Benchmarking AI and Human Text Classifications in the Context of Newspaper Frames: A Multi-Label LLM Classification Note

Alexander Tripp*

October 9, 2024

Keywords: text as data, natural language processing, statistical analysis of texts

Abstract

Amid the explosion of research using artificial intelligence and Large Language Models, I compare the multi-label classification abilities of AI models (ChatGPT and Claude) and undergraduate coders using Spanish, immigration-focused, newspaper articles. Deriving my codebook from existing research on immigration-focused media narratives, I prompt LLMs to label articles by either an 8-label schema—directly analogous to the assignment of the undergraduate coders—or a 4-label schema—a more generalized approach aggregating the 8 labels into broader thematic categories. With undergraduate coders as the benchmark, I determine that a Few Shot ChatGPT 4o model with 8 labels is currently the most reliable AI classifier, with a hamming loss of 0.1. I emphasize two additional takeaways: 1) AI models using 8 labels outperform their 4 label counterparts and 2) AI models bias toward false positives. This article is optimistic about the abilities of AI models to supplement human coders, although human coding still provides a valuable benchmark for text classification tasks. I encourage authors and publishers to consider systematic meta-analyses of these tasks so as to better understand the generalizability of AI performance across different contexts and tasks.

*PhD Candidate, Vanderbilt University, Nashville, TN, USA.
Email: alexander.j.tripp@vanderbilt.edu. ORCID: https://orcid.org/0009-0007-2574-7038.

1 Introduction

One might imagine a coding quality hierarchy. Ideally, researchers would code their own corpora, leveraging their original ideas and expertise. However, this ideal introduces bias, as a researcher's familiarity with her hypotheses could influence her results. As such, researchers instead employ coders to assess their designs with greater integrity. Direct peers or graduate students might be the next best option as coders, followed by undergraduates. This coding quality hierarchy reflects decreasing familiarity with the project's theoretical underpinnings and less research experience, alongside less bias from knowledge of the project's hypotheses. Undergraduates, while most readily available, sit at the bottom of this hierarchy due to their limited research skill set. This note attempts to determine where generative AI models place in this coding quality hierarchy as compared to undergraduates in a multi-label classification context.

Text classification has recently seen a large increase in popularity across the social sciences due to the emergence of powerful generative AIs like OpenAI's ChatGPT and Anthropic's Claude. Recent research continues to innovate on this frontier, with newer models employing strategies such as pairwise comparisons of latent traits (Wu et al. 2023) or built-in chains of thought (Wei et al. 2022) to improve performance. Across a range of settings, large language models (LLMs) perform comparably to humans.[1]

I assess the text classification abilities of three popular LLMs—ChatGPT 4 Turbo, ChatGPT 4o, and Claude Sonnet 3.5—across two primary characteristics: 4 vs. 8 labels and Zero vs. One vs. Few Shot prompts. Existing work shows meaningful differences in performance based on the provision of accurate examples (Chae and Davidson 2024), highlighting the need to evaluate these prompts under varying specifications. I compare these LLMs in particular based on their analysis of Spanish texts, incorporation of large context windows, and ease of access to researchers. This project uses Colombian newspapers covering Venezuelan immigration due to my substantive research interests.

My case provides a hard test for the conclusion of existing research—that LLMs are useful tools comparable to humans—in two ways. First, the identification of news frames requires a thoroughly validated coding schema, because they represent subtle, implied cues rather than concrete information, like the identification of sentiment or topic. Second, I use non-English texts. Traditionally, cutting-edge LLMs have been designed using English. English is the highest resourced language, meaning that it has high quality digitized text on which models can train (Nicholas and Bhatia 2023). Languages with less digitized text are less likely to be supported by these models and exhibit weaker performance. As a result, non-English researchers have been excluded from these advances.

I find that a Few Shot ChatGPT 4o model with 8 labels performs best, showing an average hamming loss of 0.10 across labels. Digging deeper, I uncover that 1) AI models using 8 labels are generally more reliable than their 4 label counterparts and 2) AI models are biased toward false positive errors.

My work contributes to scholarship on the text classification abilities of generative AI models by exploring how their output differs from undergraduates and providing a blueprint for the efficient, robust usage of LLMs by applied researchers in non-English contexts.[2] While I do not test the generalizability of this work to other prompts and/or languages, I am optimistic that it extends to

[1] I use "AI models" and "LLMs" interchangeably.

[2] I include detailed information regarding my prompt and model specifications in the appendix. See Appendix B for more information on my API calls, Appendix F for more information on how my adaptations of my codebook to AI prompts improved from my first to latest prompt, and Appendix J for examples of all my prompt specifications.

other settings based on it being a harder test. On this point, I thoroughly document my prompts, code, and output, providing a template for other researchers to apply this method. I encourage authors and publishers to consider regular, broad-ranging meta-analyses of AI performance across a variety of coding tasks and contexts, as it is important to archive progress in this booming field of research in an effort to understand the generalizability of these powerful tools.

2 Setting the Scene

2.1 The Undergraduates

In Spring 2024, I worked with a team of three undergraduate researchers—chosen based on their advanced Spanish skills—to classify a random sample of 600 Colombian newspaper articles focused on Venezuelan immigration from 2012-2023 using a set of predefined labels, each of which outline substantive frames. The undergraduates were instructed to categorize each article into all applicable labels, instead of choosing a single label that fits best.

At their broadest levels of abstraction, these labels are **humanitarianism**, **threat**, **benefit**, and **policy & integration**, as seen in Table 1.[3] The codebook subdivided each of the 4 concepts into 8 smaller, more discrete categories to provide for more fine-grained descriptive data. See Table 2 for more information on the 8 label schema. Rather than code for the subject of these newspaper articles, my project identifies how newspaper articles frame—or, package information in an effort to direct how people think about—immigration (Chong and Druckman 2007; Nicholls and Culpepper 2021).

Table 1: 4 Label Framework for Coding Articles

Label	**Description**
1: Humanitarian	Perceptions that immigrants are vulnerable, fellow humans fleeing harsh conditions
2: Threat	Fears, broadly construed, that immigration will worsen the host country's economy, health outcomes, and/or safety
3: Benefit	Perceptions that immigration can improve local and national economies
4: Policy & Integration	Neutral, technical reports about immigration policies and government action

The undergraduates parsed each newspaper article for the 8 labels, coding articles that matched as 1s and those that did not match as 0s. If the primary topic of the article was not immigration-related, the undergraduates were instructed to set all labels to 0. The output of this coding exercise is a dataset where each row represents a newspaper article and each column represents a label, with 1s if the label occurred in the article and 0 if it did not.

Since this task involved three independent coders, it inevitably produced disagreement. In this case, because each of the three undergraduates read and coded many of the same articles, there were

[3] I generate these theoretical labels in my other work.

Table 2: 8 Label Framework for Coding Articles

Label	Description
1: Humanitarian- Vulnerability	Perceptions that immigrants are uniquely vulnerable, fellow humans (Feldman and Steenbergen 2001)
2: Humanitarian- Refugee	Perceptions that immigrants are like refugees fleeing harsh, impossible conditions (Fraser and Murakami 2022)
3: Threat- Disease	Fears that immigration will bring disease (Kam and Estes 2016)
4: Threat- Economic	Fears that immigration will lead to economic competition locally and nationally (Hainmueller and Hiscox 2010)
5: Threat- Instability	Fears, broadly construed, that immigration will cause social instability
6: Threat- Violence	Fears that immigration will lead to increased crime and violence (Sniderman, Hagendoorn, and Prior 2004)
7: Benefit	Perceptions that immigration can improve local and national economies
8: Policy & Integration	Neutral, technical reports about immigration policies and government action

multiple opportunities for disagreement. Reassuringly, only 3% of all label-article cells disagreed. To resolve these disagreements for the purposes of assigning *true* labels, I first join these three output datasets together based on agreed values. Then, for cells with disagreement, I default to responses from the two coders with the highest intercoder reliability metrics.[4] I first fill in disagreed values with output from the coder in this pair with the most research experience and Spanish fluency. Then, I fill in additional disagreed cells with the second coder in this pair. For any remaining disagreed cells, I use the output from my third coder.

2.2 The AI Models

I compare the output of my undergraduate coders to three generative AI models: OpenAI's ChatGPT (models: 4 Turbo and 4o) and Anthropic's Claude (model: Sonnet 3.5). I employ these AI models within my analysis to uncover 1) which is the best classifier, 2) whether it is a sufficient replacement for human coders, and 3) how its outputs might differ from other LLMs and undergraduates. My work contributes to recent research (e.g., Brown et al. 2020 and Pangakis, Wolken, and Fasching 2023) by assessing LLMs in a non-English context, analyzing how disaggregated information in prompts

[4]See Appendix A for more information on intercoder reliability. As a robustness check, in Appendix E, I rerun my analysis and exclude disagreed values, finding similar results.

may improve their performance, and assessing how they systematically differ from undergraduate coders.

I use these models specifically, because they 1) were trained on Spanish-language corpora, 2) have large context windows, and 3) are comparatively accessible. The first two characteristics are necessary for comparability between undergraduates and AI, as both sets of coders must understand the language of the texts and make their decisions using the entirety of the data.

I focus my classification exercises in a non-English context. Recent attempts to train LLMs on multiple languages at once—dubbed multilingual language models—have expanded language support (Nicholas and Bhatia 2023). These multilingual AI models are increasingly proficient in less resourced languages, diversifying the field by allowing scholars from non-English backgrounds to employ cutting-edge models in their native languages. ChatGPT and Claude respond well to Spanish language prompts,[5] so my project explores their text classification capabilities as compared to undergraduates that also know Spanish.

Furthermore, each model has a large context window, meaning that it can fully incorporate information from longer prompts. Some newspaper articles in my dataset are over 10,000 tokens (over 7,500 words), and many existing LLMs would not be able to process these larger texts due to their smaller context windows.[6] Since undergraduate coders can read and incorporate all of the text into their decision-making, the classification models in question should do the same.

Finally, these ChatGPT and Claude models do not require any large downloads, computational power, or fine-tuning on the side of the researchers, making them more accessible to researchers less familiar with LLMs. I gear the insights and analysis of this paper toward applied researchers using LLMs as a tool, rather than those looking to build or fine-tune their own models.[7]

To create the prompts that I send to each AI model, I adapt the codebook given to my undergraduate research team and follow the best practices in prompt engineering.[8] I then combine the prompt with each newspaper article and use the relevant API to send my request.[9] Throughout the course of this project, I iteratively refined and validated my prompts to both ensure that they match the content of the codebook and to maximize model performance.[10]

I now discuss the trade-offs of using LLMs versus human coders. Human coding is highly valuable, because these coders 1) can accurately classify texts given some set of instructions, asking clarification questions when needed, 2) can explain their decision-making process, and 3) are often students gaining useful research experience and training to be the next generation of social scientists, constituting a social good. On the other hand, human coding is expensive and slow, and many researchers face a lack of institutional resources or short time horizons that leave them unable to utilize human coding. The cost and time differentials between my human and AI coders are large, but those between AI models are usually not. Human coders necessitate either course credit or hourly pay, often by semester. For the coding of 600 newspaper articles, my human coders took about 180 hours throughout one semester, while my AI models took roughly 15 minutes and cost about $10. However, I spent many hours revising the prompts before I was satisfied with the model

[5]Spanish is less resourced than English, though it is still highly digitized and broadly used. Thus, this provides an easier test for the capabilities of these models in non-English languages than, say, Quechua or Welsh.

[6]For example, the context window of the base BERT model is 512 tokens.

[7]However, recent research questions the replicability of these closed-source models (Spirling 2023).

[8]See `https://platform.openai.com/docs/guides/prompt-engineering` for ChatGPT-specific best practices, and `https://docs.anthropic.com/en/docs/build-with-claude/prompt-engineering/overview` for Claude-specific best practices. Also see Liu et al. 2023.

[9]In this API call, I set the temperature—a hyperparameter ranging from 0 to 2 where lower values indicate less randomness in responses—of each model to 0.

[10]See Appendix F.

results. Over the entire course of this project, I spent over $1,000 running and rerunning these LLMs, so the costs can quickly add up if one is not disciplined in their analytical strategy.[11]

3 Comparative Analysis

I evaluate model performance by comparing the output of my AI classification models to that of my undergraduate coders.[12] In my main analyses, I use hamming loss to determine the reliability of these models. Hamming loss shows the fraction of incorrectly predicted labels to the total number of labels in a multi-label setting, providing intuitive model comparisons based on the proportion of correct classifications in multi-label tasks (Pal, Selvakumar, and Sankarasubbu 2020). Models with hamming loss values closer to zero are the better performers, as they predict fewer incorrect labels as a proportion of all labels. So, a hamming loss of 0.25 would indicate that 25% of labels in an article were incorrectly classified compared to the standard. My analysis includes 18 LLM specifications in total: models from Claude Sonnet 3.5, ChatGPT 4 Turbo, or ChatGPT 4o, models with 4 labels or 8 labels, and Zero Shot, One Shot, or Few Shot models.

