Just Read the Codebook! Make Use of Quality Codebooks in Zero-Shot Classification of Multilabel Frame Datasets

Mattes Ruckdeschel
Leibniz Institute for Media Research
Hans-Bredow-Institute
m.ruckdeschel@leibniz-hbi.de

Abstract

The recent development of large language models (LLM) lowered the barrier to entry for using natural language processing (NLP) methods for various tasks in the related scientific field of computational social science and has led to more scrutiny of their performance on complex datasets. While in many cases the costly fine-tuning of smaller Language Models outperforms LLMs, zero and few-shot approaches on consumer hardware have the potential to deepen interdisciplinary research efforts, whilst opening up NLP research to complex, niche datasets that are hard to classify. The enormous effort involved in coding data sets has the advantage that there are precise instructions on how to code the available data. We investigate, whether highly specific, instructive codebooks created by social scientists in order to code text with a multitude of complex labels can improve zero-shot performance on (quantized) LLMs. Our findings show, that using the latest LLMs, zero-shot performance can improve by providing a codebook on two complex datasets with a total of four different topics and can outperform few-shot in-context learning setups. The approach is equally or more token-efficient, and requires less hands-on engineering, making it particularly compelling for practical research.

1 Introduction

In the last years, a new landscape of open-weight Large-Language-Models (LLMs) with various sizes arose (Zhao et al., 2023). This has led to an increased interest in the field of NLP from various scientific disciplines that work with text data, as well as in applying these new models for downstream tasks. One advantage of Text-to-Text generative models Raffel et al. (2020), which are trained to follow natural language instructions, is their ease of use for people with little engineering experience compared to fully fine-tuning smaller transformer-based language models (Devlin et al., 2019). While Zero-Shot Performance on many datasets is impressive, it often lacks behind the performance achieved by the common approach of supervised fine-tuning of smaller models such as RoBERTa (Liu et al., 2019b) for many challenging tasks (Ziems et al., 2024). Due to the large Parameter size, fine-tuning LLMs is not only resource-intensive, it also requires heaps of quality data. These necessities make Zero-Shot methods interesting for researchers with real-world Datasets for specific research goals, like content analysis of political texts or news articles, which often are particularly challenging for common NLP solutions: They often are highly unbalanced, with some fringe classes only occurring a handful of times, while overall lacking size due to the cost of coding large datasets, as well as data availability. By constructing a research-specific codebook, researchers create highly specific *codebook-specific label assumptions*, that for certain terms might differ from the definitions, which an LLM has learned in its extensive pre-training (Halterman and Keith, 2024), or in now-common supervised fine-tuning on already coded data.

An alternative to fine-tuning for generative LLMs is *In-Context-Learning* (ICL) (Brown et al., 2020), where an LLM is given training examples in their input prompt to learn how to perform a task (Dong et al., 2022). Since inductive codebook construction for content analysis is a labor-intensive task (Hsieh and Shannon, 2005), which results in an explicit, exhaustive list of labels, derived from extensive work with the text data, they distill the researchers' effort to fully grasp a scientific topic according to their research goal. We argue that codebooks provide a token-efficient source of quality information that can help LLMs in challenging zero or few-shot classification tasks. In this paper, we investigate this hypothesis and compare the performance with different ICL approaches.

Our experiments are based on two datasets:

Proceedings of the 31st International Conference on Computational Linguistics, pages 6317–6337
January 19–24, 2025. ©2025 Association for Computational Linguistics

the Argument Aspect Corpus (Ruckdeschel and Wiedemann, 2023) provides concise codebooks and coded data of argument aspects on four different topics, and the Gun Violence Frame corpus, which consists of Frame labels for news headlines regarding gun violence in the U.S..
Issue-specific Frames and aspects are important in content analysis for political science and communication science research, whilst their classification is still a challenging NLP-Task (Ali and Hassan, 2022), especially from an interdisciplinary perspective (Vallejo et al., 2024). Since frame analysis of issue-specific frames is conducted by inductively building a codebook of label definition by analyzing the data at hand (Matthes and Kohring (2008), Borah (2011) i.a.), we assume the utilization of those particular codebooks to be highly beneficial. We focus on researching language models that are practical to run for non-technical teams in order to explore, whether today's language models are able to support researchers from various backgrounds on linguistically challenging research tasks. Further, we make available a test suite, which makes it not only easy to reproduce our results but also to plug different LLMs and codebooks for further testing. [1]

In section 2, we give an overview of ICL, as well as research that tries to utilize codebook data, either to help annotate or classify new data. After describing our experimental setup in 3, we present quantitative results in section 4 and qualitatively look at some errors of the best-performing model in 5. We summarize our results and recommend future work in section 6.

2 Related Work

2.1 In-Context Learning

Brown et al. (2020) showed that it is feasible for LLMs to learn new tasks from demonstrations that are provided in the prompt input of the model, without updating the model's parameters, due to scaling and the large amounts of text data that the models have trained on in pre-training. The method, known as *In-Context Learning* (Dong et al., 2022) has subsequently been researched for improvements (Zhao et al. (2021), Wei et al. (2021), i.a.), different models (Chowdhery et al. (2023) i.a.), and compared to common few-shot learning as well as fine-tuning approaches (Mosbach et al. (2023), Gao et al. (2021), Milios et al. (2023) i.a.).

Wei et al. (2023) introduced *chain-of-thought-prompting*, which improves few-shot capabilities for large-scale LLMs from in-context examples by not only presenting the LLM with solutions to tasks but also including step-by-step reasoning that explains how to solution was derived. Suzgun et al. (2023) showed, that larger models have larger performance gains from CoT-prompting. Chen et al. (2023) investigated how many examples are necessary for ICL to be beneficial. They have found that more demonstrations only barely improve results, for ICL with or without CoT. Further, they found that negative examples tend to 'confuse' LLMs, as they are not reliably able to distinguish between positive and negative samples.

There is also some novel work by Agarwal et al. (2024) and Bertsch et al. (2024) on many-shot in context learning, which is made possible due to newer models with larger context windows (Gemini et al., 2024). While they showed that providing many examples to ICL prompts can improve the quality of predictions on various tasks, the downsides are increased inference time and the need for a large set of labeled examples. Interestingly, with longer context instructions, ICL becomes more robust to changing prediction quality due to example order, which was a reported issue for older LLMs (Lu et al., 2022). Zhao et al. (2024) found, that for instruction-following, more examples do not lead to better performance regarding ICL, and that instruction fine-tuning still significantly outperforms ICL.

2.2 Using Codebook Data

The idea of using codebook data to guide instruction-based LLMs to classify or annotate unseen data has been investigated by some other researchers before. Xiao et al. (2023) explored the feasibility of using GPT-3 (Brown et al., 2020) to code children's questions on their complexity and structure with the help of an expert codebook developed by psychologists. They have found that codebook-based prompting yields higher agreement with expert coding than examples-based prompting. Our study differs from this approach by having a larger label space, as well as using open-weight language models that can be used on consumer hardware, instead of a proprietary API. Halterman and Keith (2024) investigated the classification performance of Mistral7B (Jiang et al., 2023) when using expert codebooks from the field

[1] The test suite can be found at: https://github.com/Leibniz-HBI/Codebook-Paper

of political science on three different datasets. They found that the codebook design matters and that adding a structure that is easy to interpret by LLMs improves zero-shot and few-shot results, yet instruction-tuning using QLORA on codebook data improves zero-shot classification results. Ziems et al. (2024) thoroughly investigated the possibilities of using state-of-the-art LLMs on a variety of CSS tasks and datasets. They find zero-shot performance to be able to match the performance of fine-tuned RoBERTa-Models on some relatively easy tasks, such as stance detection on various datasets. They hypothesize that zero-shot performance is strong on tasks where an LLM's label definitions overlap with the codebook label assumption of the given task. For stance detection or emotion detection tasks, this is likely. They found that for more complex tasks, such as event argument extraction, zero-shot performance was low and LLMs are unlikely to be able to comply with codebook instructions for complex tasks. Atreja et al. (2024) investigated how prompt design affects compliance and accuracy of LLMs for annotation tasks. They found that the compliance is model and prompt-dependent. For the models that they have investigated, adding definitions of classes helped only for ChatGPT, and was detrimental to Falcon7b (Almazrouei et al., 2023) and PaLM2, on a news frame dataset on the topic of gun violence. It has to be noted, that the codebook information that was used for their experiments was significantly shortened from the original codebooks (Liu et al., 2019a). This is likely due to the context window length of the investigated models but differs from the idea of using the entire codebook as is to instruct LLMs.
Due to the fast pace of LLM research, it is unclear whether these findings still hold for newer models. We build upon this research by investigating whether findings from older models like GPT-3 and GPT3.5 also hold for newer LLMs, particularly those with open weights which are of interest to many researchers, and whether we can improve on the recommendation for using codebook data in zero-shot classification.

