 (90.3%), and Claude-2 (90.6%) all achieve similarly high performance on the *unique test set* with 13 categories. OpenAI’s older model, GPT-3.5, then trails them somewhat (84.5%). But the gap is most stark for the other two LLMs. Llama-2’s accuracy was only 80.3%, with PaLM-2 showing a similarly poor score of 79.7%. Llama-2’s performance loss is most dramatic when looking at 50 categories where it achieves only 51.0% accuracy on the *overall test set* and 50.4% accuracy on the *unique test set*. Appendix G shows where this model underperformed.

The supervised models also show performance variation on the “novel question” scenario. The neural network trained on BERT embeddings (e.g., 74.3% accuracy on 50 categories) outperforms DistilRoBERTa on all metrics (67.7% accuracy for 50 categories) which in turn outperforms a traditional SVM approach (64.8%). Using more sophisticated supervised approaches for this task has a meaningful payoff.

The “existing question” scenario we consider is one where a researcher has access to a large historical training dataset and is considering labeling new text using a model. To understand this scenario, we compare the accuracy of a human coder, the LLMs, and supervised models trained on 576,000 labeled cases. We compare their performance on the *new test set* of responses that did not occur in the historical training data. Figure 2 shows the accuracy of these methods on the *new test set*.

The best LLMs fall slightly below human performance when measured on 13 categories (Claude-1.3 is 3.2 percentage points behind the human coder), and the gap widens to 7.7 percentage points when accuracy is measured for 50 categories. Unlike the “new question” scenario, DistilRoBERTa has the highest performance for supervised models. While its performance is significantly behind the best LLMs (3.2 percentage points), the difference in performance is small enough that DistilRoBERTa may plausibly equal LLM performance on other similar tasks. Nonetheless, it is notable that a few-shot LLM approach can match or exceed the performance of a modern supervised algorithm trained on 576,000 cases.

The LLMs show a similar pattern to the previous scenario. Claude-1.3, Claude-2, and GPT-4 perform the best across all metrics. GPT-3.5 performs substantially worse than GPT-4, and PaLM-2 and Llama-2 see a further (often large) performance loss.

The supervised models show a different pattern of performance in the “existing question” scenario than they did in the “new question” scenario, with the neural network