While I write my prompts in English, I provide examples for the One Shot and Few Shot models in Spanish.[13] I find no notable differences across model metrics for models run with fully Spanish prompts versus those with instructions in English and examples in Spanish.[14] This provides some evidence that the language of the prompt instructions is not hugely influential, making it easier for researchers to conduct analyses for languages in which they are not fluent.

3.1 How do LLMs Compare to Undergraduate Coders?

In Figure 1, I contrast the hamming loss pooled across labels for each AI model specification as compared to my undergraduate coders.[15] The graph is ordered so that better performing models are at the top and worse performing models are at the bottom.

I find that the Few Shot ChatGPT 4o model with 8 labels has the lowest average hamming loss across LLMs.[16] The Few Shot ChatGPT 4o model with 8 labels has an average hamming loss of 0.10, incorrectly predicting 10% of the labels as compared to the benchmark of my undergraduate coders. Compared to recent literature, a hamming loss of 0.10 is fairly competitive with other state-of-the-art LLMs, conditional on the fact that media frames are more difficult to systematically identify (El Rifai, Al Qadi, and Elnagar 2022; Nicholls and Culpepper 2021).

ChatGPT 4 Turbo appears slightly worse than 4o across the board, and it is considerably more expensive to run than 4o and Claude Sonnet 3.5. Sonnet 3.5 performs worse than 4o and 4 Turbo, though it shows less variation in its performance across the 4 and 8 labels specifications. Additionally, its Few Shot specification does not improve over its Zero and One Shot models to the same degree as ChatGPT 4o and 4 Turbo.

[11] Note that Claude Sonnet 3.5 and ChatGPT 4o are much less expensive than ChatGPT 4 Turbo.

[12] See Appendix C for an in-depth discussion of the costs involved in running these models.

[13] See Appendix J for examples of each prompt and information on how I constructed them.

[14] Specifically, I rerun a Few Shot ChatGPT 4o model with an 8 label specification using a prompt fully in Spanish. See Appendix J.

[15] In Figure 4 of Appendix H, I expand this analysis out to show the results of these models across each label. In Appendix G, I perform a similar analysis using my own coding of these newspaper articles as a benchmark.

[16] See Table 3 in Appendix D for other metrics used to compare the LLM outputs with reference to just the undergraduate codings. Also see Figures 2 through 15 in Appendix H for figures visualizing the results for each of these model metrics.

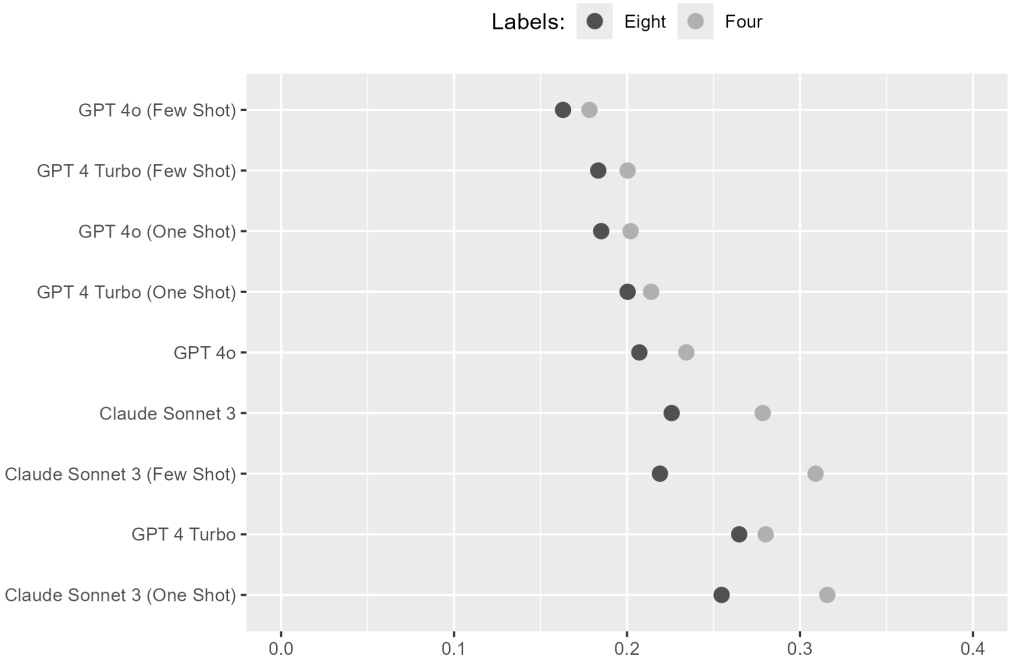

Figure 1: Few Shot ChatGPT 4o model with 8 labels performs best

I break this analysis down by label and label specification in Figure 2 below, where values closer to zero again indicate better performance. For each label, the AI models vary from one another by hamming loss metrics of about .15 on average, and the performance for some labels is much better than others. For example, in the 8 label specification, LLMs identify the Disease Threat, Economic Threat, and Economic Benefit labels quite well—with hamming loss scores of less than 0.10—while they struggle to identify other labels, such as Instability and P & I. In the 4 label specification, the Benefit label performs best, and the P & I label shows the greatest variation. In the next two sections, I explore what might be contributing to these differences in LLM performance via prompt structure and a potential false positive bias.

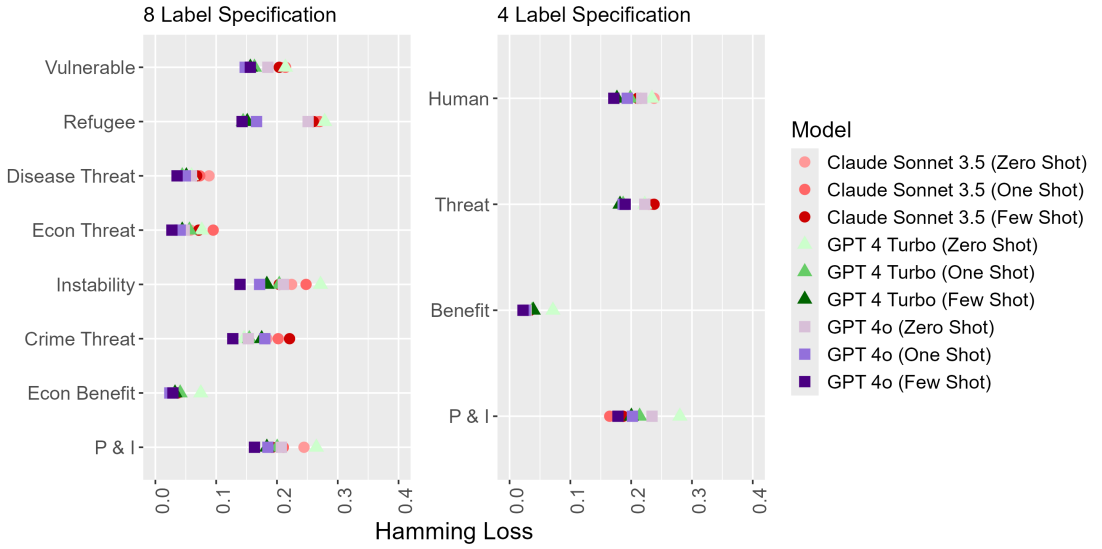

Figure 2: LLM performance substantially varies across label

3.2 How does Prompt Structure Affect LLM Outputs?

In Figure 1, the 8 label models outperform their 4 label counterparts. This begs the question of how prompts—and more specifically, the structure of the information in the prompts—influence LLM output.

The 8 label models may perform better because they provide more clearly structured information to the AI models than the 4 label prompts.[17] With this in mind, it is important to note that the four label specifications contain the same information and examples as the 8 label specifications, barring slight changes to the names of the labels.[18]The AI models may exploit the more structured explanations of each label, using this information to classify articles more akin to the undergraduate coders than the aggregated information found in the 4 label prompts.

Moreover, the One Shot and Few Shot models—by providing correctly identified examples of each label from the undergraduate coders—display better metrics than the zero-shot models, in line with scholarly expectations that more detailed sets of information improve AI performance across the board (Chae and Davidson 2024).

These findings have notable implications for the best practices of text classification tasks for both undergraduates and AI models. There exists a trade-off for the instruction of undergraduate coders in classification: codebooks must be balanced in terms of length, structure, and detail, as undergraduates will 1) become fatigued if the instructions are too long or 2) be unclear as to their assignment if the instructions are ill-defined. I disaggregate my original conception of four overarching frames into eight more specific frames in my codebook to reflect a balance in the level of

[17]They also more closely resemble the original task of the author and undergraduates.

[18]For example, Benefit in the 4 label AI prompt and Econ Benefit in the 8 label AI prompt are the exact same, as they both correspond to economic benefits. On the other hand, Human in the 4 label specification contains the same information as Vulnerable and Refugee in the 8 label specification.

detail, structure, and cognitive load. Adapting this codebook to AI prompts thus provides additional evidence that more structured information improves text classification performance for AI models. Writ large, one might be optimistic about using Few Shot LLMs to supplement undergraduate codings.

3.3 LLM Classifiers are Prone to False Positives

I now explore how the output of LLMs might be systematically different from undergraduate coders by focusing on LLMs' seeming bias toward false positives. Those models that display the worst performance metrics—in this case, higher values for hamming loss—simultaneously exhibit the false positive bias more egregiously (e.g., the Zero Shot ChatGPT 4 Turbo model).

To better depict this trend, I display the proportion of articles that undergraduates coded as 1s subtracted from the proportion of articles that LLMs coded as 1s in Figure 3 for both the 4 and 8 label specifications. A value of zero on these graphs indicates no false positives by LLMs in comparison to undergraduate coders, whereas higher positive values indicate more false positive errors.[19] On average, the 8 label models do not show this false positive bias as strongly as the 4 label models, but the general trend remains that the performance of AI models in this analysis seem to be correlated with their level of false positive bias.

Figure 3: LLMs are biased toward false positives

In order to better understand why these false positives may be occurring, I conducted close readings of several newspaper articles coded as 1s by LLMs and 0 by undergraduates.[20] Anecdotally,

[19]Note that in the 8 label specification, undergraduates coded articles as 1s 10% of the time; in the 4 label specification, 23% of the time. For the 4 label undergraduate model, I deterministically aggregate each of their 8 label codings into the respective 4 label categories. In their original task, undergraduates did not code articles into the 4 label schema.

[20]See Appendix L for more information on these close readings. This appendix also explores correlation matrices

I find that the LLMs seemed to pick up on more subtle label cues than the human coders did, but they had trouble differentiating certain labels, such as the Vulnerable and Refugee labels in the 8 label specification. Moreover, the LLMs were more likely to make multiple coding errors within the same article, as opposed to mistakenly coding one label and getting the others correct.

While existing literature notes the problematic, black-box nature of LLMs (see Bisbee et al. 2024 and Spirling 2023), I am unaware of any existing work that explains this false positive bias. [21] This systematic bias thus constitutes an area of future research.[22]

4 Discussion and Conclusion

In this note, I analyze the reliability of LLMs in comparison to undergraduates, providing evidence that a Few Shot ChatGPT 4o model with 8 labels is the most reliable classifier. I offer two additional takeaways to practitioners and the burgeoning literature on LLM classifiers. First, LLMs using more structured prompts—those with their information broken down into more discrete parts—perform better than those using more general prompts. Second, these LLMs are biased toward false positive errors.

Use cases for social science applications of LLMs and text classification models vary in their difficulty and complexity. While AI models have recently approached human abilities on some tasks that are easier for humans, these models still struggle in other, more difficult applications. This project represents one of those more difficult applications, showing that the best performing AI model misclassifies labels about 10% more often than undergraduate coders. Despite the systematic biases of these models, they only lag slightly behind the undergraduate coders on the more complicated task of identifying newspaper frames. With this in mind, I place LLMs near the same level—or perhaps, slightly lower—than undergraduates on the coding quality hierarchy.