3 Experimental Setup

For all experiments, we use the Ollama-API [2] to query LLMs with prompts. Since we are interested in making NLP research available to researchers regardless of their technical background, we do all our tests on a single Geforce RTX 3090 TI with 24 GB of VRAM. We report on the achieved F-score and accuracy of the models and compare our results to results from the literature. All experiments were conducted over three runs and mean values are reported.

3.1 Datasets

We use two Datasets in our research. Both datasets contain issue-specific frames, or aspects, which we want to classify on a sentence level. In both datasets, multilabel annotations are permitted, but most of the data points have only one label assigned. In accordance with other tests on the datasets, we omitted sentences labeled with the negative labels *OTHER* and *NO THEME* respectively.

AAC: The Argument aspect corpus (AAC) (Ruckdeschel and Wiedemann, 2023) consists of sentences coded with their aspects, over four topics. As the fourth topic, *abortion*, was added only in a newer version and no baselines are reported, we omit it from our analysis. The codebooks are written in markdown. Label descriptions are short lists of mentionings and keywords that indicate the occurrence of an aspect. The variation in length is indicative of the coders' assumed complexity of various labels and might be a hint as to how much a codebook label assumption differs from a universal label assumption.
GVFC: The Gun Violence Frame Classification Corpus (GVFC) (Liu et al., 2019a) consists of news headlines regarding gun violence in America. The codebook is provided as a Word document and gives examples and definitions in a table. The codebook defines three coding tasks: relevance of a headline to gun violence, whether the focus of the story is on an event or the larger issue of gun violence, and finally the main theme, or frame or the news story according to the headline, which is the code that we want to predict. Each headline can have at most two frames assigned to it. We made several changes to the codebook that we used for our experiments. We deconstructed the table into a simpler list, as shown in table 1. We have also omitted the examples from our codebook, to test the effect of adding label definitions in isolation. Table 1 shows further dataset statistics as well as sample explanations for codebook labels. While the intercoder-agreement for the AAC topics is sufficiently good, the intercoder-agreement for the GVFC is remarkably high, which implies that it was easier for human coders to make labeling

[2] https://ollama.com/

Topic	α_K	N	Example codebook description
Minimum Wage	0.65	194	[ECONOMIC IMPACT] * Effects of minimum wage on the economy in general * Effects on the spending power of consumers * Higher demand for goods and services * mentions of good/bad for the economy; ripple effect
Nuclear Energy	0.65	224	[HEALTH EFFECTS] * worker safety or worker health issues * mentions of people dying due to radiation; cancer; radiation sickness * NOTE: Casualties of explosions are to be labeled with [ACCIDENTS/SECURITY]
Marijuana Legalization	0.64	215	[DRUG ABUSE] * mentions of drug abuse in general; overdose; drugged drives * people drive under the influence of drugs
Gun Violence	0.90*	2609	[PUBLIC_OPINION] The story is about the public's, including a certain community's reactions to gun-related issues, including: - Public opinion polls related to guns - Protests - Mourning victims of gun violence - The public's emotional responses

Table 1: General stats for the AAC topics and the GVFC (* refers to Frame A)

decisions on the dataset.

3.2 Model selection

We test three different open-weights models in our experiment on their ability to do frame classification on a sentence level and on their ability to utilize codebook information to improve their results.
Mistral:7b (Jiang et al., 2023) is an LLM with 7 Billion parameters, which uses sliding window attention (Beltagy et al., 2020) which makes processing longer sequence length computationally feasible. It was also used by Halterman and Keith (2024).
Llama 3.1:8B is an LLM with 8 Billion parameters, that outperforms models with similar size on many common benchmarks (Dubey et al., 2024). The improvements over the previous iteration of LLama (Touvron et al., 2023) are due to the focus on quality data and longer training. They used Grouped-Query-Attention (GQA) (Ainslie et al., 2023) which helped with inference speed.
Gemma2:27 is a decoder-only LLM with 27 Billion parameters. Compared to the other models, it has a fairly low context length of 8k tokens. In order to use this large model on our Hardware, we used a 4-bit-quantized (Dettmers et al., 2023) version of the model.
Table 2 shows the different properties of the three models we used for preliminary testing. As you can see, the model size is in the same range, although parameter counts and quantization techniques differ. All three models can be run on moderate hardware, thus they all are of interest for fairly low-resource settings.

3.3 ICL Methods

We test several scenarios for utilizing ICL in order to align LLMs with the coding task on The AAC. When providing ICL examples, we always give at least one example per class, instead of random sampling over all labels. This is to be neutral to class distribution. We used the suggestions by (Ziems et al., 2024) on how to design prompts for CSS tasks, except for the fact that we have inserted the instruction and codebook before the example. All prompts share the same introductory sentence, as well as the same list of constraints at the end. [3] We tested the following setups:
Label definitions: The LLM is only given label definitions. The topic is mentioned in the introductory sentence.
Codebook definitions: The LLM is given the full codebook. For all experiments, we tried to alter the codebooks as little as possible. We standardized Label names to be in capital letters, in order to make capturing label responses from LLM output easier.
Simple ICL examples: We used up to 2 examples per class and created ICL examples from them. They are structured as *INPUT:*`\n` *<TEXT>* `\n` *Output:* `\n` *the correct label is <LABEL>*
Automated CoT style examples: We prompted the LLMs we investigated to come up with explanations for gold labels from the text and gold label of examples, given codebook data. We manually checked the generated responses for compliance

[3]Detailed prompts can be found in the Appendix, section A.1.

Model Name	Parameters	Context Window	Quantization	Size
mistral:7b-instruct-v0.3-fp16	7B	32K tokens	FP 16	14GB
llama3.1:8b-instruct-fp16	8B	128K tokens	FP 16	16GB
gemma2:27b-instruct-q4_0	27B	8K tokens	Q4_0	15GB

Table 2: Model specs for our experiments. We tested different quantization sizes, different model architectures, and different parameter numbers. Model sizes were comparable between all models.

Component	Minimum Wage	Nuclear Energy	Marijuana Legalization
RAW	179	163	184
Codebook	1533	780	593
ICL 1	595	468	618
CoT 1	2384	1993	2379
ICL 2	1183	1018	1123
CoT 2	4913	4249	4709

Table 3: Token lengths for experiment components from the Gemma Tokenizer. The used CoT explanations for labels add a large amount of additional tokens to the input prompt. The large token count for *Minimum Wage* is due to more Markdown structuring

with the task and found that Mistral and LLama regularly argued about why several labels were not fitting instead of reasoning about the correct label. We opted to only test Gemmas CoT style examples with all models due to their consistent focus on the correct label and mentioning of relevant text sequences.

We tested all combinations of Codebook inclusion and ICL styles. Table 3 shows the token length for all used prompt components. The numbers for the other models' tokenizers are comparable. Note that the token count for the minimum wage codebook is significantly longer than for the other two codebooks. This is because this codebook uses more special characters for Markdown structuring, and has significantly more information per label category. We did not alter this difference in our experiment. The token-wise most expensive addition to the experiments is the use of long CoT ICL examples. The table shows that compared to automated CoL style examples, codebooks add little tokens to the prompt, while ICL examples are also very token-efficient, as less input tokens increase the inference speed for the models.