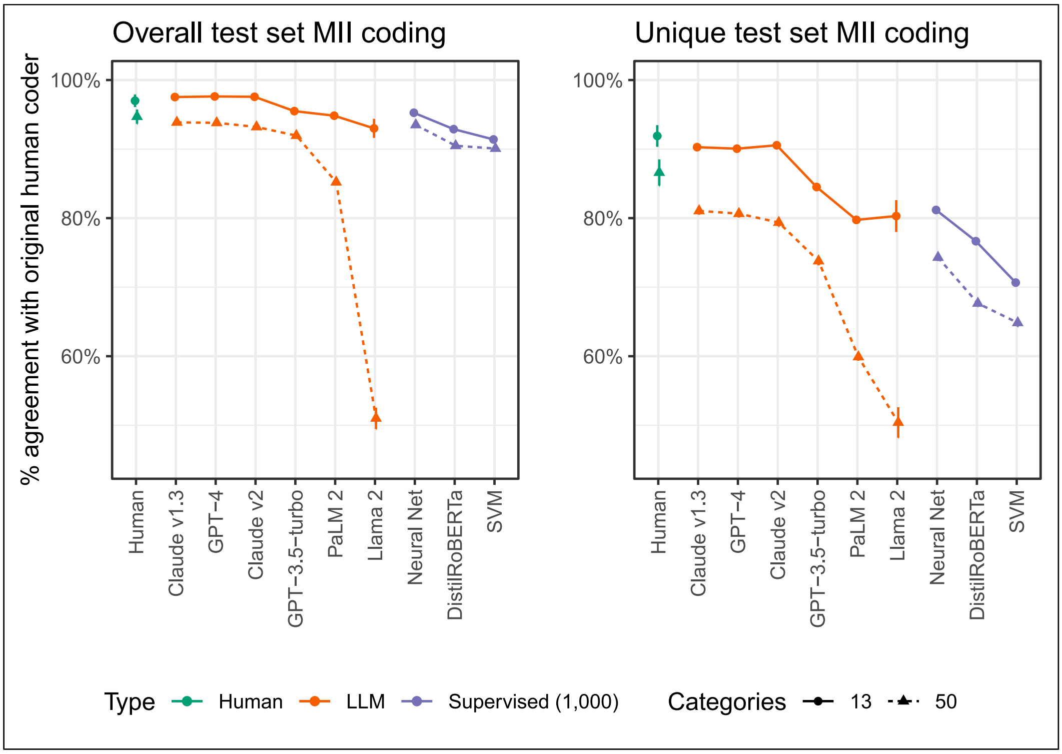

Figure 1. Percentage of responses that agree with original human coder across *overall* and *unique* test sets. Supervised models trained on 1000 cases from BESIP waves 21–23.

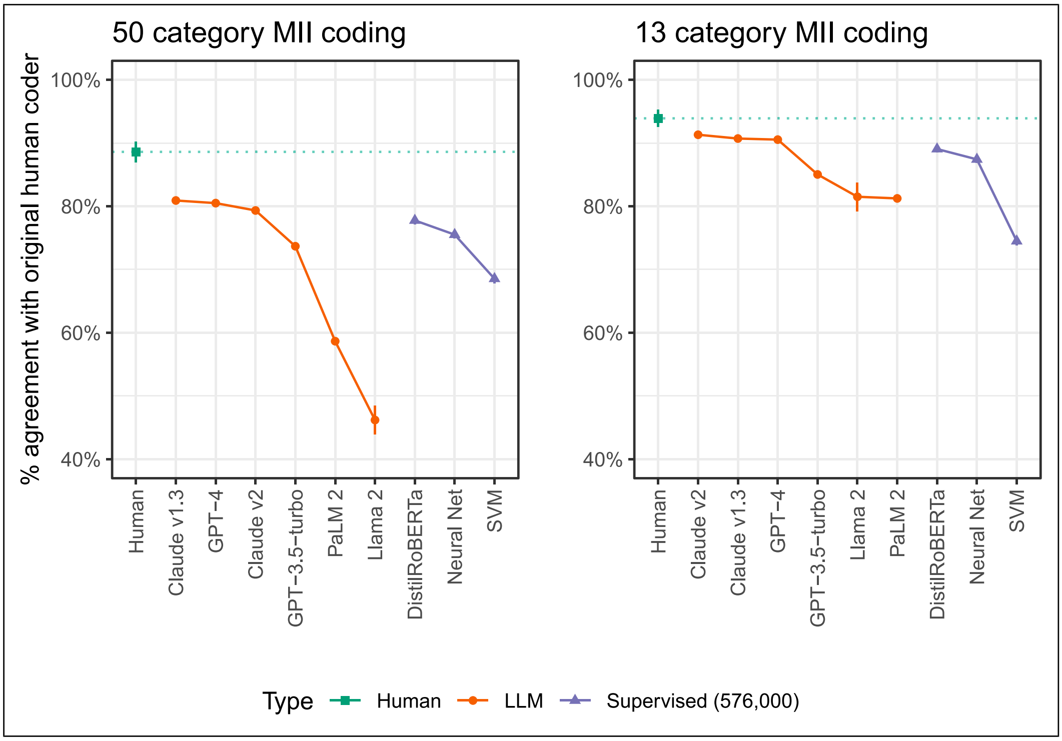

Figure 2. Percentage of responses that agree with original human coder across the *new test set* (text responses that did not appear in waves 1–20). Supervised models trained on 576,000 BESIP waves 1–20 cases.

falling slightly behind DistilRoBERTa with the traditional SVM showing significantly lower performance again.

Conclusions

The best LLMs can code MII data almost as well as a trained human. Like humans, LLMs understand natural language instructions. As such, we can teach them much as we would a typical research assistant. Yet, unlike humans, LLMs are fast, cheap, and plentiful. Employing them, thus, suggests greater productivity at a fraction of the usual cost.