While these figures are likely to change as AI models improve, this article provides a framework for assessment of their performance as compared to both undergraduates and researchers. In this light, I encourage publishers and researchers to both incentivize and conduct regular tests of LLM performance across a variety of social science tasks. A centralized meta-analysis hub, for instance, might aggregate research such as this and contribute more broadly to our understandings of the generalizability of LLMs across languages and tasks, as well as the reliability and replicability of their results over time.

Acknowledgements I would like to thank James Bisbee, Joshua Clinton, Cindy Kam, Brenton Kenkel, and Jennifer Larson for their helpful feedback on this project.

Funding Statement None.

Disclosure Statement The authors report there are no competing interests to declare.

Data Availability Statement Data will be made available upon acceptance of the article.

for articles coded by undergraduates and the Few Shot ChatGPT 4o model.

[21] It may be possible to overcome these issues with fine-tuned models or different hyperparameters.

[22] For an exploration into what predicts alignment between undergraduate and AI coders, see Appendix I.

References

Bisbee, James et al. (2024). "Synthetic Replacements for Human Survey Data? The Perils of Large Language Models". In: *Political Analysis*, pp. 1–16.

Brown, Tom et al. (2020). "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems* 33, pp. 1877–1901.

Chae, Youngjin and Thomas Davidson (2024). "Large Language Models for Text Classification: From Zero-Shot Learning to Instruction-Tuning". In: *SocArXiv*.

Chong, Dennis and James N. Druckman (2007). "Framing Theory". In: *Annual Review of Political Science* 10.1, pp. 103–126.

El Rifai, Hozayfa, Leen Al Qadi, and Ashraf Elnagar (2022). "Arabic Text Classification: The Need for Multi-Labeling Systems". In: *Neural Computing and Applications* 34.2, pp. 1135–1159.

Feldman, Stanley and Marco R Steenbergen (2001). "The humanitarian foundation of public support for social welfare". In: *American Journal of Political Science*, pp. 658–677.

Fraser, Nicholas A. R. and Go Murakami (2022). "The Role of Humanitarianism in Shaping Public Attitudes Toward Refugees". In: *Political Psychology* 43.2, pp. 255–275. ISSN: 1467-9221.

Hainmueller, Jens and Michael J Hiscox (2010). "Attitudes toward highly skilled and low-skilled immigration: Evidence from a survey experiment". In: *American Political Science Review* 104.1, pp. 61–84.

Kam, Cindy D. and Beth A. Estes (2016). "Disgust Sensitivity and Public Demand for Protection". In: *The Journal of Politics* 78.2, pp. 481–496. ISSN: 0022-3816. (Visited on 07/09/2024).

Liu, Pengfei et al. (2023). "Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing". In: *ACM Computing Surveys* 55.9, 195:1–195:35.

Nicholas, Gabriel and Aliya Bhatia (2023). "Lost in Translation: Large Language Models in Non-English Content Analysis". In: *Center for Democracy & Technology*. URL: `https://cdt.org/insights/lost-in-translation-large-language-models-in-non-english-content-analysis/`.

Nicholls, Tom and Pepper D. Culpepper (2021). "Computational Identification of Media Frames: Strengths, Weaknesses, and Opportunities". In: *Political Communication* 38.1-2, pp. 159–181.

Pal, Ankit, Muru Selvakumar, and Malaikannan Sankarasubbu (2020). "Multi-Label Text Classification using Attention-based Graph Neural Network". In: *Proceedings of the 12th International Conference on Agents and Artificial Intelligence*, pp. 494–505.

Pangakis, Nicholas, Samuel Wolken, and Neil Fasching (2023). *Automated Annotation with Generative AI Requires Validation*.

Rul, Céline Van den (2019). "[NLP] Basics: Measuring The Linguistic Complexity of Text". In: *Medium*. URL: `https://towardsdatascience.com/linguistic-complexity-measures-for-text-nlp-e4bf664bd660`.

Sniderman, Paul M., Louk Hagendoorn, and Markus Prior (Feb. 2004). "Predisposing Factors and Situational Triggers: Exclusionary Reactions to Immigrant Minorities". In: *American Political Science Review* 98.1, pp. 35–49. ISSN: 1537-5943, 0003-0554.

Spirling, Arthur (2023). "Why open-source generative AI models are an ethical way forward for science". In: *Nature* 616.7957, pp. 413–413.

Wei, Jason et al. (2022). "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models". In: *Advances in Neural Information Processing Systems* 35, pp. 24824–24837.

Wu, Patrick Y et al. (2023). "Large Language Models Can Be Used to Estimate the Latent Positions of Politicians". In: *arXiv*.

A Disagreements and Intercoder Reliability

Using the output of my three coders, I rely on the codings of my two coders with the highest intercoder reliability scores for those coding outputs that do not agree.[23] I measure intercoder reliability using Cohen's kappa. I calculate these scores for both the 8 label model and the 4 label model (by aggregating the 8 labels into their respective 4 label categories). The Cohen's kappa for coders 1 and 2 is 0.69 in the 8 label model and 0.71 in the 4 label model. These values indicate moderate reliability. Since each of my coders was randomly assigned to code half of all articles in the corpus, I utilize coder 3 to substitute for the values not coded by coders 1 and 2 (bringing the total articles coded by humans from 300 to 590). In the 8 label model, coder 3 has a Cohen's kappa score of 0.53 with coder 1 and 0.62 with coder 2. In the 4 label model, coder 3 has a Cohen's kappa score of 0.60 with coder 1 and 0.63 with coder 2. Each of these Cohen's kappa statistics are significant with p-values less than 0.05, meaning that the observed agreements are unlikely to result from random chance.

B API Calls

I program and send my calls to the ChatGPT and Anthropic APIs using the httr package in R. However, when first sending these API calls, I faced server timeout errors if the prompts exceeded around 4,000 characters. To get around this error, I attempted to increase the length of the timeout for both my R session and my call to ChatGPT/Claude, to no avail. Instead, I found that sending a very short initial prompt—in my case, I pasted "Nada." instead of a full-length newspaper article at the end of my prompt instructions—would jump-start the process. After this faux article jump-started the API calls, I faced no timeout errors regarding the length of later prompts, even if they well-exceeded the 4,000 character pseudo-limit.

In later iterations for the ChatGPT calls, I used its batching feature, which was more cost effective and often faster than calling the API from R.

C Costs

The start-up costs for running models through ChatGPT or Claude are moderate, and I intend to illuminate the necessary steps to perform these classification tasks through my discussion of API calls, prompts, and the replication code. After creating API accounts for the service of interest, one must then purchase $50-$200 of API credits and wait for a certain amount of time to pass (around 2 weeks) after account creation to access higher rate limits. After this point, researchers can send large API calls without issue.

Additionally, One Shot and Few Shot models require examples of the ground truth for training the models. Generating these examples is often a time-consuming process, especially if the researcher does not speak the necessary language(s). It took several hours to validate the examples provided by my undergraduate researchers alone, so it will likely take many more to identify them on one's own.

[23] Note that 97% of coding outputs are in agreement.

D Additional Model Statistics

In Table 3, I display the raw values for area under the curve (AUC; the accuracy of binary classifiers), recall (proportion of true 1s correctly identified by the model), F1 (harmonic mean of precision and recall), hamming loss, Intraclass Correlation Coefficient (ICC; the strength of coder agreement across model and human classification), and percentage of articles coded as 1s for each LLM in comparison to the undergraduate coders. Most models fall between 0.74 and 0.81 for AUC, 0.78 and 0.92 for recall, 0.85 and 0.94 for F1, 0.23 and 0.10 for hamming loss, and 0.31 and 0.55 for the ICC. I highlight each cell with the best model statistic value in dark gray.

Table 3: Model Comparison Metrics

Model	No. Labels	Prompt Type	AUC	Recall	F1	Hamming Loss	ICC	% 1s[a]
GPT 4 Turbo	8	Zero Shot	0.77	0.83	0.89	0.17	0.32	22
	4	Zero Shot	0.77	0.80	0.86	0.20	0.48	29
	8	One Shot	0.79	0.89	0.92	0.13	0.42	16
	4	One Shot	0.75	0.87	0.90	0.16	0.46	21
	8	Few Shot	0.77	0.90	0.93	0.12	0.43	15
	4	Few Shot	0.77	0.88	0.90	0.15	0.49	21
GPT 4o	8	Zero Shot	0.81	0.86	0.91	0.14	0.43	19
	4	Zero Shot	0.79	0.82	0.88	0.18	0.48	27
	8	One Shot	0.80	0.90	0.93	0.12	0.46	16
	4	One Shot	0.78	0.87	0.90	0.15	0.52	23
	8	Few Shot	0.78	0.92	0.94	0.10	0.48	13
	4	Few Shot	0.80	0.88	0.91	0.14	0.55	22
Claude Sonnet 3.5	8	Zero Shot	0.82	0.83	0.90	0.16	0.38	23
	4	Zero Shot	0.80	0.84	0.88	0.17	0.49	26
	8	One Shot	0.83	0.83	0.89	0.17	0.37	23
	4	One Shot	0.78	0.87	0.89	0.16	0.51	22
	8	Few Shot	0.82	0.85	0.90	0.16	0.39	21
	4	Few Shot	0.78	0.85	0.89	0.17	0.48	23

[a] Percentage of articles in the sample coded as 1s. The mean percentage of articles coded as 1s by humans is 10% for the 8 label models and 23% for the 4 label models. So, the best value for this metric would be that closest to the human coded values.

E Model Statistics without Disagreed Values

In Table 4, I display the AUC, recall, F1 score, hamming loss, ICC, and percentage of articles coded as 1s for each LLM in comparison to the undergraduate coders, excluding all articles for which the human coders disagreed in labeling. Most models fall between 0.75 and 0.85 for AUC, 0.80 and 0.94 for recall, 0.86 and 0.95 for F1 score, 0.21 and 0.09 for hamming loss, and 0.31 and 0.58 for the ICC. These model metrics are largely similar to the full dataset, and I still conclude that the Few Shot ChatGPT 4o with 8 labels is the best classification model based on its optimal scores for recall, F1 score, hamming loss, and percentage classified as 1s. However, note that the Claude Sonnet 3.5 models perform better with the disagreed values excluded across most metrics.

Table 4: Model Comparison Metrics without Disagreed Values

Model	No. Labels	Prompt Type	AUC	Recall	F1	Hamming Loss	ICC	% 1s[a]
GPT 4 Turbo	8	Zero Shot	0.79	0.85	0.91	0.15	0.34	20
	4	Zero Shot	0.78	0.82	0.88	0.19	0.40	27
	8	One Shot	0.79	0.91	0.93	0.11	0.43	14
	4	One Shot	0.76	0.89	0.91	0.14	0.48	19
	8	Few Shot	0.79	0.92	0.94	0.10	0.45	13
	4	Few Shot	0.78	0.90	0.92	0.13	0.51	19
GPT 4o	8	Zero Shot	0.80	0.88	0.92	0.13	0.43	17
	4	Zero Shot	0.79	0.85	0.90	0.16	0.50	24
	8	One Shot	0.80	0.91	0.94	0.10	0.48	14
	4	One Shot	0.79	0.89	0.91	0.14	0.53	21
	8	Few Shot	0.77	0.94	0.95	0.09	0.49	11
	4	Few Shot	0.80	0.90	0.92	0.12	0.58	20
Claude Sonnet 3.5	8	Zero Shot	0.83	0.85	0.91	0.15	0.38	20
	4	Zero Shot	0.81	0.85	0.90	0.16	0.52	24
	8	One Shot	0.83	0.85	0.91	0.15	0.38	20
	4	One Shot	0.80	0.88	0.91	0.14	0.55	21
	8	Few Shot	0.83	0.87	0.92	0.14	0.40	19
	4	Few Shot	0.79	0.87	0.90	0.15	0.51	22

[a] Percentage of articles in the sample coded as 1s. The mean percentage of articles coded as 1s by humans is 9% for the 8 label models and 16% for the 4 label models. So, the best value for this metric would be that closest to the human coded values.

F Prompt Comparisons and Explanations

In this section of the appendix, using a Zero Shot ChatGPT 4o model with an 8 label specification, I compare my prompts across two dimensions: earlier versus later versions of my prompts and the inclusion of explanations. My prompts went through a series of adjustments to improve performance over the course of this project (though most of these changes concerned the structure of the output rather than the substantive content of the labels). I compare the performance of my models across one of the earliest versions of my prompt and the latest version in Figures 4 and 5 along metrics for hamming loss and the proportion of articles classified as 1s. These graphs show that the latest version of my prompt has considerably improved the performance across the hamming loss and proportion classified as 1s metrics.