4 Results

We first present classification results on both datasets for codebook inclusion. Afterward we present the results of ICL-Experiments on Gemma, which was found to be the only model that benefited consistently from more information in the prompt.

4.1 Codebook Experiments

In our first experiments, we investigated, how the three different models were performing on the tasks without ICL additions, and whether adding the codebook improved performance. Table 4 shows the performance differences from all three models when utilizing the codebook, as well as a baseline performance from a fine-tuned RoBERTa model for the AAC dataset. We can see that in most cases, utilizing the codebook has improved the performance regardless of the chosen model and topic. A closer look reveals that for LLama, most of the F-score improvements come from a higher recall, and precision only improved for *Minimum Wage*. This is indicative of two failure modes:

First, LLama assigned more labels when the Codebook was added to the prompt. Second, it became less reliable by following the instruction that labels that are not found should not be mentioned (see section C.2 for an example). Since our automatic evaluation relied on checking for mentioned labels, only a qualitative look into the data revealed that failure mode Llama only improved its precision on *Minimum wage*, while precision lowered for the other two topics when adding the codebook. This is an indication that for LLama, the additional input from the codebook added more confusion instead of aligning it more with the task by giving more detailed instructions.

Mistral performed better for *Nuclear Energy*. and *Marijuana Legalization* when utilizing the codebook, but not for *Minimum Wage*, where only very minor differences in performance can be observed. This may be due to the additional structural noise, or due to the significantly larger amount of tokens in the codebook for *Minimum Wage*, indicating, that a more concise codebook is necessary for Mistral to benefit from codebook information.

Overall, Gemma performs the best of all models, with an improvement in F-score of 0.11 for *Minimum Wage*, 0.02 for *Nuclear Energy*, and 0.13

	No Codebook				Codebook			
Model	**Precision**	**Recall**	**F1**	**Mean Labels Assigned**	**Precision**	**Recall**	**F1**	**Mean Labels Assigned**
Minimum Wage								
Baseline	**0.84**	0.68	**0.75**					
Gemma	0.48	0.73	0.58	1.91	0.61	**0.80**	0.69	1.63
Llama	0.28	0.67	0.40	2.95	0.32	0.77	0.45	3.01
Mistral	0.32	0.75	0.44	2.97	0.30	0.74	0.43	3.09
Nuclear Energy								
Baseline	**0.84**	0.62	0.72					
Gemma	0.61	**0.81**	0.70	1.79	0.66	0.80	**0.72**	1.64
Llama	0.39	0.81	0.53	2.76	0.30	0.85	0.44	3.83
Mistral	0.41	0.74	0.53	2.40	0.57	0.66	0.61	1.57
Marijuana Legalization								
Baseline	**0.82**	0.69	**0.74**					
Gemma	0.43	0.70	0.54	1.94	0.59	**0.77**	0.67	1.56
Llama	0.28	0.70	0.40	3.00	0.22	0.85	0.35	4.63
Mistral	0.31	0.72	0.43	2.79	0.56	0.69	0.62	1.49

Table 4: Micro average performance on the AAC for all three tested models. The baseline is the performance from a fine-tuned RoBERTa-Large model, as described in (Ruckdeschel and Wiedemann, 2022). Bold font denotes the best overall results, while a green background denotes a better result between utilizing and omitting the codebook.

for *Marijuana Legalization*. The performance increase for *Nuclear Energy* was relatively low, but already without a codebook, Gemma was almost on par with the RoBERTa-large baseline. The label definition that Gemma learned in pre-training might have been more in line with the codebook definitions for this topic. Gemma also assigned fewer labels on average when adding codebook information. This shows that it was able to utilize additional information but still followed the constraints to only mention relevant labels. Further, the performance gains on *Minimum Wage* imply Gemma's robustness to additional structural noise.

Table 5 shows the micro F-1 performance on the GVFC for three different prompting strategies regarding codebook additions. The short codebook is taken from (Atreja et al., 2024), the full codebook utilizes the entire label definitions, as shown in Table 1. Here we can see that the smaller models benefit most from more concise codebooks. For both Mistral and LLama, the elaborate codebook does bring only a marginal performance gain, while the gain from the smaller codebook is more significant. The classification performance increases for both additions of a codebook, but only for Gemma does adding more detailed label definitions bring additional performance gains.

In essence, Gemma performed best from all models and benefited the most from using additional codebook information. It showed robustness for structural noise and codebook lengths and was able to follow constraints, even with elaborate label instructions. Both smaller models benefit more from concise notebooks but benefit nonetheless. The experiments show the value of codebook information for zero-shot performance, but that there is still research to be done on how to optimally construct codebooks for LLM utilization.

Codebook Style	Precision	Recall	F1
Gemma			
No	0.41	0.72	0.52
Short	0.49	0.73	0.59
Full	**0.51**	**0.76**	**0.61**
Llama			
No	0.32	**0.74**	0.45
Short	**0.48**	**0.74**	**0.58**
Full	0.35	0.72	0.47
Mistral			
No	0.33	0.52	0.4
Short	**0.58**	**0.57**	**0.47**
Full	0.31	0.8	0.45

Table 5: Performance on the GVFC for three different codebook style. Bold font indicates best performance per model.

4.2 ICL Experiments

Table 6 shows the performance of different ICL additions with and without codebook utilization for Gemma. We can see that solely relying on the codebook is the most beneficial for *Minimum Wage* and *Marijuana Legalization*. Without codebook information, Gemma benefits from at least one ICL example per class. However, on no topic does it significantly outperform solely relying on the codebook information. Only for *Nuclear Energy* does it result in comparable performance. Here, the best performance comes from utilizing ICL examples and codebook information together.

Our automatically created CoT-style examples did significantly decrease performance. Most likely they were significantly too long for Gemma to utilize the additional reasoning they provide. Qualitatively looking at the data revealed that Gemma was focusing only on select labels from the codebook and arguing for their presence on many samples, as long as no label is explicitly mentioned in the sentence. You can find an example in the Appendix C.2. It is unclear, why those particular labels were picked repeatedly.

While the previous experiments have indicated Gemma's ability to make use of more information in its prompt, we can see from this experiment, that concise codebook information is more beneficial to classification performance than simply adding examples. ICL examples, with their rigid structure, seem to restrict the LLMs' ability to 'think' about the results, as it tends to create shorter responses, that do not elaborate on the chosen labels, resulting in a lowered recall. CoT-Style responses on the other hand make the LLM less able to follow instructions and more prone to answer with labels that are not in the codebook. Additionally, it significantly increases inference time, as prompts become very large.

This experiment affirms the benefit of using codebook data. It is similarly token-efficient as ICL examples, without forcing Gemma into short responses, which tend to be detrimental to classification performance. Yet the information is still short enough for Gemma to fully grasp it without inhibiting its ability to follow instructions and constraints.

5 Qualitative Analysis

We have thoroughly investigated failure modes of all models. We look at positive and negative example responses from Gemma+Codebook in table7. [4]. Most errors came from labeling additional labels. There were three prevalent cases in the dataset. **Misinterpretation** of the codebook often occurred when the codebook mentioned keywords that hint at a label (here, *cost of living*), and Gemma failed to semantically parse the sentence correctly. According to the codebook, the label *PRICES* should only be assigned, when price increases due to the introduction of a minimum wage are mentioned. **Overinterpretation** occurred when Gemma speculated about the texts' meaning beyond what is explicitly stated. This error is also a source of low intercoder-agreement between human coders.

We have also seen cases, where Gemma applied the codebook more thoroughly than human coders. In the third example, Gemma's reason to include drug policy is comprehensible and can be interpreted as a miss by coders.