The best LLMs matched or outperformed modern supervised learning approaches in all cases. They were also much easier to work with. As discussed above, LLMs can follow natural language instructions. Modern supervised methods, in contrast, do not. Instead, they need time, expertise, and computational power to achieve their results. Most users are, thus, likely better off using a top-performing LLM.

One option that we did not test was to fine-tune an LLM on our training data. This may allow for even higher performance. Yet neither OpenAI nor Anthropic supported fine-tuning their models when we ran our analysis. The poorly performing PaLM-2 or Llama-2 would have to serve as the base model. Whatever the case, our few-shot approach is more attractive since it requires little technical expertise and incurs no training costs.

That said, the choice of LLM had a large influence on accuracy. GPT-4, Claude-1.3, and Claude-2 performed best and outperformed OpenAI's prior-generation LLM, GPT-3.5. But, at the time of writing, GPT-4's performance improvement over GPT-3.5 incurs a 20-fold cost increase. Conversely, Anthropic's Claude models are currently free, while matching or exceeding GPT-4's performance, making them the clear winners in performance-cost tradeoff.

MII coding might be a best-case scenario for coding open-text data. Public opinion surveys target the general public and do not assume deep political expertise. The most important issue on many respondents' minds is likely to be widely discussed and therefore well-represented in the web, book, and Wikipedia data that many LLMs are trained on. Nonetheless, it is likely that many other open-text coding tasks in social science share these characteristics.

It is still an open question how many other tasks the best LLMs can assist social scientists with. We do not know how they perform when coding the *sentiment* of open-text data. We also do not know how well they can code open-text data when it requires specialized domain knowledge, for instance, coding job titles into common occupational or industry schema. It is worth learning whether LLMs can remove the need for these arduous tasks or whether they will always need a trained human coder. We should also investigate whether LLMs can group text data into common topics. At present, this requires structural topic modeling. But, if possible, we could use LLMs both to write a coding schema and to code our text.

Despite their promise, researchers must exercise caution with LLMs. Like any computer-aided method, we must manually validate the results. Likewise, the statistical adage, "garbage in, garbage out" remains relevant—if survey data collection or coding schema is poor, results will mirror this. Non-English data may pose challenges too. While some LLMs train on multilingual data, English is hugely over-represented which may explain LLMs' lower performance on some non-English tasks (Lai et al., 2023). Researchers must also consider disclosure risks when using cloud-based LLMs. We discuss these issues in appendix H.

Scholars have also raised concerns about being "trapped" in a closed-source AI ecosystem (Spirling, 2023). This is reasonable because companies like OpenAI, Anthropic, and Google have their own interests and may tailor their models to maximize profitability. However, in our case, switching APIs was as simple as changing 15 lines of code, so adjusting pipelines to use open-source models like Llama-2 is already straightforward.[5]

Replicability is another potential concern. This is most true for closed-source models that might change or disappear without warning. This is a clear limitation of relying on proprietary models. Yet, even in the worst-case scenario, we can still detect changes by relabeling a sample of previous responses. Further, this problem applies to human labeling too. For example, a research assistant might make different judgments at different points in time. As such, only future open-source models offer the promise of full replicability, given that no open-source model currently performs acceptably on this task.

LLMs' ability to code open-text responses creates new research possibilities. First, it allows researchers to cheaply apply new schemas to existing data. For example, we could develop a comparative coding of issue agendas, regardless of the schemas used by the original studies. Second, it makes open-ended survey questions a more attractive research method. Previously, their use was limited by cost even where they might even be methodologically better. Our findings remove much of this burden. Allowing respondents to express their views without the constraint of closed categories might even lead to discoveries we could not have anticipated.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study is supported by Economic and Social Research Council (ES/S015671/1 and ES/S012435/1).

Disclosure

The views expressed herein are those of the authors and do not reflect the position of the United States Military Academy, the Department of the Army, or the Department of Defense. We are grateful to colleagues at the 2023 Social Research Association conference for providing comments on an earlier draft.

ORCID iDs

Jonathan Mellon https://orcid.org/0000-0001-6754-203X
Jack Bailey https://orcid.org/0000-0001-8517-5018
Ralph Scott https://orcid.org/0000-0002-9498-8034
Phillip Schmedeman https://orcid.org/0000-0002-2463-3879

Supplemental Material

Supplemental material for this article is available online.