In addition to comparing the performance of my earlier versus later prompts, I analyze how requesting explanations for the model's output might improve performance. I find that asking the models to justify their responses slightly improves performance in the earlier prompt while reducing performance in the later prompt.

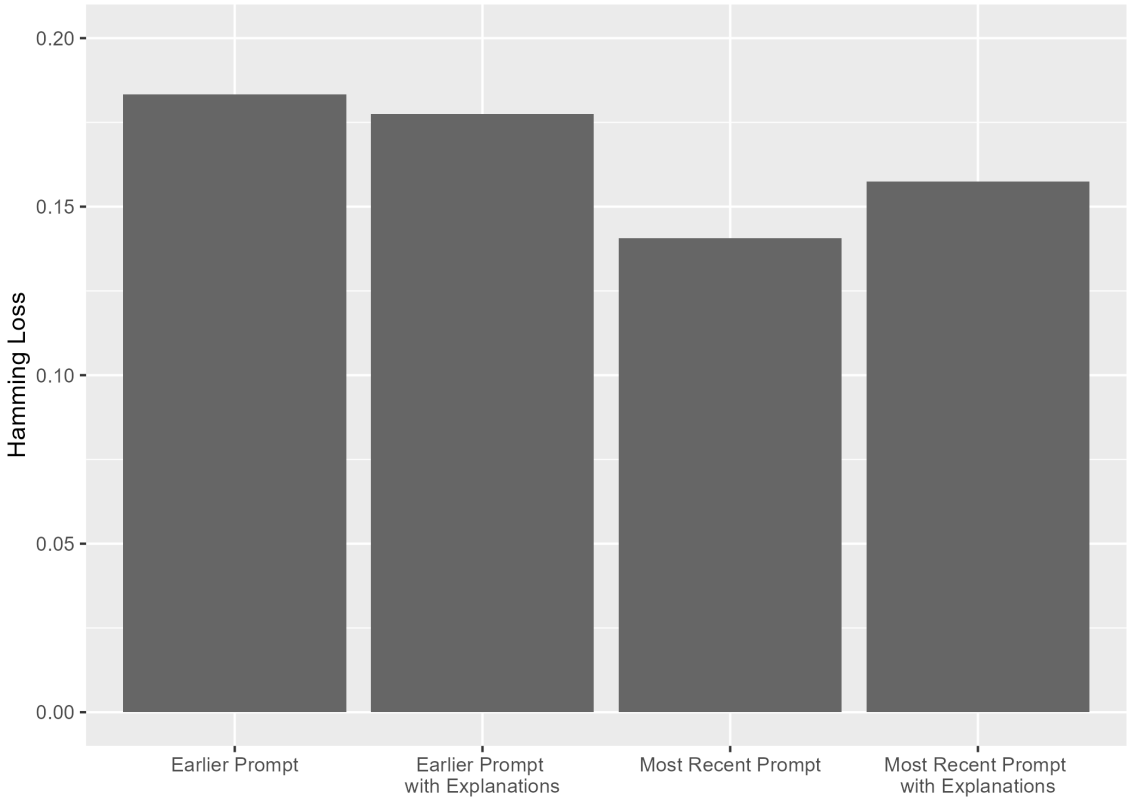

Figure 4: Hamming Loss by Prompt on ChatGPT 4o

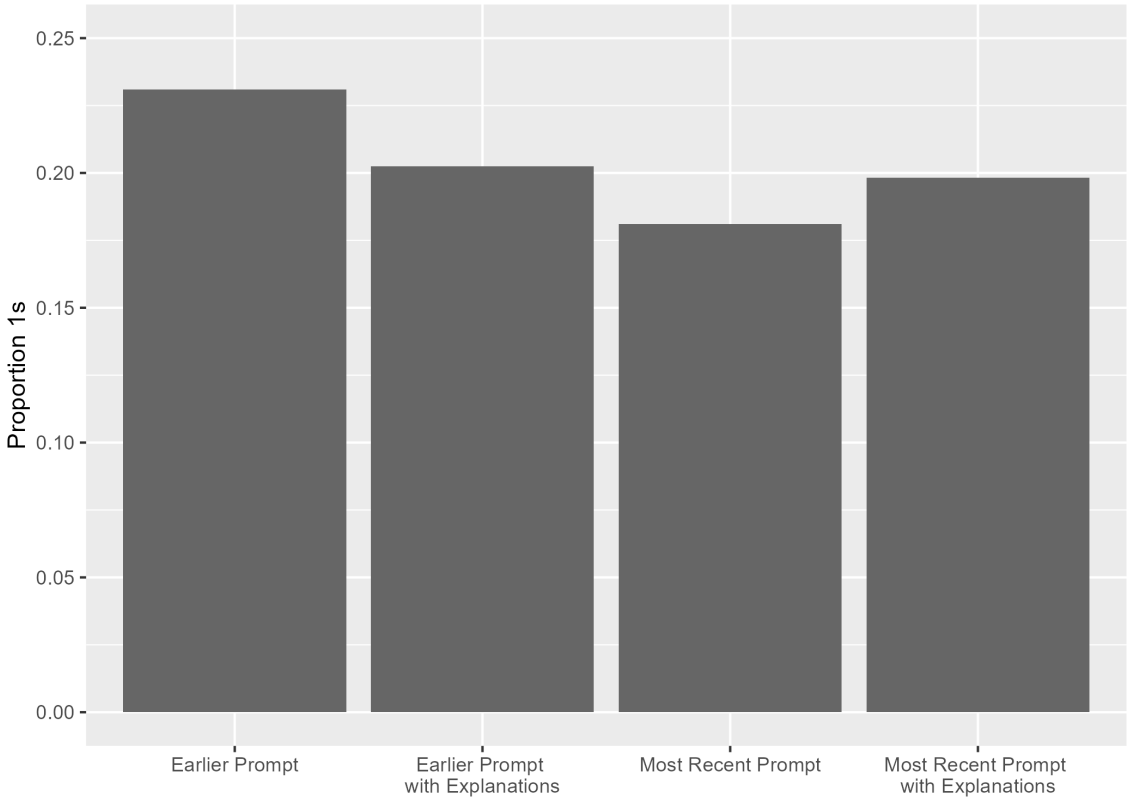

Figure 5: Proportion 1s by Prompt on ChatGPT 4o

G Comparison to the Author's Codings

To provide another point of comparison, I code the same set of newspaper articles following the undergraduate codebook and analyze the output of both undergraduates and LLMs with my codings as the benchmark. Figure 6 displays the hamming loss of each model in comparison to my codings, with the results pooled across labels. I find that the 8 label ChatGPT 4o model performs best. More generally, the 8 label models perform better than the 4 label models, and the Few Shot models perform better than the One and Zero Shot models.

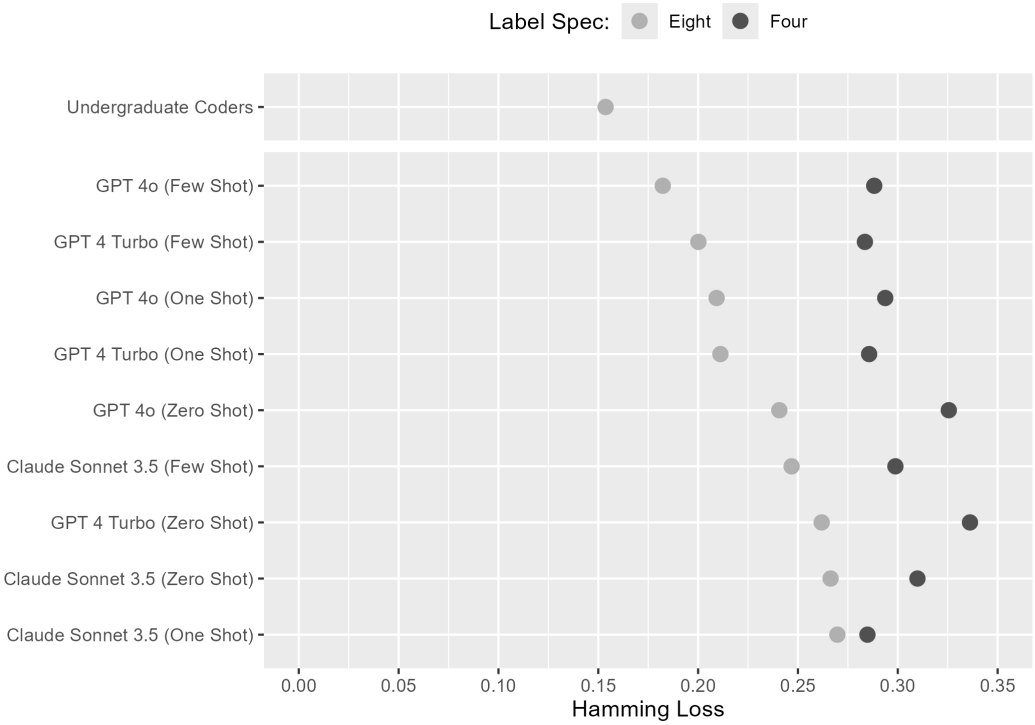

Figure 6: Few Shot ChatGPT 4o Model with 8 Labels Performs Best

Moving forward with the best performing model, I break out the results of undergraduates and ChatGPT 4o across labels in Figure 7. With my codings as a benchmark, the output of ChatGPT 4o only differs from that of the undergraduates by an average hamming loss of 0.03. This is strong evidence in support of using these LLMs as supplements or replacements to human coders.[24] This analysis comes to the same conclusions of my comparisons with undergraduates and LLMs—these AI models are performing at a similar level to undergraduate coders.

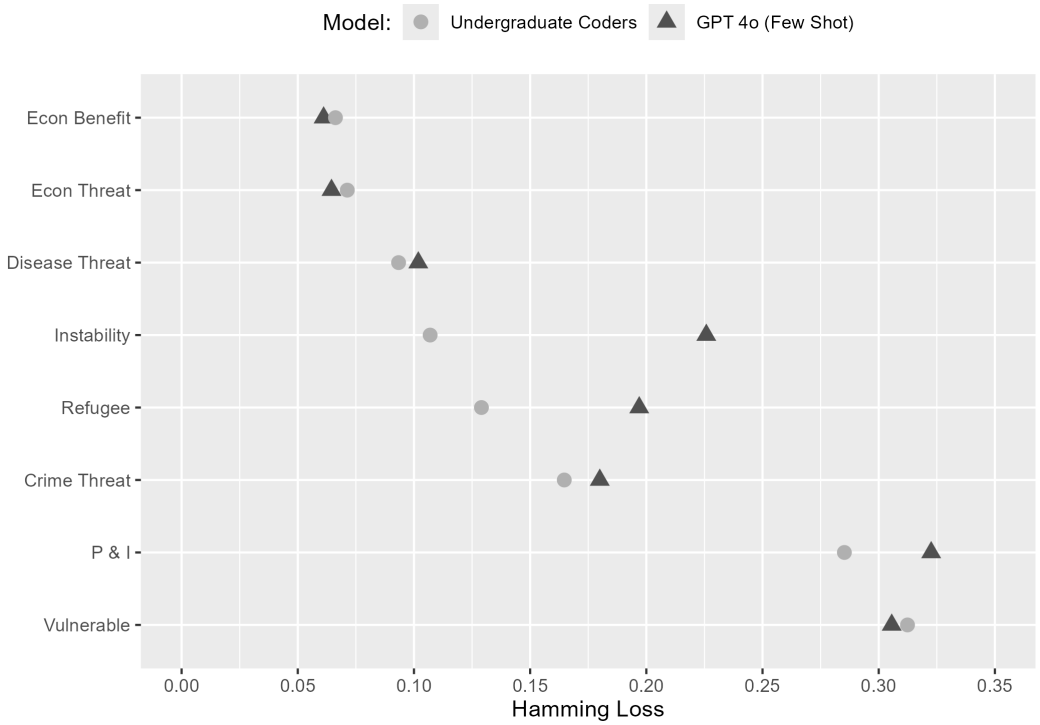

Figure 7: On average, ChatGPT 4o Performs 3% Worse than Undergraduates

[24]Though notice the variation across labels. In some, such as Econ Benefit and Threat, ChatGPT 4o outperforms the undergraduates, while in others, such as Instability and Refugee, ChatGPT 4o lags far behind the undergraduates.

H Supplementary Model Visualizations

In this section, I visualize how the LLMs compare across each label for a variety of model metrics, using the undergraduate codings as a baseline. I then display AUC, F1, ICC, and recall 1) across models, 2) across all labels, disaggregated by the 4 and 8 label specifications, and 3) the differences in each metric when going from a 4 to 8 label specification. Most of these graphs supplement the primary takeaways from the main text: the Few Shot GPT 4o model with 8 labels performs best, 8 label models outperform 4 label models, and most models are systematically different from human coders.