Qualitatively looking at the data also found hard-to-label sentences, which are ambiguous and rely on missing context. In the fourth example in table 7, *dangerous* needs to be interpreted in order to come to a labeling decision. Both labeling decisions are equally correct. The problem of coding due to different perspectives (Romberg, 2022) is not always fully resolvable and shows that some 'misclassifications' by the LLM are valid as well.

6 Conclusion

In this paper, we showed that social science codebooks are a quality source of information for this generation's open-weights LLMs to improve in complex classification tasks that are drawn from CSS research. The researched models are all small enough to run on common consumer hardware, opening up research opportunities for researchers with various backgrounds and resources. The information from codebooks is especially token-efficient, which makes more accurate results achievable for larger datasets than other common few-shot techniques such as common ICL and CoT-ICL. Only on one of three tested datasets did ICL examples achieve equally good results. Though modern LLMs have remarkable context lengths, more information is not necessarily better for complex classi-

[4]For additional errors from other models and settings, see appendix C.2

ICL Type	No Codebook			Codebook		
	Minimum Wage					
	Precision	Recall	F1	Precision	Recall	F1
No ICL	0.48	**0.73**	0.58	0.61	**0.8**	**0.69**
ICL, 1	0.61	0.65	0.63	0.61	0.46	0.53
ICL, 2	0.66	0.63	**0.65**	0.76	0.45	0.57
COT-ICL 1	**0.93**	0.13	0.23	**0.90**	0.13	0.22
COT-ICL 2	0.81	0.25	0.38	0.24	0.12	0.16
	Nuclear Energy					
No ICL	0.61	**0.81**	0.70	0.66	**0.80**	0.72
ICL, 1	0.70	0.76	**0.73**	**0.76**	0.72	**0.74**
ICL, 2	0.64	0.67	0.65	0.75	0.63	0.68
COT-ICL 1	**0.83**	0.33	0.47	0.67	0.44	0.53
COT-ICL 2	0.71	0.19	0.30	0.68	0.19	0.30
	Marijuana Legalization					
No ICL	0.43	**0.70**	0.54	0.59	**0.77**	**0.67**
ICL, 1	0.57	0.58	0.58	0.66	0.63	0.66
ICL, 2	0.63	0.60	**0.61**	**0.70**	0.64	**0.67**
COT-ICL 1	**0.74**	0.31	0.44	0.50	0.27	0.35
COT-ICL 2	0.51	0.31	0.38	0.65	0.20	0.30

Table 6: Performance differences between different types of ICL additions. No ICL denotes the baseline without any additions to the prompt. Green denotes an improvement in micro-f1 score when adding to the prompt, red denotes a decline. Bold text denotes the best performance.

Type	Example
Misinterpretation ✗	*As the cost of living has jumped by leaps and bounds minimum wage has barely made an impact* Prediction: SOCIAL_JUSTICE/INJUSTICE , PRICES PRICES:The sentence mentions the "cost of living has jumped"
Overinterpretation ✗	*Parkland Shooter's Rifle Magazines Were Etched With Swastikas: Reports* Prediction: RACE/ETHNICITY , POLITICS POLITICS: The event is likely being used as a political talking point, fueling debates about gun control, hate crimes, and extremism.
Correct label addition ✓	*Marijuana prohibition makes no more sense than alcohol prohibition did in the early 1900s* PREDICTION: LEGAL_DRUGS , DRUG_POLICY This sentence argues against marijuana prohibition by comparing it to the historical failure of alcohol prohibition. This draws upon elements of **DRUG POLICY** as it discusses a government policy (prohibition)
Ambiguous Sentences	*They know it is both far too dangerous and far too expensive .* COSTS , HEALTH_EFFECTS , ACCIDENTS/SECURITY HEALTH EFFECTS: The phrase "far too dangerous" directly references potential health risks associated with nuclear energy.

Table 7: Examples of three types of misclassifications from Gemma utilizing codebook data. The first row in italics is the input text, the second row shows the predicted labels (red: wrong, green: correct, yellow: missing), and the third row is an edited response from the LLM

fication tasks, which makes it even more important to have concise information to align LLMs with coding instructions. A combination of codebook information and other ICL methods adds much more complexity to the LLMs' input, and inference time to their output and needs further research. In our experiments, CoT-ICL did decrease the performance significantly. While we suspect their lengths to be detrimental to performance, more optimized CoT-ICL examples might still prove beneficial. Another interesting topic for future work is the optimization of codebook structure and length. It might be worthwhile to investigate, whether adding more information to more generic labels (such as *POLICY*) is needed, to guide LLMs towards the research-specific codebook label assumption of particular research interests. Another interesting finding in that regard is that the higher intercoder-agreement from the GVFC did not translate to higher gains when including the corpus. Also, while intercoder-agreement doesn't differ much within the AAC, the performance gains were very different for the three topics. A more thorough investigation between intercoder-agreement on specific frames and their codebook description might reveal strategies to further improve codebooks for zero-shot classification and may help to explain, what makes some tasks easier for zero-shot classification by LLMs compared to others.

Further, we have shown that results for codebook research are very model-dependent, stressing that insights need to be reevaluated on many datasets and models periodically in a model landscape that is ever-changing and quickly improving.

7 Limitations

Due to the explosion of available open-weights models and the many flavours of quantization, our research is not exhaustive, but an indication of the possibilities that codebook learning can bring for advanced LLMs on consumer hardware.

A more thorough and hands-on approach for selecting CoT-Examples might yield better results than the automated generation approach that we have chosen. Another valid option to achieve better results might be the use of larger, and/or proprietary models such as GPT-4. We opted against this in order to be faithful to the task of running on consumer hardware, but a comparison could be insightful nonetheless, especially with larger versions of the same models.

It is also important to state that the tests on both datasets have been conducted on a pre-filtered dataset that ignores data points with negative labels. Another pre-filtering step to find relevant data points is necessary in order to classify real-world data. Predicting negative labels like *irrelevant* or *NO_THEME* in the same step as predicting multiple class labels is a more challenging task.

While incorporating codebooks in zero-shot classification can enhance results, the laborious task of creating codebooks by inductive coding remains necessary. Since creating codebooks still requires coding numerous examples, this approach needs to outperform supervised fine-tuning on the coded data. In real-world settings, however, this alleviates the challenges arising from imbalanced datasets and can lower the technical barriers for interdisciplinary research, so that zero-shot-classification with quality codebooks can still be advantageous for some tasks.

Acknowledgments

This work was funded by the Bundesministerium für Bildung und Forschung (BMBF) as part of the project *FLACA: Few-shot learning for automated content analysis in communication science* (project no. 16DKWN064B).

References

Rishabh Agarwal, Avi Singh, Lei M. Zhang, Bernd Bohnet, Luis Rosias, Stephanie C. Y. Chan, Biao Zhang, Aleksandra Faust, and Hugo Larochelle. 2024. Many-shot In-Context Learning. In *ICML 2024 Workshop on In-Context Learning*.

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebron, and Sumit Sanghai. 2023. GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4895–4901, Singapore. Association for Computational Linguistics.

Mohammad Ali and Naeemul Hassan. 2022. A Survey of Computational Framing Analysis Approaches. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9335–9348, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The Falcon Series of Open Language Models. *Preprint*, arXiv:2311.16867.

Shubham Atreja, Joshua Ashkinaze, Lingyao Li, Julia Mendelsohn, and Libby Hemphill. 2024. Prompt Design Matters for Computational Social Science Tasks but in Unpredictable Ways. *Preprint*, arXiv:2406.11980.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. *Preprint*, arXiv:2004.05150.

Amanda Bertsch, Maor Ivgi, Uri Alon, Jonathan Berant, Matthew R. Gormley, and Graham Neubig. 2024. In-Context Learning with Long-Context Models: An In-Depth Exploration. *Preprint*, arXiv:2405.00200.