Notes

1. Structural topic models offer one way to categorize open-text data (Roberts et al., 2014) but can produce categories that do not reflect the theory of concern or that are linguistically distinct but substantively similar.
2. Labeling unique text responses. Supervised algorithms were trained on 1000 randomly selected cases.
3. 81,266 open-text responses would be costly for a human to relabel, so they labeled a random sample of 1000 responses from each test set. Llama-2's high cost led us to follow this approach for that LLM as well.
4. BESIP provides 50- and 13-category codes to users (see Appendix A for the mapping).
5. Our replication package provides code for researchers to use or adapt the LLM and supervised approaches in this paper.

References

Amin M, Cambria E and Schuller B (2023) *Will Affective Computing Emerge from Foundation Models and General AI? A First Evaluation on ChatGPT*. arXiv. https://arxiv.org/abs/2303.03186

Bevan S, Jennings W and Wlezien C (2016) An analysis of the public's personal, national and EU issue priorities. *Journal of European Public Policy* 23(6): 871–887.

Brown T, Mann B, Ryder N, et al. (2020) *Language Models Are Few-Shot Learners*. arXiv. https://arxiv.org/abs/2005.14165

Dennison J (2019) A review of public issue salience: concepts, determinants and effects on voting. *Political Studies Review* 17(4): 436–446.

Ferrario B and Stantcheva S (2022) Eliciting people's first-order concerns: text analysis of open-ended survey questions. *SSRN Electronic Journal* 112: 163–169.

Fieldhouse E, Green J, Evans G, et al. (2022) *British Election Study Internet Panel Waves*. 1–23.

Gu YH, Piao X, Yin H, et al. (2022) Domain-specific language model pre-training for Korean tax law classification. *IEEE Access* 10: 46342–46353.

He Z and Schonlau M (2020) Coding text answers to open-ended questions: human coders and statistical learning algorithms make similar mistakes. *Methods, Data, Analyses* 15(1): 103–120.

Huang F, Kwak H and An J (2023) Is ChatGPT better than human annotators? Potential and limitations of ChatGPT in explaining implicit hate speech. In: Companion Proceedings of the ACM Web Conference 2023, pp. 294–297. https://arxiv.org/abs/2302.07736

Jurka TP, Collingwood L, Boydstun AE, et al. (2013) RTextTools: A supervised learning package for text classification. *The R Journal* 5(1): 6–12. DOI: 10.32614/RJ-2013-001.

Kaplan J, McCandlish S, Tom H, et al. (2020) *Scaling Laws for Neural Language Models*. arXiv. https://arxiv.org/abs/2001.08361

Lai V, Ngo N, Veyseh A, et al. (2023) *ChatGPT beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning*. arXiv. https://arxiv.org/abs/2304.05613

LeCun Y, Bengio Y and Hinton G (2015) Deep learning. *Nature* 521(7553): 436–444.

Liu Y, Ott M, Goyal N, et al. (2019) *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. arXiv. https://arxiv.org/abs/1907.11692

Ornstein J, Blasingame E and Truscott J (2022) *How to Train Your Stochastic Parrot: Large Language Models for Political Texts*. github.io.

Ramos L and Chang O (2023) Sentiment analysis of Russia-Ukraine conflict tweets using RoBERTa. *Uniciencia* 37(1): 1–11.

Roberts ME, Stewart BM, Tingley D, et al. (2014) Structural topic models for open-ended survey responses. *American Journal of Political Science* 58(4): 1064–1082.

Sanh V, Chaumond J and Wolf T (2020) *Model Card for DistilRoBERTa Base*. Available at: https://github.com/n-yassir/distilroberta-base/blob/main.

Schuldt JP and Roh S (2014) Media frames and cognitive accessibility: what do 'Global Warming' and 'Climate Change' evoke in partisan minds? *Environmental Communication* 8(4): 529–548.

Spirling A (2023) Why open-source generative AI models are an ethical way forward for science. *Nature* 616(7957): 413.