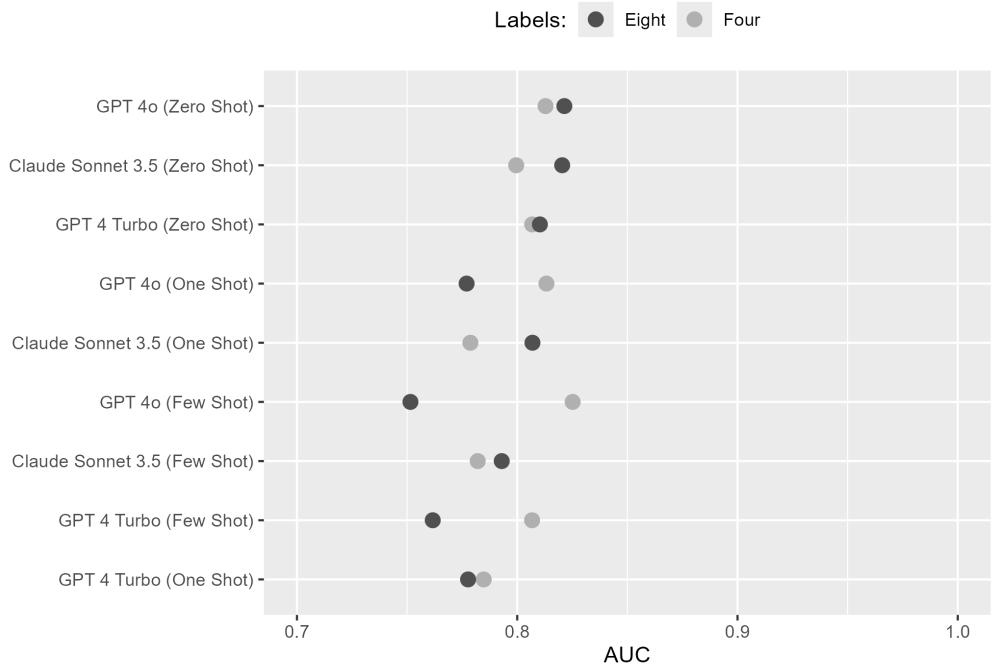

Figure 8: AUC Across Models

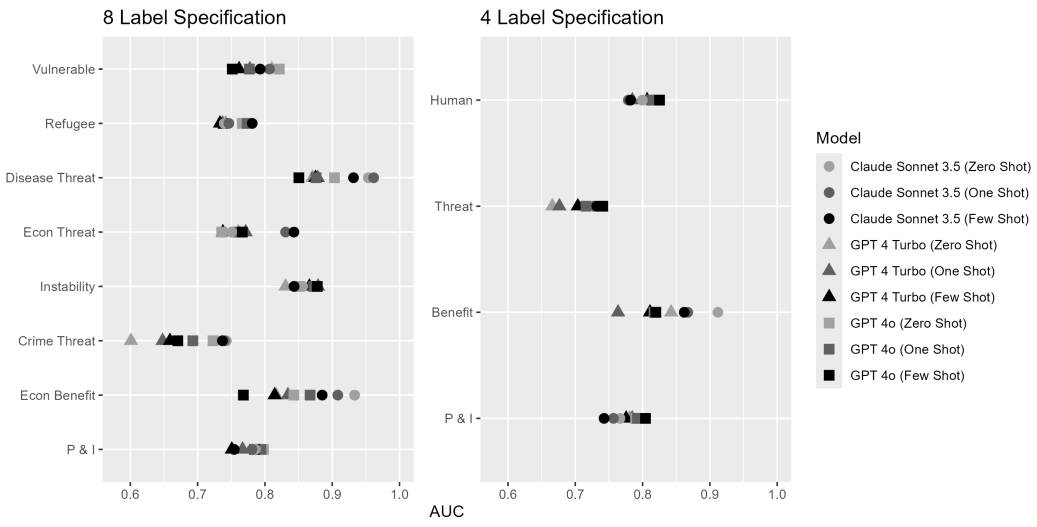

Figure 9: AUC Across Labels

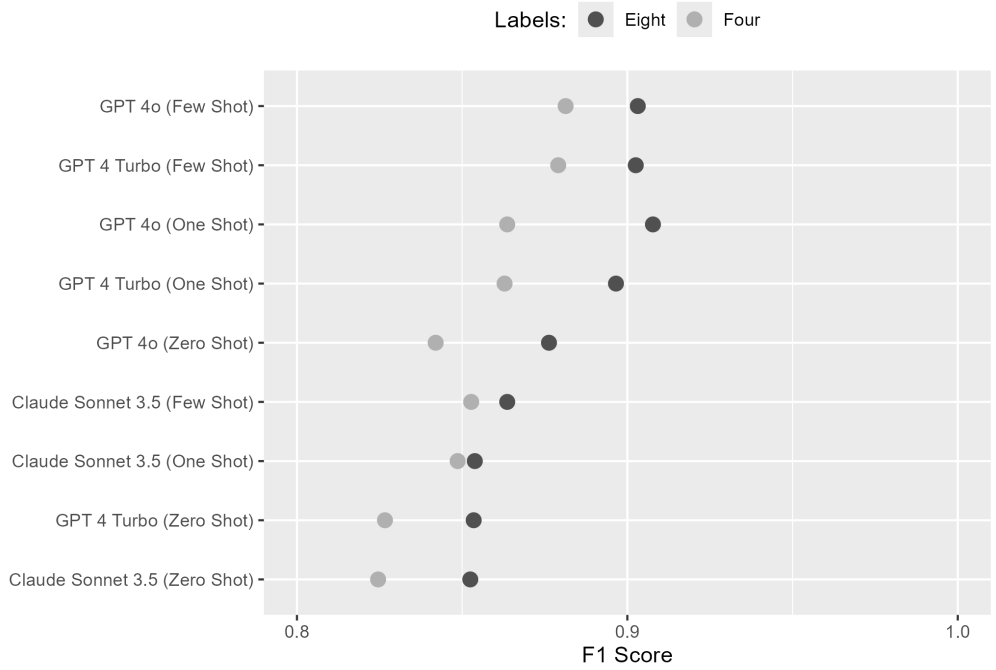

Figure 10: F1 Score Across Models

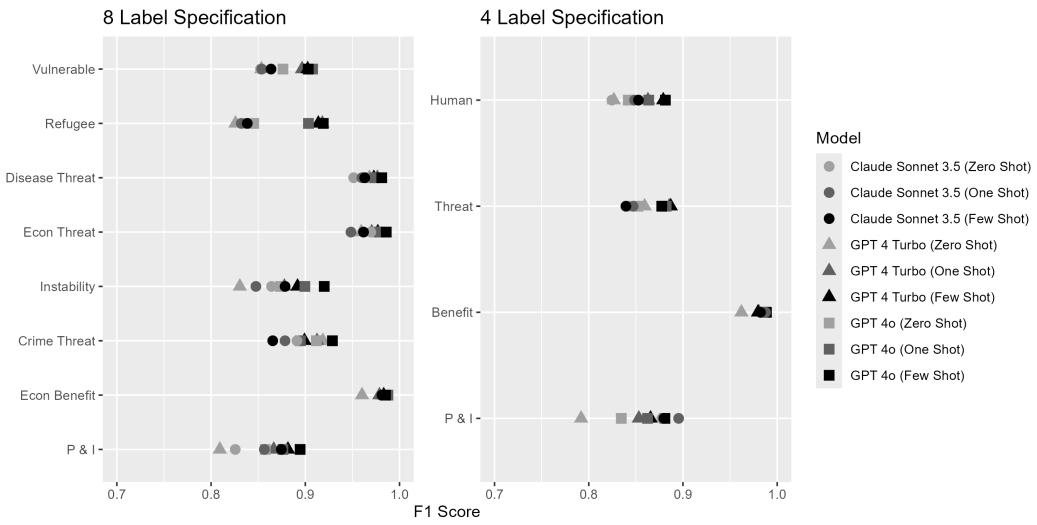

Figure 11: F1 Score Across Labels

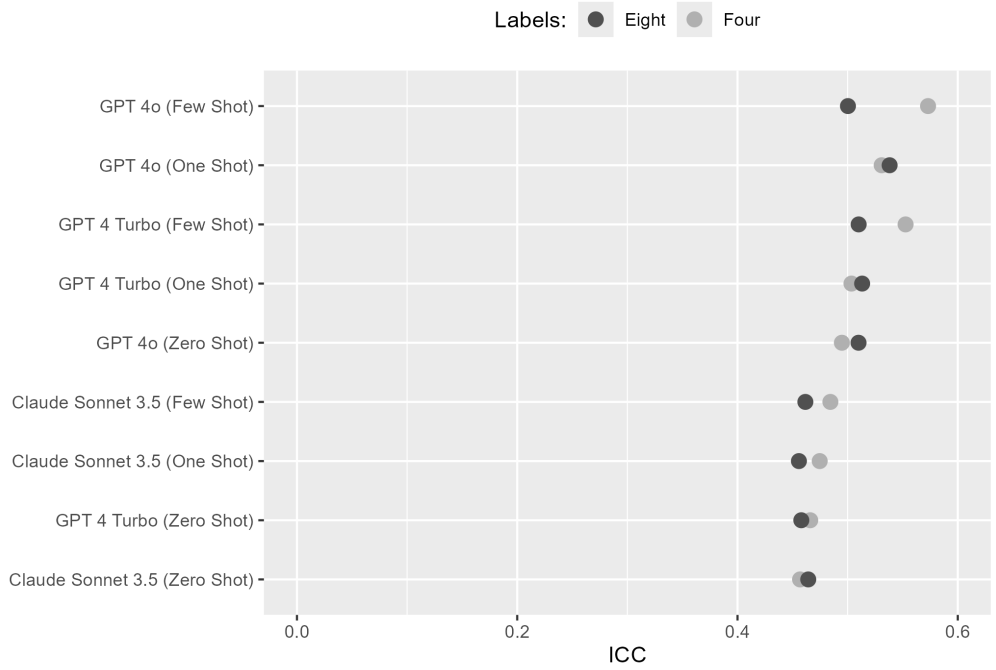

Figure 12: ICC Across Models

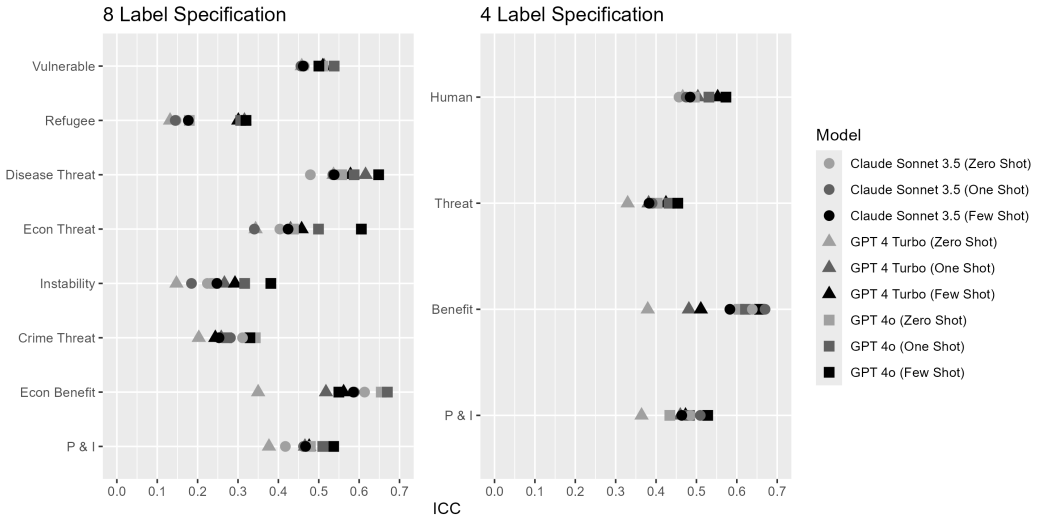

Figure 13: ICC Across Labels

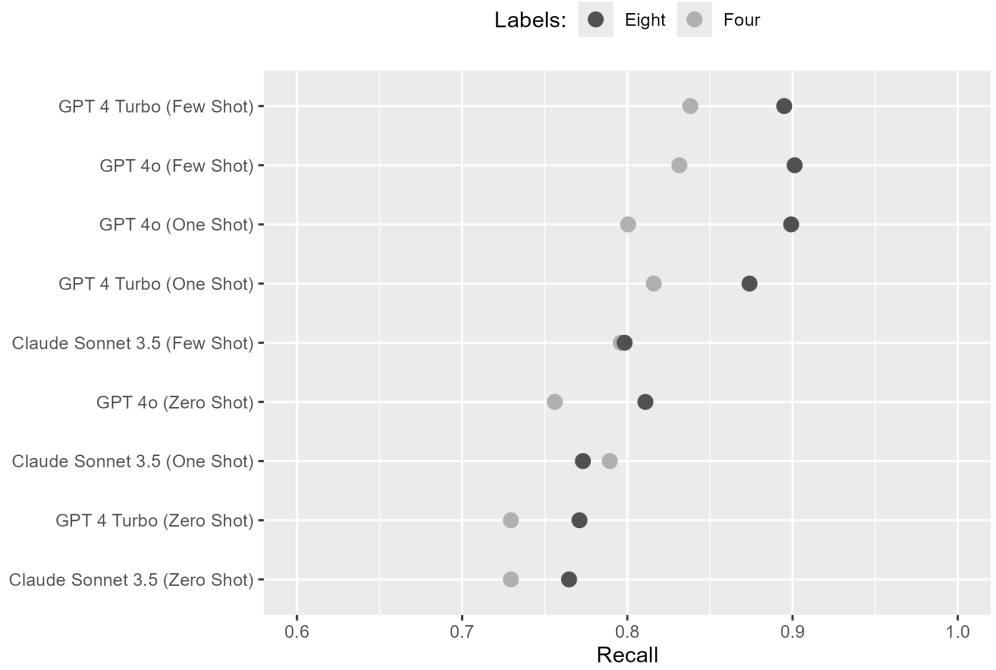

Figure 14: Recall Across Models

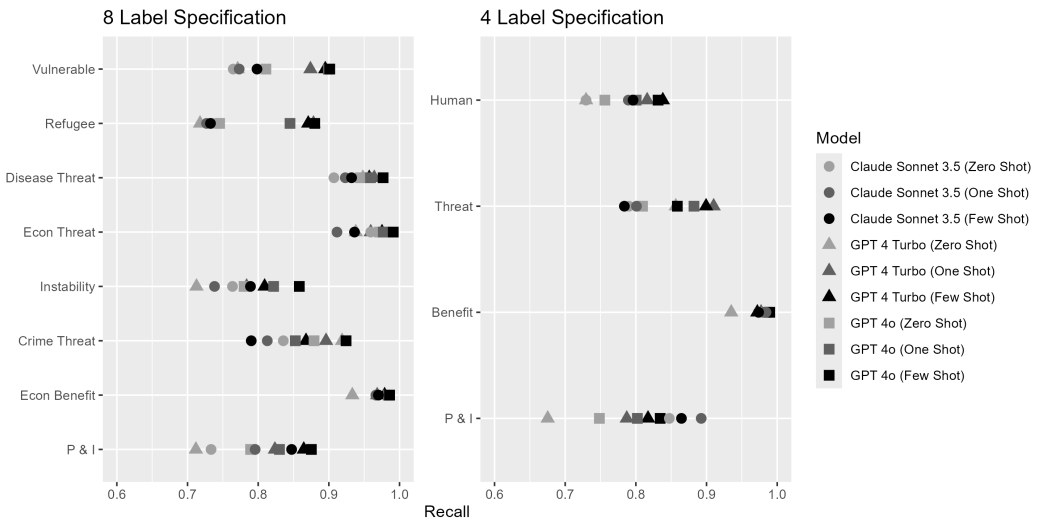

Figure 15: Recall Across Labels

I Predicting Alignment Between Human and AI Output

In this section, I attempt to understand the variation across labels for both the 8 and 4 label specifications of the best performing model: ChatGPT 4o (Few Shot). Tables 5 and 6 regress the probability of correct AI predictions on article word counts, complexity (using the Type-Token Ratio—which is the number of unique tokens divided by the total number of tokens—via R's quanteda package), and human coder agreement (Rul 2019).

In the 8 label model of Table 5, I find that articles in which human coders agreed on the final label were more likely to have aligned codings between humans and AI. While longer articles do not have any consistent effects across models, greater complexity largely predicts more alignment between human and AI labels (except for Label5, in which complexity has a large, negative effect on the probability of alignment).

Table 5: Predicting AI-Human Aligned Responses in 8 Label Models

	Label1	Label2	Label3	Label4	Label5	Label6	Label7	Label8
Word Count (By 100)	0.003	−0.0002	0.002	0.001	−0.001	0.0005	0.001	0.004
	(0.003)	(0.002)	(0.001)	(0.001)	(0.002)	(0.002)	(0.001)	(0.003)
Text Complexity	0.28	0.58	0.40	0.13	−0.80	0.44	0.22	0.16
	(0.26)	(0.25)	(0.14)	(0.12)	(0.25)	(0.25)	(0.12)	(0.27)
Human Agreement	0.21	0.17	0.02	0.05	0.04	0.09	0.02	0.16
	(0.04)	(0.04)	(0.02)	(0.02)	(0.04)	(0.04)	(0.02)	(0.04)
Intercept	0.50	0.43	0.73	0.86	1.23	0.58	0.84	0.59
	(0.15)	(0.14)	(0.08)	(0.07)	(0.14)	(0.14)	(0.07)	(0.15)
Observations	589	589	589	589	589	589	589	589
Adjusted R^2	0.04	0.04	0.01	0.01	0.02	0.01	0.003	0.02

In the 4 label model of Table 6, I find largely similar results for the human agreement variable. However, in these models, I do not see the same positive results—in which greater text complexity predicts greater alignment—but rather a null effect.

Table 6: Predicting AI-Human Aligned Responses in 4 Label Models

	Label1	Label2	Label3	Label4
Word Count (By 100)	0.002	0.003	0.0003	0.003
	(0.003)	(0.003)	(0.001)	(0.003)
Text Complexity	0.22	0.05	0.08	0.15
	(0.28)	(0.29)	(0.11)	(0.28)
Human Agreement	0.22	0.11	0.02	0.15
	(0.05)	(0.05)	(0.02)	(0.05)
Intercept	0.51	0.67	0.92	0.59
	(0.15)	(0.16)	(0.06)	(0.16)
Observations	589	589	589	589
R^2	0.04	0.01	0.003	0.02

J Prompts

I paste below examples of the 4 and 8 label prompts used for ChatGPT. The Claude prompts only differ in that I pasted the newspaper article before the instructions rather than after, as I did here.

4 label specification (Zero Shot):

> Consider the following themes that may be present in newspaper articles:
>
> 1: Humanitarianism, empathy, and perspective-taking. Does the article portray immigrants as vulnerable victims of circumstance or focus on their poverty, victimhood, and lack of choice? Does the article attempt to make the audience see life from an immigrant's point of view, encouraging further thinking and reflection on the issue? Does it provoke sympathy for immigrants? Does the article mention that some immigrants are refugees fleeing persecution or extreme economic deprivation?
>
> 2: Economic, disease, and violent threats. Does the article relate issues of health and disease spread to immigration? Does the article mention how immigration may lead to worse economic outcomes, such as job loss, lower salaries, or less access to public services for citizens? Does the article mention how immigration may lead to increases in violence, crime, or physical insecurity? Does the article associate immigrants with crime, violence, or illicit activities? Does the connect political or societal instability to immigration? Does it use the size of immigration flows to provoke worry or fear (for example, language like "hordes of immigrants" or "massive immigration flows")?
>
> 3: Economic benefits. Does the article mention how immigration may improve economic conditions?
>
> 4: Immigration policies and integration. Does the article bring up the relationship between immigration policies or government actions and immigrant integration?
>
> I will provide you with a newspaper article, and you will determine whether it matches any of the above themes. Label each theme with a 1 if it occurs in the article and a 0 if not. For example, if only theme 3 is mentioned, your response should be: "Theme1": 0, "Theme2": 0, "Theme3": 1, ... If the article is not primarily concerned with immigration, code all themes as 0s.
>
> Your response should be structured like this: ["Theme1": 0, "Theme2": 0, "Theme3": 1, ...]
>