Porismita Borah. 2011. Conceptual Issues in Framing Theory: A Systematic Examination of a Decade's Literature. *Journal of Communication*, 61(2):246–263.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, and Prafulla Dhariwal. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Jiuhai Chen, Lichang Chen, Chen Zhu, and Tianyi Zhou. 2023. How Many Demonstrations Do You Need for In-context Learning? In *Findings of the Association* for Computational Linguistics: EMNLP 2023*, pages 11149–11159, Singapore. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, and Adam Roberts. 2023. PaLM: Scaling Language Modeling with Pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient Fine-tuning of Quantized LLMs. *Advances in Neural Information Processing Systems*, 36:10088–10115.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2022. A Survey on In-context Learning. https://arxiv.org/abs/2301.00234v4.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, and Ahmad Al-Dahle. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making Pre-trained Language Models Better Few-shot Learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.

Team Gemini, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, and Anmol Gulati. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Andrew Halterman and Katherine A. Keith. 2024. Codebook LLMs: Adapting Political Science Codebooks for LLM Use and Adapting LLMs to Follow Codebooks. *Preprint*, arXiv:2407.10747.

Hsiu-Fang Hsieh and Sarah E. Shannon. 2005. Three Approaches to Qualitative Content Analysis. *Qual Health Res*, 15(9):1277–1288.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, and Diego de las Casas. 2023. Mistral 7B. *Preprint*, arXiv:2310.06825.

Siyi Liu, Lei Guo, Kate Mays, Margrit Betke, and Derry Tanti Wijaya. 2019a. Detecting Frames in News Headlines and Its Application to Analyzing News Framing Trends Surrounding U.S. Gun Violence. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 504–514, Hong Kong, China. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. RoBERTa: A robustly optimized BERT pretraining approach.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Jörg Matthes and Matthias Kohring. 2008. The Content Analysis of Media Frames: Toward Improving Reliability and Validity. *Journal of Communication*, 58(2):258–279.

Aristides Milios, Siva Reddy, and Dzmitry Bahdanau. 2023. In-Context Learning for Text Classification with Many Labels. In *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP*, pages 173–184, Singapore. Association for Computational Linguistics.

Marius Mosbach, Tiago Pimentel, Shauli Ravfogel, Dietrich Klakow, and Yanai Elazar. 2023. Few-shot Fine-tuning vs. In-context Learning: A Fair Comparison and Evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12284–12314, Toronto, Canada. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Julia Romberg. 2022. Is your perspective also my perspective? enriching prediction with subjectivity. In *Proceedings of the 9th Workshop on Argument Mining*, pages 115–125, Online and in Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Mattes Ruckdeschel and Gregor Wiedemann. 2022. Boundary Detection and Categorization of Argument Aspects via Supervised Learning. In *Proceedings of the 9th Workshop on Argument Mining*, pages 126–136, Online and in Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Mattes Ruckdeschel and Gregor Wiedemann. 2023. Argument Aspect Corpus.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. 2023. Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, Toronto, Canada. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *Preprint*, arXiv:2302.13971.

Gisela Vallejo, Timothy Baldwin, and Lea Frermann. 2024. Connecting the Dots in News Analysis: Bridging the Cross-Disciplinary Disparities in Media Bias and Framing. In *Proceedings of the Sixth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS 2024)*, pages 16–31, Mexico City, Mexico. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2021. Finetuned Language Models are Zero-Shot Learners. In *International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Preprint*, arXiv:2201.11903.

Ziang Xiao, Xingdi Yuan, Q. Vera Liao, Rania Abdelghani, and Pierre-Yves Oudeyer. 2023. Supporting Qualitative Analysis with Large Language Models: Combining Codebook with GPT-3 for Deductive Coding. In *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23 Companion, pages 75–78, New York, NY, USA. Association for Computing Machinery.

Hao Zhao, Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. 2024. Is In-Context Learning Sufficient for Instruction Following in LLMs? *Preprint*, arXiv:2405.19874.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A Survey of Large Language Models. https://arxiv.org/abs/2303.18223v13.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate Before Use: Improving Few-shot Performance of Language Models. In *Proceedings of the 38th International Conference on Machine Learning*, pages 12697–12706. PMLR.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can Large Language Models Transform Computational Social Science? *Computational Linguistics*, 50(1):237–291.

A Appendix

A.1 Prompts

Longer prompts are formatted in markdown to make them easier to read.

A.1.1 Minimum Wage

Simple Prompt:

You are a well-trained social scientist with a speciality for coding data for scientific research.
You have been handed the following markdown codebook as instructions to code example sentences for their aspects.
these are the labels for a coding task regarding arguments in the debate about minimum wage:
ECONOMIC_IMPACT,
UN/EMPLOYMENT_RATE,
TURNOVER,
LOW-SKILLED,
YOUTH_AND_SECONDARY_WAGE_EARNERS,
CAPITAL_VS_LABOR,
SOCIAL_JUSTICE/INJUSTICE,
PRICES,
WELFARE,
GOVERNMENT,
COMPETITION/BUSINESS_CHALLENGES,
MOTIVATION/CHANCES,
In some cases, more than one label applies.
Do not mention labels, that do not apply.
Assign one or more of the labels from the codebook to the following text:

A.1.2 Codebook Prompt:

You are a well-trained social scientist with a speciality for coding data for scientific research. You have been handed the following markdown codebook as instructions to code example sentences for their aspects.

B Aspect Annotation in short argumentative text units.

The task is to find the aspect of an argumentative sentence or short text unit. An aspect in this task is defined as a sub-topic of discourse in the broader topic of the debate about minimum wage. It is possible, that more than one aspect is present in a sentence, please mark all applicable aspects. Below, the aspect categories are further explained. Bulletpoints explain the categories in detail. The list of words below the bulletpoints give some terms.

B.0.1 [MOTIVATION/CHANCES]

- Incentives and opportunities for workers to start a job/career, futher educate themselves and move up in a company
- Influence of minimum wage on education in general or for particular groups (for example for children in certain household or neighbourhoods)
- Eagerness of workers to do their job or do their job well aspiration (of workers) motivation incentives to education incentives to work chances and opportunities for workers investment in education upward mobility (of workers) worker morale

Examples

- "If we raise the minimum wage, low skilled workers might become complacent and don't want to move up"
- "Paying employee better will make them more motivated to work for your company."

B.0.2 [COMPETITION/BUSINESS CHALLENGES]

- How a minimum wage affects the competition between businesses
- Small businessses having trouble due to minimum wage when competing with larger companies
- Also: Businesses have to close because they cannot afford to pay minimum wage
- Also: Businesses see a shrinking profit margin or bevome unprofitable due to minimum wage
- Mentions of higher labour costs which makes it hard for businesses to afford new staff or pay existing stuff. *Might overlap with* `UN/EMPLOYMENT RATE`

B.1 labor costs payability competition between (SME/big) companies effects on SME businesses end of retail

B.1.1 [PRICES]

- Minimum wages effect on consumer prices or cost increases for the public in general
- "Higher prices cancel positive effects of minimum wage."
- Mentions of inflation as an effect of a rising minimum wage consumer prices inflation cost of daycare, housing, living costs [costs are rising! bad effect for workers] negative net effect on impoverished households

---

B.1.2 [SOCIAL JUSTICE/INJUSTICE]

- Mentions of general social justice or general fairness in society
- Mentions of the fairness of worker compensation
- "Workers should be able to afford paying their bills"
- Affects of a minimum wage on poverty at large
- Mentions of inequality or an (increasing) wage gap
- "Wages have to fairly compensate workers"
- Mentions of a living wage, or a wage high enough to afford living in a cetain area.
- Also: Fairness of a federal minimum wage for areas with different living costs (e.g. cities vs. rural areas)
- Mentions of the decrease of the actual minimum wage over the years/decades due to inflation or rise in living costs in general.

living wage living costs [costs must be covered! good effect for workers] social justice dignity inequality/ equality economical struggle (for individuals) standard of living / poverty line poverty (on the macro level) poverty reduction effects on classes / mileus (black neighborhood, workers, small town people)