> Do not include any output other than the JSON structured response, and your responses should be in English. The newspaper article is below:

4 label specification (One Shot):

> Consider the following themes that may be present in newspaper articles. For each theme, I provide an explanation and then an example text embodying that theme.
>
> 1: Humanitarianism, empathy, and perspective-taking. Does the article portray immigrants as vulnerable victims of circumstance or focus on their poverty, victimhood, and lack of choice? Does the article attempt to make the audience see life from an immigrant's point of view, encouraging further thinking and reflection on the issue? Does

it provoke sympathy for immigrants? Does the article mention that some immigrants are refugees fleeing persecution or extreme economic deprivation?

Example: "El drama de más de 100 migrantes, refugiados en parque Las Banderas. Muchos de estos migrantes habían clamado por ayuda, sobre todo, porque había niños y adultos mayores con síntomas de gripa, debido a la intensas lluvias de las últimas semanas."

2: Economic, disease, and violent threats. Does the article relate issues of health and disease spread to immigration? Does the article mention how immigration may lead to worse economic outcomes, such as job loss, lower salaries, or less access to public services for citizens? Does the article mention how immigration may lead to increases in violence, crime, or physical insecurity? Does the article associate immigrants with crime, violence, or illicit activities? Does the connect political or societal instability to immigration? Does it use the size of immigration flows to provoke worry or fear (for example, language like "hordes of immigrants" or "massive immigration flows")?

Example: "Aunque oficialmente no se ha reconocido un monto exacto, se calcula que cada paciente crónico extranjero atendido le cuesta a Colombia entre 200 y 220 millones de pesos al año. El Ministro recalcó que estas cifras son parciales y no dimensionan el gasto total para el país. Por eso pidió a todas las entidades territoriales y a las EPS, a través de una circular de febrero pasado, detallar los costos que ha generado este año la atención a extranjeros."

3: Economic benefits. Does the article mention how immigration may improve economic conditions?

Example: "Si hay alguna característica que uno puede buscar en un inmigrante es, que es gente aspiracional, gente que busca algo mejor. Van a trabajar, a, sacrificarse, a luchar."

4: Immigration policies and integration. Does the article bring up the relationship between immigration policies or government actions and immigrant integration?

Example: "'Sus políticas de integración, salud y educación han sido pioneras en la región', añadió el embajador. Colombia sigue necesitando mucho apoyo de la cooperación internacional para atender a los migrantes."

I will provide you with a newspaper article, and you will determine whether it matches any of the above themes. Label each theme with a 1 if it occurs in the article and a 0 if not. For example, if only theme 3 is mentioned, your response should be: "Theme1": 0, "Theme2": 0, "Theme3": 1, ... If the article is not primarily concerned with immigration, code all themes as 0s.

Your response should be structured like this: ["Theme1": 0, "Theme2": 0, "Theme3": 1, ...]

Do not include any output other than the JSON structured response, and your responses should be in English. The newspaper article is below:

4 label specification (Few Shot):

Consider the following themes that may be present in newspaper articles. For each theme, I provide an explanation and then a few example texts embodying that theme.

1: Humanitarianism, empathy, and perspective-taking. Does the article portray immigrants as vulnerable victims of circumstance or focus on their poverty, victimhood, and lack of choice? Does the article attempt to make the audience see life from an immigrant's point of view, encouraging further thinking and reflection on the issue? Does it provoke sympathy for immigrants? Does the article mention that some immigrants are refugees fleeing persecution or extreme economic deprivation?

Example: "El drama de más de 100 migrantes, refugiados en parque Las Banderas. Muchos de estos migrantes habían clamado por ayuda, sobre todo, porque había niños y adultos mayores con síntomas de gripa, debido a la intensas lluvias de las últimas semanas."

Example: "El papa aprovechó para recordar "las prolongadas penalidades y angustias" de la crisis humanitaria de Venezuela, agravadas por la pandemia, así como a "todos aquellos que han dejado el país en busca de mejores condiciones de vida", al referirse a los millones de venezolanos que han tenido que emigrar hacia otras naciones."

Example: "Más de un millón de personas han huido del país en la última década, el 90 por ciento en los últimos cuatro años. La desesperación, el empobrecimiento y la irritación de los venezolanos está creciendo aceleradamente, ocurren brotes espontáneos de violencia todos los días. Existe una anarquía que deja la sensación de que no hay gobierno."