B.1.3 [WELFARE]

- Mentions of how minimum wage affects welfare spending.
- The relationship between welfare and minimum wage.
- "Full-time workers should not have to rely on welfare"

B.2 government benefits/subsedies/welfare (for minimum wage receivers) social benefits and tax credits

B.2.1 [ECONOMIC IMPACT]

- Effects of minimum wage on the economy in general
- Effects on the spending power of consumers
- Higher demand for goods and services good/bad for the economy ripple effect

---

B.2.2 [TURNOVER]

- Effects of minimum wage on turnover
- Effects of turnover like training costs or vacant positions employee turnover (job stability) cost of hiring and training

---

B.2.3 [CAPITAL VS LABOR]

- Mentions of the power disparity between workers and company owners
- "Class Warfare"
- mentions of predatory capitalism or predatory business practices
- Mentions of the role of unions in setting wages or the lack of strong unions or worker organization
- *Often occurs together with* `GOVERNMENT` *or* `SOCIAL JUSTICE/INJUSTICE` power disparity (capital vs labor) working class explicit reference to injustice of the capitalist system worker exploitation sweatshop collective bargaining

---

B.2.4 [GOVERNMENT]

- Mentions of state intervention into wage setting
- "State vs. free market."
- Also: general agreement/disagreement with state intervention
- *can occur together with* `CAPITAL VS LABOR` government regulation / intervention / legislation government regulation vs free market

---

B.2.5 [UN/EMPLOYMENT RATE]

- Effects on minimum wage on employment or unemployment rates
- Mentions of job creation or jon loss, in genral or in particular companies/sectors, due to minimum wage(increases)
- Minimum wage effects on the job market (labour supply and demand)
- Mentions of job loss due to automation or outsourcing *Can occur together with* `COMPETITION/BUSINESS CHALLENGES`

job loss / creation (on the individual/company level) labor supply outsourcing automation

Examples Businesses are sometimes forced to *hire fewer employees* because they must pay minimum wage the *unemployment rate* will rise because all businesses must pay minimum wage

---

B.2.6 [LOW-SKILLED]

- Effects of minimum wage on low-skilled workers
- Effects of minimum wage on entry level employment, e.g. "If the minimum wage is too high, nobody will hire people who are entering the work force"
- *Note: Distinction to* `YOUTH AND SECONDARY WAGE EARNERS`*: The mention of young people entering the workforce and looking for a first job should be labeled as* `LOW-SKILLED`. `YOUTH AND SECONDARY WAGE EARNERS` *is for the mentions of teenagers who want to earn some extra money*
- Mentions of the working conditions of low-skilled workers

low-skilled workers unskilled workers entry-level workers youth un-/employment

---

B.2.7 [YOUTH AND SECONDARY WAGE EARNERS]

- Mentions of part time workers
- Mentions of workers, who are not earning a primary household income
- Mentions of teenagers, who want to earn some extra money (see distinction at `LOW-SKILLED`)
- Mentions of the working conditions for secondary wage earners and tennagers

students teenagers part time workers

In some cases, more than one label applies. Do not mention labels, that do not apply. Assign one or more of the labels from the codebook to the following text:

B.2.8 Nuclear Energy

Simple Prompt:

You are a well-trained social scientist with a speciality for coding data for scientific research.
You have been handed the following markdown codebook as instructions to code example sentences for their aspects.
these are the labels for a coding task regarding arguments in the debate about nuclear energy:
ACCIDENTS/SECURITY,
TECHNOLOGICAL_INNOVATION,
ENVIRONMENTAL_IMPACT,
WASTE,
ENERGY_POLICY,
RELIABILITY,
WEAPONS,
PUBLIC_DEBATE,
COSTS,
HEALTH_EFFECTS,
FOSSIL_FUELS,
RENEWABLES,
In some cases, more than one label applies.
Do not mention labels, that do not apply.
Assign one or more of the labels from the codebook to the following text:

Codebook Prompt:
You are a well-trained social scientist with a speciality for coding data for scientific research. You have been handed the following codebook as instructions to code example sentences for their aspects. Aspect Annotation in short argumentative text units. The task is to find the aspect of an argumentative sentence or short text unit. An aspect in this task is defined as a sub-topic of discourse in the broader topic of the debate about nuclear energy. It is possible, that more than one aspect is present in a sentence, please mark all applicable aspects. Below, the aspect categories are further explained. Bulletpoints explain the categories in detail. The list of words below the bulletpoints give some terms which are often found as aspect terms or in sentences containing the aspect.

[WASTE]

- nuclear waste
- radioactive waste
- waste storage
- used fuel
- byproducts of nuclear energy generation

[ACCIDENTS/SECURITY]

- Mentions of Reactorsecurity and Reactor accidents
- Accidents
- Terrorist attacks

[HEALTH EFFECTS]

- Mentions of people dying due to radiation
- **NOTE:** Casualties of explosions are to be labeled with [ACCIDENTS/SECURITY]
- Mentions of worker safety or worker health issues
- cancer
- radiation sickness

[Environmental Impact]

- Pollution
- Emission
- Carbon footprint
- relies on supply of fresh water
- sustainability
- Ressource consumption
- space consumption
- Green house gases
- Uninhabitable environment
- Contamination

[COSTS]

- financial costs
- time costs
- cost of developing and deploying nuclear energy Examples: "high construction costs have hindered the development of nuclear power in many countries"

[WEAPONS]

- diversion of nuclear material

[RELIABILITY]

- Also: Efficiency
- need to meet a high and steady energy demand
- nuclear energy depends on steady supply of uranium
- ressource dependency
- complexity of site requirements

[TECHNOLOGICAL INNOVATION]

- innovation of nuclear energy plants, get more efficient, less waste etc.
- SMR
- FSR
- Fusion reactors
- thorium ractors
- sodium cooled reactors
- FS-MSR reactors
- Helium-3 reactors
- New generation reactors

- Third generation reactors
- Modern reactors

[RENEWABLES]

- Mentions of Renewable energy sources such as wind, solar and hydroelectrical power
- Advantages/disadvantages of nuclear power to renewables
- Comparisions between nuclear and renewables as replacements for current power generation

[FOSSIL FUELS]

- Mentions of fossil fuels for power generation such as coal, gas or oil.
- Advantages/disadvantages of nuclear power to fossil fuels

[ENERGY POLICY]

- Proposals, Demands and actions of political actors
- Subsidies
- Loans from the Government
- Governmental plans
- dependency on foreign powers
- wars over ressources
- Effects on specific communities/groups
- Centralized energy vs. locally produced Energy
- taxpayers
- "We need nuclear energy, to not be dependent on foreign powers for our Energy"

[PUBLIC DEBATE]

- mentions of reception of nuclear energy in the public
- mentions of how the discourse about nuclear energy is portrayed in the media
- mentions of how public opinion is influenced
- mentions of the complexity of the issues and peoples' misinformation about the topic
- mentions of constructed hysteria
- Protests
- People Against Nuclear Energy (PANE)
- media influence
- disinformation
- scare phrases
- hysteria
- mobilization

In some cases, more than one label applies. Do not mention labels, that do not apply. Assign one or more of the labels from the codebook to the following text:

B.2.9 Marijuana Legalization

Simple Prompt:

You are a well-trained social scientist with a speciality for coding data for scientific research.
You have been handed the following markdown codebook as instructions to code example sentences for their aspects.
these are the labels for a coding task regarding arguments in the debate about nuclear energy:
LEGAL_DRUGS,
HEALTH/PSYCHOLOGICAL_EFFECTS,
NATIONAL_BUDGET,
HARM,
MEDICAL_MARIJUANA,
GATEWAY_DRUG,
CHILD_AND_TEEN_SAFETY,
ADDICTION,
DRUG_ABUSE,
PERSONAL_FREEDOM,
ILLEGAL_TRADE,
COMMUNITY/SOCIATAL_EFFECTS,
DRUG_POLICY,
In some cases, more than one label applies.
Do not mention labels, that do not apply.
Assign one or more of the labels from the codebook to the following text:

Codebook Prompt:

You are a well-trained social scientist with a speciality for coding data for scientific research. You have been handed the following codebook as instructions to code example sentences for their aspects.

C Aspect Annotation in short argumentative text units.

The task is to find the aspect of an argumentative sentence or short text unit. An aspect in this task is defined as a sub-topic of discourse in the broader topic of the debate about marijuana legalization. It is possible, that more than one aspect is present in a sentence, please mark all applicable aspects. Below, the aspect categories are further explained. Bulletpoints explain the categories in detail. The list of words below the bulletpoints give some terms which are often found as aspect terms or in sentences containing the aspect.

[ILLEGAL TRADE]

- "profits from illegal mj sales benefit criminals"
- "Illegal trade is violent"
- "legalization of mj limit profits for criminals" Cartels Drug Dealers

[CHILD AND TEEN SAFETY]

- "teenagers will have easier access to mj"
- "children will be exposed to mj use"

[COMMUNITY/SOCIETAL EFFECTS]

- Mentions of increase/decrease in cannabis/drug use due to legalization
- "morally wrong to enable people to use drugs"
- "legalization decriminalizes a lot of people"
- "legalization would reduce/increase crime in general"

[HEALTH/PSYCHOLOGICAL EFFECTS]

- Mentions of effects of mj on the body and mind *"Mj consumption makes relaxed" *"Mj consumption makes creative" *"Mj consumption makes violent" *"Mj consumption makes lazy and stupid" effects on the brain

[MEDICAL MARIJUANA]

- "MJ as a treatment for various diseases" [DRUG ABUSE]

- "people drive under the influence"

- mentions of drug abuse in general overdose drugged drives [NATIONAL BUDGET]

- "More money due to tax revenue"

- "More costs, due to more treatment of drug addicts"

[DRUG POLICY]

- mentions of a government call to action
- assessment of specific policies market regulation

[ADDICTION]

- marijuana is addicting

[HARM]

- vs other drugs?

PERSONAL FREEDOM

- "if I smoke weed it only affects myself"
- second hand smoke
- victimless (crime)

[GATEWAY DRUG]

- Marijuana use leads to the use of harder drugs

[LEGAL DRUGS]

- "alcohol is more harmful than mj"
- "alcohol and tobacco have deep cultural roots in society"

In some cases, more than one label applies. Do not mention labels, that do not apply.

C.0.1 Gun Violence Frame Corpus

simple prompt:
You are a well-trained social scientist with a speciality for coding data for scientific research. You have been handed the following codebook as instructions to code headlines for their theme regarding gun violence.
These are the labels for the coding task:
GUN_RIGHTS
GUN_CONTROL
POLITICS
MENTAL_HEALTH
PUBLIC/SCHOOL_SAFETY
RACE/ETHNICITY
PUBLIC_OPINION
SOCIAL/CULTURAL_ISSUES
ECONOMIC_CONSEQUENCES
NO_THEME
In some cases, more than one label applies.
Do not mention labels, that do not apply.
Assign one or more of the labels from the codebook to the following headline:

short codebook:

I want you to perform a data annotation task. Your task is to carefully read the headline of a news article and determine the frame(s) of the news article. Each news headline must be assigned one or more of the following 9 frame classes: GUN_RIGHTS GUN_CONTROL POLITICS MENTAL_HEALTH PUBLIC/SCHOOL_SAFETY RACE/ETHNICITY PUBLIC_OPINION SOCIAL/CULTURAL_ISSUES ECONOMIC_CONSEQUENCES NO_THEME Annotation guidelines: For this task, additional instructions for each of the frame class are provided below:

1. GUN_RIGHTS: The story is related to the Con- stitution, the second amendment, and protection of individual liberty and gun ownership as a right,
2. GUN_CONTROL: The story is about issues related to regulating guns through legislation and other institutional measures.
3. POLITICS: The story is mainly about the political issues around guns and shootings.
4. MENTAL_HEALTH: The story is about issues related to individuals' mental illnesses or emotional well-being, or the mental health

system as a whole.
5. PUBLIC/SCHOOL_SAFETY: Issues related to institutional and school safety
6. RACE/ETHNICITY: The story is about gun issues related to certain ethnic group(s)
7. PUBLIC_OPINION: The story is about the public's, including a certain community's reactions to gun-related issues.
8. SOCIAL/CULTURAL_ISSUES: Societal-wide factors that are related to gun violence.
9. ECONOMIC_CONSEQUENCES: The story is about financial losses or gains, or the costs involved in gun-related issues.

You must follow the instructions mentioned above when providing your response. Do not provide a response that does not align with the instructions. In your output, respond with the frame class the headline belongs to. In your response, you may provide one additional class if you believe the headline belongs to multiple classes. Headline: <>

full codebook:

You are a well-trained social scientist with a speciality for coding data for scientific research. You have been handed the following codebook as instructions to code headlines for their themes. News Framing of U.S. Gun Violence Codebook Instructions: To code each news story, take a look at the headline and then answer the following question: **What is the main theme of this news story?** Below, the themes are further explained.

[GUN_RIGHTS]

The story is related to the Constitution, the second amendment, and protection of individual liberty and gun ownership as a right, including:

- Meaning of the 2nd amendment
- The irrefutability of one's right to own guns
- Gun ownership as critical to democracy and protecting oneself

[GUN_CONTROL]

The story is about issues related to regulating guns through legislation and other institutional measures.

- Enforcing and/or expanding background checks
- Limiting sale of guns and/or related dangerous equipment (e.g., AR15s, semi-automatic rifles, bump stocks, large-capacity ammo)
- Increasing age limits on gun purchases
- Implementing licensing and gun safety training programs

[POLITICS]

The story is mainly about the political issues around guns and shootings, including:

- Political campaigns and upcoming elections (e.g., using guns as a wedge issue or motivating force to get people to the polls)
- Fighting between the Democratic and Republican parties, or politicians
- Political money contributions from gun lobbies (e.g., NRA)
- One political party or one politician's stance on gun violence. Therefore, as long as the news headline mentions a politician's name, it often indicates the theme of politics.
- Often times, the politicians' names or the party names should be mentioned.

[MENTAL_HEALTH]

The story is about issues related to individuals' mental illnesses or emotional well-being, or the mental health system as a whole, including:

- Predicting and preventing mental health breakdowns
- Treating mental illness
- Creating measures to ensure mentally ill people do not have access to guns
- Descriptions of individuals' behavioral / personality traits that indicate instability, impulsivity, anger, etc.

[PUBLIC/SCHOOL_SAFETY]

Issues related to institutional and school safety, including:

- Awareness and monitoring of "troubled" individuals by law enforcement (e.g., local police, FBI)
- Safety measures in schools to prevent or mitigate shootings (e.g., police/safety officers in the school, armed teachers, metal detectors, clear backpacks)
- Note that a headline simply mentioning "school shooting" does not necessarily mean it uses this safety measure frame.

[RACE/ETHNICITY]

The story is about gun issues related to certain ethnic group(s), including:

- Angry, isolated white men as primary perpetrators of domestic gun violence
- Immigrants from Mexico bringing in guns from across the border
- Muslim "terrorists"

- Gun violence in African American communities

[PUBLIC_OPINION]

The story is about the public's, including a certain community's reactions to gun-related issues, including:

- Public opinion polls related to guns
- Protests
- Mourning victims of gun violence
- The public's emotional responses

[SOCIAL/CULTURAL_ISSUES]

Societal-wide factors that are related to gun violence, including:

- Violence in media (e.g., TV/movies and video games)
- Social pressures that may incite someone to violence (e.g., cliques/bullying and isolation)
- Breakdown in family structures, so there is a lack of familial support and stability
- Breakdown in community structures (e.g., religious organizations, other civic-oriented groups), so there is a lack of community support and stability

[ECONOMMIC_CONSEQUENCES]

The story is about financial losses or gains, or the costs involved in gun-related issues, including:

- The actual sales of firearms
- The financial consequences of gun regulation (e.g., lost tax revenue, or gun manufacturing companies moving to a different state)
- The financial state of gun-related lobbying groups (e.g., the NRA)
- Federal budget for gun-related programs

Note:

Code up to two dominant themes

Enter NO_THEME if there is no theme identified.