2: Economic, disease, and violent threats. Does the article relate issues of health and disease spread to immigration? Does the article mention how immigration may lead to worse economic outcomes, such as job loss, lower salaries, or less access to public services for citizens? Does the article mention how immigration may lead to increases in violence, crime, or physical insecurity? Does the article associate immigrants with crime, violence, or illicit activities? Does the connect political or societal instability to immigration? Does it use the size of immigration flows to provoke worry or fear (for example, language like "hordes of immigrants" or "massive immigration flows")?

Example: "Aunque oficialmente no se ha reconocido un monto exacto, se calcula que cada paciente crónico extranjero atendido le cuesta a Colombia entre 200 y 220 millones de pesos al año. El Ministro recalcó que estas cifras son parciales y no dimensionan el gasto total para el país. Por eso pidió a todas las entidades territoriales y a las EPS, a través de una circular de febrero pasado, detallar los costos que ha generado este año la atención a extranjeros."

Example: "El presidente de Colombia Iván Duque anunció que se cerrarán los, siete pasos fronterizos con Venezuela, con el objetivo de tratar de detener la propagación del coronavirus en el país."

Example: "Lo que ha representado 'problemas de desorden social en la región fronteriza, casos de trata de personas, victimización violenta, extorsión y despojos, el desarrollo de toda una economía informal en torno a la masiva migración y un importante ejército de reserva de mano de obra barata para la economía legal, informal e ilegal', según recoge Pares en su informe."

3: Economic benefits. Does the article mention how immigration may improve economic conditions?

Example: "Si hay alguna característica que uno puede buscar en un inmigrante es, que es gente aspiracional, gente que busca algo mejor. Van a trabajar, a, sacrificarse, a luchar."

Example: "Colombia tendrá que destinar entre 0,23 y 0,41% de su PIB en el corto plazo para atender a los migrantes venezolanos que huyen de crisis en su país, aunque bien gestionada la ola migratoria puede darle réditos económicos a mediano y largo plazo."

Example: "Pero si esos refugiados estuvieran legalizados, podrían pagar impuestos, trabajar, y aportar a la economía de ese país."

4: Immigration policies and integration. Does the article bring up the relationship between immigration policies or government actions and immigrant integration?

Example: "'Sus políticas de integración, salud y educación han sido pioneras en la región', añadió el embajador. Colombia sigue necesitando mucho apoyo de la cooperación internacional para atender a los migrantes."

Example: "Una de esas medidas, acordadas el pasado 4 de agosto, fue la creación de una cédula fronteriza. El documento, que deben portar tanto los ciudadanos venezolanos como colombianos que residen en la frontera, tiene contenida información fundamental de las actividades que desarrollan y los motivos de su paso entre ambos países."

Example: "Los venezolanos que residen en Colombia y cuentan con el Permiso por Protección Temporal (PPT) podrán obtener su licencia de conducción sin ningún problema. Esto bajo una resolución del Ministerio de Transporte expedida en los últimos días del mandato de Iván Duque. Todo migrante venezolano que tenga con el PPT lo podrá utilizar para realizar trámites asociados con la oficina de tránsito, es decir, podrán iniciar un proceso para solicitar su licencia de conducción y así transitar en vehículos de manera legal por el país."

I will provide you with a newspaper article, and you will determine whether it matches any of the above themes. Label each theme with a 1 if it occurs in the article and a 0 if not. For example, if only theme 3 is mentioned, your response should be: "Theme1": 0, "Theme2": 0, "Theme3": 1, ... If the article is not primarily concerned with immigration, code all themes as 0s.

Your response should be structured like this: ["Theme1": 0, "Theme2": 0, "Theme3": 1, ...]

Do not include any output other than the JSON structured response, and your responses should be in English. The newspaper article is below:

8 label specification (Zero Shot):

Consider the following themes that may be present in newspaper articles:

1: Humanitarianism, empathy, and perspective-taking. Does the article portray immigrants as vulnerable victims of circumstance or focus on their poverty, victimhood, and lack of choice? Does the article attempt to make the audience see life from an immigrant's point of view? Does it provoke sympathy for immigrants?

2: Persecution, poverty, and push factors. Does the article mention that some immigrants are refugees fleeing persecution or extreme economic deprivation?

3: Health issues and threat of disease. Does the article relate issues of health and disease spread to immigration?

4: Economic issues, resource scarcity, and threat. Does the article mention how immigration may lead to worse economic outcomes, such as job loss, lower salaries, or less access to public services for citizens?

5: Issues of crime, violence, and illegal activities. Does the article mention how immigration may lead to increases in violence, crime, or physical insecurity? Does the article associate immigrants with crime, violence, or illicit activities?

6: General instability and threat. Does the connect political or societal instability to immigration? Does it use the size of immigration flows to provoke worry or fear (for example, "hordes of immigrants" or "massive immigration flows")?

7: Economic benefit. Does the article mention how immigration may improve economic conditions?

8: Immigration policies and integration. Does the article bring up the relationship between immigration policies or government actions and immigrant integration?

I will provide you with a newspaper article, and you will determine whether it matches any of the above themes. Label each theme with a 1 if it occurs in the article and a 0 if not. For example, if only theme 3 is mentioned, your response should be: "Theme1": 0, "Theme2": 0, "Theme3": 1, ... If the article is not primarily concerned with immigration, code all themes as 0s.

Your response should be structured like this: ["Theme1": 0, "Theme2": 0, "Theme3": 1, ...]

Do not include any output other than the JSON structured response, and your responses should be in English. The newspaper article is below:

8 label specification (One Shot):

Consider the following themes that may be present in newspaper articles. For each theme, I provide an explanation and then an example text embodying that theme.

1: Humanitarianism, empathy, and perspective-taking. Does the article portray immigrants as vulnerable victims of circumstance or focus on their poverty, victimhood, and lack of choice? Does the article attempt to make the audience see life from an immigrant's point of view? Does it provoke sympathy for immigrants?

Example: "A pesar de que las víctimas que son residentes han sido maltratadas y tienen procesos muy duros, las que están en el exterior son doblemente victimizadas porque se han tenido que ir a otro país y vivir todo el proceso del migrante, adicional al proceso de ser víctima vulnerable."

2: Persecution, poverty, and push factors. Does the article mention that some immigrants are refugees fleeing persecution or extreme economic deprivation?

Example: "El drama de más de 100 migrantes, refugiados en parque Las Banderas. Muchos de estos migrantes habían clamado por ayuda, sobre todo, porque había niños y adultos mayores con síntomas de gripa, debido a la intensas lluvias de las últimas semanas."

3: Health issues and threat of disease. Does the article relate issues of health and disease spread to immigration?

Example: "El presidente de Colombia Iván Duque anunció que se cerrarán los, siete pasos fronterizos con Venezuela, con el objetivo de tratar de detener la propagación del coronavirus en el país."

4: Economic issues, resource scarcity, and threat. Does the article mention how immigration may lead to worse economic outcomes, such as job loss, lower salaries, or less access to public services for citizens?

Example: "Desde que llegó hace un mes, reparte las horas entre buscar empleo y buscar comida. "No hay trabajo para los cucuteños menos para el venezolano", dice este exsargento del Ejército."

5: Issues of crime, violence, and illegal activities. Does the article mention how immigration may lead to increases in violence, crime, or physical insecurity? Does the article associate immigrants with crime, violence, or illicit activities?

Example: "Es cierto que grupos criminales se estén aprovechando de estos venezolanos. Resulta que en el año 2016 se han capturado, por diferentes delitos, 242 ciudadanos de Venezuela en el área metropolitana de Cúcuta."

6: General instability and threat. Does the connect political or societal instability to immigration? Does it use the size of immigration flows to provoke worry or fear (for example, "hordes of immigrants" or "massive immigration flows")?

Example: "El creciente flujo de refugiados hacia Colombia muestra el caos humanitario que se avecina. Ante la tragedia, los mandatarios de la región han resuelto enterrar la cabeza como avestruces."

7: Economic benefit. Does the article mention how immigration may improve economic conditions?

Example: "Si hay alguna característica que uno puede buscar en un inmigrante es, que es gente aspiracional, gente que busca algo mejor. Van a trabajar, a sacrificarse, a luchar."

8: Immigration policies and integration. Does the article bring up the relationship between immigration policies or government actions and immigrant integration?

Example: "'Sus políticas de integración, salud y educación han sido pioneras en la región', añadió el embajador. Colombia sigue necesitando mucho apoyo de la cooperación internacional para atender a los migrantes."

I will provide you with a newspaper article, and you will determine whether it matches any of the above themes. Label each theme with a 1 if it occurs in the article and a 0 if not. For example, if only theme 3 is mentioned, your response should be: "Theme1": 0, "Theme2": 0, "Theme3": 1, ... If the article is not primarily concerned with immigration, code all themes as 0s.

Your response should be structured like this: ["Theme1": 0, "Theme2": 0, "Theme3": 1, ...]

Do not include any output other than the JSON structured response, and your responses should be in English. The newspaper article is below:

8 label specification (Few Shot):

Consider the following themes that may be present in newspaper articles. For each theme, I provide an explanation and then a few example texts embodying that theme.

1: Humanitarianism, empathy, and perspective-taking. Does the article portray immigrants as vulnerable victims of circumstance or focus on their poverty, victimhood, and lack of choice? Does the article attempt to make the audience see life from an immigrant's point of view? Does it provoke sympathy for immigrants?

Example: "A pesar de que las víctimas que son residentes han sido maltratadas y tienen procesos muy duros, las que están en el exterior son doblemente victimizadas porque se han tenido que ir a otro país y vivir todo el proceso del migrante, adicional al proceso de ser víctima vulnerable."

Example: "También tuvo un recuerdo en su mensaje de Navidad para los 'desplazados, los emigrantes y refugiados, y los que hoy son objeto de la trata de personas' y lamentó que muchos pueblos 'sufren por las ambiciones económicas de unos pocos y la avaricia voraz del dios dinero que lleva a la esclavitud.'"

Example: "El papa aprovechó para recordar "las prolongadas penalidades y angustias" de la crisis humanitaria de Venezuela, agravadas por la pandemia, así como a "todos aquellos que han dejado el país en busca de mejores condiciones de vida", al referirse a los millones de venezolanos que han tenido que emigrar hacia otras naciones."

2: Persecution, poverty, and push factors. Does the article mention that some immigrants are refugees fleeing persecution or extreme economic deprivation?

Example: "El drama de más de 100 migrantes, refugiados en parque Las Banderas. Muchos de estos migrantes habían clamado por ayuda, sobre todo, porque había niños y adultos mayores con síntomas de gripa, debido a la intensas lluvias de las últimas semanas."

Example: "El 80% de esas personas refugiadas está en una situación de refugio de largo plazo, es decir que en más de cinco años no han podido volver a su país de origen, y esto se da o porque el conflicto sigue o porque hay explosiones de otros conflictos."

Example: "Más de un millón de personas han huido del país en la última década, el 90 por ciento en los últimos cuatro años. La desesperación, el empobrecimiento y la irritación de los venezolanos está creciendo aceleradamente, ocurren brotes espontáneos de violencia todos los días. Existe una anarquía que deja la sensación de que no hay gobierno."

3: Health issues and threat of disease. Does the article relate issues of health and disease spread to immigration?

Example: "El presidente de Colombia Iván Duque anunció que se cerrarán los, siete pasos fronterizos con Venezuela, con el objetivo de tratar de detener la propagación del coronavirus en el país."

Example: "Los médicos descubrieron entonces que los migrantes se contagiaban en los ríos, donde caen las heces contaminadas con cercarias de las aves migratorias que transitan por el Darién."

Example: "Esto los tiene en alerta roja con una ocupación hospitalaria en las unidades de cuidados intensivos que llega al 98 por ciento y que se agravaría si los migrantes venezolanos continúan su paso sin control."

4: Economic issues, resource scarcity, and threat. Does the article mention how immigration may lead to worse economic outcomes, such as job loss, lower salaries, or less access to public services for citizens?

Example: "Desde que llegó hace un mes, reparte las horas entre buscar empleo y buscar comida. "No hay trabajo para los cucuteños menos para el venezolano", dice este exsargento del Ejército."

Example: "Aunque oficialmente no se ha reconocido un monto exacto, se calcula que cada paciente crónico extranjero atendido le cuesta a Colombia entre 200 y 220 millones de pesos al año. El Ministro recalcó que estas cifras son parciales y no dimensionan el gasto total para el país. Por eso pidió a todas las entidades territoriales y a las EPS, a través de una circular de febrero pasado, detallar los costos que ha generado este año la atención a extranjeros."

Example: "¿El problema es ver a los refugiados como una carga para los Estados? Así es. En el caso de Ecuador, este lunes el Gobierno dio una rueda de prensa en la que explicaban que los 56.000 refugiados colombianos le cuestan al Estado 30 millones de dólares."