Make your decision based on the explicit expression.

Do not infer or over interpret.

C.1 Example Generated CoT responses

C.1.1 generation prompt

The generation prompt included the codebook per dataset and the following intro and constraints as follows:
You are a well-trained social scientist with a speciality for coding data for scientific research. Table \ref{tab:app-reasons} shows example CoT style reasonings that were automatically created using Gemma for the topics of the AAC.

You have been handed the following codebook as instructions to code example sentences for their aspects:

CODEBOOK

After thoroughly studying the codebook, You are given a text and its correct gold label .

Please provide a step-by-step reasoning, why according to the codebook, this label is correct.

End the reasoing with "therefore, the correct label is "

C.2 Example Errors

Here we list some more errors that have occurred, either for ICL examples or for methods other than Gemma plus codebook. Other than in the paper body, we do not edit the text.

Experiment	Example
Minimum Wage	*As a matter of record that con is largely unfounded because studies have shown that the higher wages actually helps to increase spending in the economy and increases overall growth of the economy* Let's break down why the label **ECONOMIC_IMPACT** is correct for the given text: 1. **Focus on Economic Effects:** The sentence explicitly discusses the impact of higher wages on the economy. 2. **Keywords:** Phrases like "increase spending in the economy" and "increases overall growth of the economy" directly point to economic consequences. 3. **Codebook Alignment:** The codebook defines **ECONOMIC_IMPACT** as encompassing the effects of minimum wage on the economy in general, including factors like spending power and overall growth. Therefore, the correct label is **ECONOMIC_IMPACT**.
Nuclear Energy	*Wastes are contained and managed , not released .* Let's break down why the label [WASTE] is correct for the sentence "Wastes are contained and managed, not released." Focus on Sub-Topic: The sentence explicitly discusses the handling of nuclear waste. Keywords: The words "wastes," "contained," and "managed" directly align with the description of the [WASTE] aspect in the codebook: [WASTE] - nuclear waste, radioactive waste, waste storage, used fuel, byproducts of nuclear energy generation." Therefore, the correct label is [WASTE]
Marijuana Legalization	*Like sex , alcohol , or cigarettes , marijuana is one of life ' s little pleasures for some people.* Here's a step-by-step reasoning for the given text and label: 1. **Identify Key Phrases:** The sentence compares marijuana to "sex, alcohol, or cigarettes" and refers to it as "one of life's little pleasures." 2. **Match with Codebook Categories:** These phrases directly align with the **[LEGAL_DRUGS]** category in the codebook. This category explicitly mentions comparisons between marijuana and other legal substances like alcohol and tobacco, often highlighting arguments about their relative harm or social acceptance. 3. **Reasoning:** The sentence argues that marijuana, like other socially accepted indulgences (sex, alcohol, cigarettes), can be a pleasurable experience for some individuals. This comparison supports the stance that marijuana should be treated similarly to these legal substances. Therefore, the correct label is **[LEGAL_DRUGS]**.

Table 8: Examples for generated CoT Reasonings

C.2.1 Failure to not list irrelevant Labels

- experiment: LLama+Codebook Minimum Wage
- Text: Discussing the fact that higher labor costs often cause higher prices , Christina writes " Often , the customers paying those prices — including some of the diners at McDonald 's and the shoppers at Walmart — have very low family incomes .
- Response: Based on the provided codebook, I would assign the following labels to the text:

PRICES: This label is applicable as the text mentions that higher labor costs often cause higher prices.
SOCIAL JUSTICE/INJUSTICE: This label is also applicable as the text implies that low-income families are affected by these price increases, which can be seen as a social justice issue.
Additionally, I would consider assigning the label **WELFARE**, as it mentions the potential impact of higher prices on low-income families. However, this is not explicitly stated in the codebook as a possible category, so I will only assign two labels.
Note that other categories like **ECONOMIC IMPACT** or **COMPETITION/BUSINESS CHALLENGES** could be applicable as well, but they are more specific and related to the context of labor costs and prices.

C.2.2 Listing all Labels

- experiment: Mistral+Codebook Minimum Wage
- Text: With an estimated 9 million people making as little as $ 9/hour in these two states, the immediate effects of the wage increases are undeniable .
- Response:

1. MOTIVATION/CHANCES: "If we raise the minimum wage, low skilled workers might become complacent and don't want to move up"
2. COMPETITION/BUSINESS CHALLENGES: "Businesses are sometimes forced to hire fewer employees because they must pay minimum wage", "the unemployment rate will rise because all businesses must pay minimum wage"
3. PRICES: "Higher prices cancel positive effects of minimum wage."
4. SOCIAL JUSTICE/INJUSTICE: "Workers should be able to afford paying their bills", "Affects of a minimum wage on poverty at large", "Mentions of inequality or an (increasing) wage gap"
5. WELFARE: "Full-time workers should not have to rely on welfare"
6. ECONOMIC IMPACT: "Effects of minimum wage on the economy in general", "Effects on the spending power of consumers", "Higher demand for goods and services"
7. TURNOVER: "employee turnover (job stability)", "cost of hiring and training"
8. CAPITAL VS LABOR: "Power disparity (capital vs labor)", "mentions of predatory capitalism or predatory business practices", "Mentions of the role of unions in setting wages or the lack of strong unions or worker organization"
9. GOVERNMENT: "State intervention into wage setting", "government regulation / intervention / legislation", "government regulation vs free market"
10. UN/EMPLOYMENT RATE: "Effects on minimum wage on employment or unemployment rates", "Mentions of job creation or jon loss, in genral or in particular companies/sectors, due to minimum wage(increases)"
11. LOW-SKILLED: "With an estimated 9 million people making as little as $ 9/hour in these two states"
12. YOUTH AND SECONDARY WAGE EARNERS: Not mentioned in the provided text.

C.2.3 Hallucinations of Labels

- experiment:Gemma+Codebook, CoT,1 Marijuana Legalization
- Text: 2.1 million children in the United States live with at least one parent who was dependent on or abused illicit drugs .
- Response: The correct label for this input is **CHILD_WELLBEING**. Here's why:

1. **Focus on Children:** The sentence explicitly states "2.1 million children in the United States live with at least one parent who was dependent on or abused illicit drugs." This directly centers the discussion around the well-being of children.

2. **Parental Drug Use as a Threat:** The phrase "dependent on or abused illicit drugs" highlights a situation that can significantly impact a child's upbringing, safety, and overall development.

3. **Codebook Alignment:** The codebook likely defines **CHILD_WELLBEING** as encompassing issues related to the health, safety, and overall welfare of children. Parental drug use falls squarely within this category due to its potential negative consequences on children.

Let me know if you have any other text snippets you'd like me to analyze!

C.2.4 Arguing for wrong label

- experiment:Gemma+Codebook, CoT,1 Minimum Wage
- Text: In 2014 , supporters of minimum wage cited news organizations who reported the state with the highest minimum-wage garnered more job creation than the rest of the United States .
- Response: Let’s break down why **ECONOMIC_IMPACT** is the most suitable label for this text:

- **Focus on Job Creation:** The sentence highlights a comparison between states with different minimum wages and their respective job creation rates. This points towards an analysis of the economic consequences of minimum wage policies.

- **"More job creation"**: This phrase directly implies an economic outcome – a change in employment levels – that is often used as a metric to assess the effectiveness of economic policies.

Therefore, the correct label is **ECONOMIC_IMPACT**.