5: Issues of crime, violence, and illegal activities. Does the article mention how immigration may lead to increases in violence, crime, or physical insecurity? Does the article associate immigrants with crime, violence, or illicit activities?

Example: "Es cierto que grupos criminales se estén aprovechando de estos venezolanos. Resulta que en el año 2016 se han capturado, por diferentes delitos, 242 ciudadanos de Venezuela en el área metropolitana de Cúcuta."

Example: "Teniendo en cuenta que el ingreso de venezolanos ilegales a la ciudad ha sido relacionado con el incremento de la inseguridad, el Mandatario dijo que solicitaron una unidad especial de Migración Colombia para realizar algunas deportaciones."

Example: "Lo que ha representado 'problemas de desorden social en la región fronteriza, casos de trata de personas, victimización violenta, extorsión y despojos, el desarrollo de toda una economía informal en torno a la masiva migración y un importante ejército de reserva de mano de obra barata para la economía legal, informal e ilegal', según recoge Pares en su informe."

6: General instability and threat. Does the connect political or societal instability to immigration? Does it use the size of immigration flows to provoke worry or fear (for example, "hordes of immigrants" or "massive immigration flows")?

Example: "El creciente flujo de refugiados hacia Colombia muestra el caos humanitario que se avecina. Ante la tragedia, los mandatarios de la región han resuelto enterrar la cabeza como avestruces."

Example: "La ciudad, de 350.000 habitantes, está colapsada. Y si no fuera por las remesas de su hermano, Tilus y su familia estarían en la calle, como otros migrantes."

Example: "La crisis no termina y estamos tratando con consecuencias, de otra serie de problemas políticos que si no se solucionan este flujo, no logrará controlarse."

7: Economic benefit. Does the article mention how immigration may improve economic conditions?

Example: "Si hay alguna característica que uno puede buscar en un inmigrante es, que es gente aspiracional, gente que busca algo mejor. Van a trabajar, a sacrificarse, a luchar."

Example: "Colombia tendrá que destinar entre 0,23 y 0,41% de su PIB en el corto plazo para atender a los migrantes venezolanos que huyen de crisis en su país, aunque bien gestionada la ola migratoria puede darle réditos económicos a mediano y largo plazo."

Example: "Pero si esos refugiados estuvieran legalizados, podrían pagar impuestos, trabajar, y aportar a la economía de ese país. "

8: Immigration policies and integration. Does the article bring up the relationship between immigration policies or government actions and immigrant integration?

Example: "'Sus políticas de integración, salud y educación han sido pioneras en la región', añadió el embajador. Colombia sigue necesitando mucho apoyo de la cooperación internacional para atender a los migrantes."

Example: "Una de esas medidas, acordadas el pasado 4 de agosto, fue la creación de una cédula fronteriza. El documento, que deben portar tanto los ciudadanos venezolanos como colombianos que residen en la frontera, tiene contenida información fundamental de las actividades que desarrollan y los motivos de su paso entre ambos países."

Example: "Los venezolanos que residen en Colombia y cuentan con el Permiso por Protección Temporal (PPT) podrán obtener su licencia de conducción sin ningún problema. Esto bajo una resolución del Ministerio de Transporte expedida en los últimos días del mandato de Iván Duque. Todo migrante venezolano que tenga con el PPT lo podrá utilizar para realizar trámites asociados con la oficina de tránsito, es decir, podrán iniciar un proceso para solicitar su licencia de conducción y así transitar en vehículos de manera legal por el país."

I will provide you with a newspaper article, and you will determine whether it matches any of the above themes. Label each theme with a 1 if it occurs in the article and a 0 if not. For example, if only theme 3 is mentioned, your response should be: "Theme1": 0, "Theme2": 0, "Theme3": 1, ... If the article is not primarily concerned with immigration, code all themes as 0s.

Your response should be structured like this: ["Theme1": 0, "Theme2": 0, "Theme3": 1, ...]

Do not include any output other than the JSON structured response, and your responses should be in English. The newspaper article is below:

K Fixed vs. Ordered Prompts

In this section, I test for the phenomenon in Bisbee et al. 2024 that AI models are more likely to select the first label in a text classification task. I rerun my AI models, randomizing the order in which the AI sees each label in the prompt. In Figure 16 below, I compare the mean proportion of articles coded as 1s by the AI models with fixed and randomized label orderings. This figure shows small differences in the proportion of articles, with the models with randomized label orderings seemingly classifying articles as 1s at higher rates than the fixed label ordering models. However, none of these differences are significant at a 95% confidence level. Furthermore, I compare differences between each label in the fixed and randomized label order models. While there occur some significant differences in the proportions of articles coded as 1s, I do not find that the first labels in the fixed order prompts are any more likely to be chosen than those in the random order prompts.

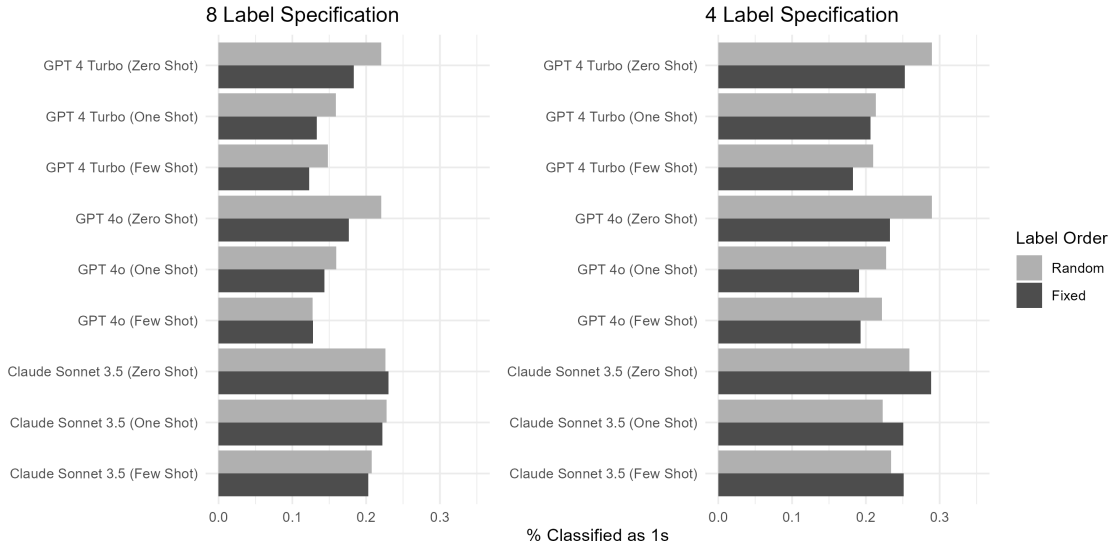

Figure 16: Fixed versus Ordered Prompt Differences

L Correlations and Close Readings

I display here a depiction of the differences in the correlation matrices of the undergraduate and ChatGPT 4o (Few Shot) output for the 8 label specification. Darker purple values indicate that the LLM label correlations are smaller than their undergraduate equivalents, while darker red values indicate that they are larger. Cells with pluses in them indicate that both correlations—for the undergraduate and ChatGPT 4o output—were positive, while minuses indicate that both were negative. The tilde symbol indicates that the correlations were different signs. Most differences in correlations are small, though note the larger decrease (a difference of nearly 0.5) in the correlations for Label 2 of each model. Furthermore, most correlations for both AI models and undergraduates are positive, though a fair number are negative or mixed, especially for labels 7 and 8.

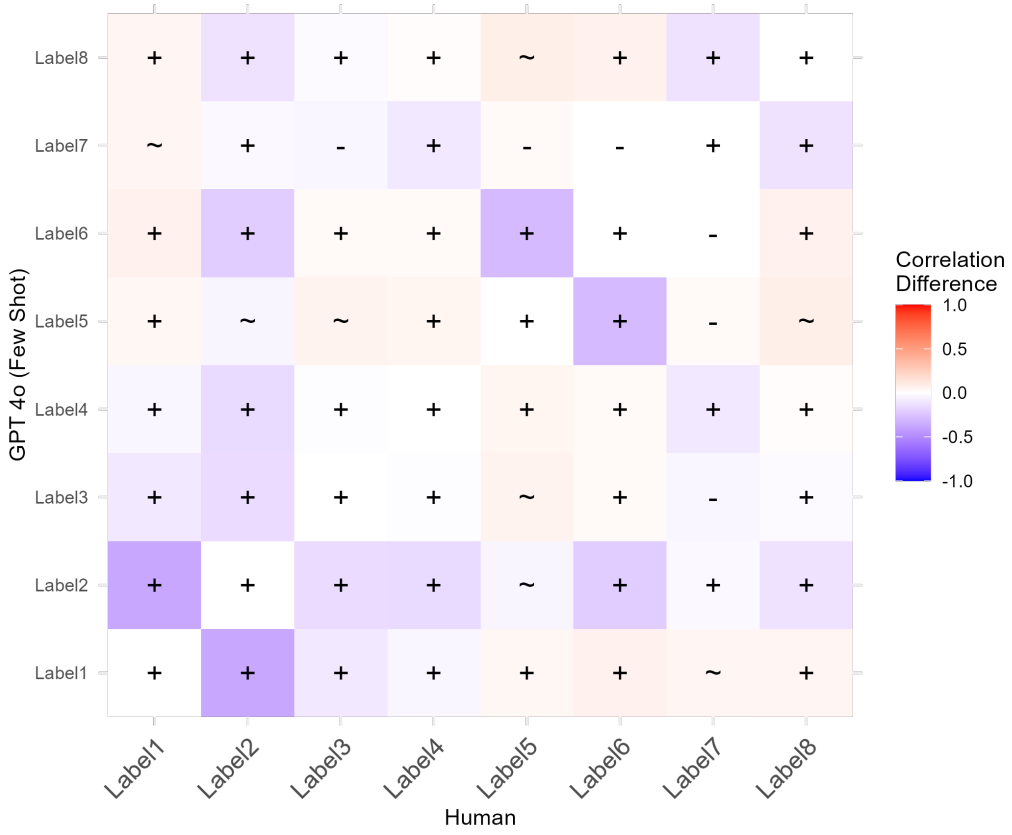

Figure 17: Differences in Correlation Matrices

Based on these correlations and the false positive biases of the AI models, I perform close readings of randomly chosen articles for each label that were coded as 1s by the Few Shot ChatGPT 4o model with 8 labels but 0s by the human coders (performed after my own codings of the articles). Anecdotally, I find that the model picks up on more subtle label cues than the human coders did,

but it had trouble differentiating certain labels, such as labels 1 and 2. Furthermore, the AI model had more miscodings among the more prevalent labels, suggesting that these miscodings might scale with the prevalence of the labels in the corpus. The miscodings seemed more beneficial in this analysis for labels 6 and 8, picking up on more subtle frames that my human coders could detect and correctly coding these documents according to my readings. However, for labels 2, 3, 4, and 5, the miscodings may have been more detrimental, with the AI model seemingly coding this mislabeled articles at random. Notably, many of the articles that I read were miscoded in multiple labels. That is, those articles that were miscoded were more likely to have several miscoded values. This may indicate that the LLMs were not sufficiently coding the articles not focused on immigration as 0s.

I generate three takeaways from my close readings. 1) One should be as specific and structured as possible in the codebook and validate the results of the classification task, given that the AI models may code articles in a way that correlates labels more closely than humans would. 2) Small-scale, close readings can be helpful in ensuring that one's codebook is specific enough for each label, includes the necessary topics and keywords for each label, and includes all relevant labels. For example, in this project, I remove the cultural benefit label from my original codebook due to its low prevalence in the corpus and poor intercoder reliability scores for the human coders. However, I noticed that articles with this theme were categorized into labels 4 and 5 multiple times, suggesting that its inclusion might benefit a more substantive analysis. 3) Since the Few Shot ChatGPT model with 8 labels is the best comparison to human coders in my context, I suggest using it for classification tasks. However, if one wants to descriptively estimate the prevalence of these labels, it may be more accurate to aggregate them up to the broader, 4 label framework (after classification with the more discrete, 8 label specification) before assessing their prevalence over time. This will minimize any error that may arise when AI models—unsure as to the correct labels—classify article into multiple, similarly correlated labels (such as labels 1 and 2).[25]

[25]Here, I found that the AI models would code many articles into labels 1 and 2, despite the fact that most of these should have been coded as just label 1. I suspect that this is due to the fact that labels 1 and 2 are highly correlated and deal with humanitarianism from different angles.