Towards Automated Stance Detection in Congressional Hearings with Large Language Models

Anthony Nikolaidis[1]
Supervisor(s): Stephanie Tan[1]**, Edgar Salas Gironés**[1]
[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 22, 2025

Name of the student: Anthony Nikolaidis
Final project course: CSE3000 Research Project
Thesis committee: Stephanie Tan, Edgar Salas Gironés, Odette Scharenborg

***Abstract*—U.S. congressional hearing transcripts offer a valuable window into national policy discourse, but they are prohibitively large for manual analysis. This study explores the use of large language models (LLMs) for multi-speaker, multi-target stance detection, a task that involves identifying each speaker's position on multiple topics within a single hearing. To this end, a novel annotation framework is introduced to produce stance labels for a small corpus of hearings from the House Oversight and Government Reform Committee. The study then evaluates the classification performance of zero- and few-shot prompting and investigates how chain-of-thought reasoning influences the results. The evaluation is conducted using OpenAI's GPT-4o and o3 models. Initial experimental results indicate that combining chain-of-thought with few-shot prompting yields the highest performance, suggesting a promising direction for automating stance analysis using LLMs in complex political discourse.**

***Index Terms*—Stance detection, congressional hearings, large language models**

I. Introduction

What do 6,000 hours[1] of congressional deliberation reveal about America's policy priorities–and who's advocating for what? Across the 117th Congress (2021-2022), just under 2,000 hearing transcripts[2] were released, documenting extensive discussions on issues ranging from healthcare and education to technology and national security [2]. These transcripts represent a rich resource for understanding political discourse in the United States. However, the sheer volume of material makes manual analysis impractical, creating a need for automated methods.

In an effort to address this need, this paper explores how Large Language Models (LLMs) can be used for multi-target stance detection in U.S. congressional hearing transcripts. Compared to other domains, annotated datasets for congressional hearings are limited and often difficult to produce [3], [4]. This challenge positions LLMs as the optimal architectural choice for this task, as they have demonstrated impressive capabilities in zero-shot and few-shot settings, where minimal or even no task-specific training examples are needed [3], [5], [6].

Multi-target stance detection is defined by Küçük and Can [7] as a classification problem in which, for a single author's text and a set of related targets, each target is assigned a label–support, oppose, or neutral–while recognizing that the stance chosen for one target may influence the stances assigned to the others. In the context of hearings, an author is a speaker (e.g., a Committee member or witness), the text is a collection of their utterances, and the targets are topics identified within each transcript.

Hearings are inherently multi-speaker, however, motivating an extension of that definition that considers every speaker as an author and treats all speakers' utterances, i.e. the entire transcript, as the text. Consequently, the task becomes assigning a stance label to each speaker-topic pair while acknowledging that one speaker's expressed position may inform the inferred stances of others.

Given this adapted formulation of the multi-target stance detection problem, the research question that the paper aims to answer is: *how do different LLM prompting strategies compare in their ability to perform multi-target stance detection across multiple speakers towards specific topics in U.S. congressional hearing transcripts?* This overarching question is addressed through the following subquestions:

- How do zero-shot and few-shot prompting strategies compare in terms of performance?
- How does chain-of-thought prompting influence performance when paired with zero-shot and few-shot approaches?

To answer these questions, the paper makes two main contributions. First, it introduces a new framework for annotating topic-based stances across speakers in congressional hearings. This provides the foundation for evaluating classification performance. Second, it presents an empirical comparison of zero- and few-shot prompting, with and without chain-of-thought reasoning, for the task of multi-speaker, multi-target stance detection.

The remainder of the paper is structured as follows. Section II reviews related work on stance detection in the domain of political discourse. Section III describes the framework employed in the study and the experimental setup. Section IV presents and discusses the key findings. Section V outlines the limitations of the study. Section VI suggests directions for future research. Section VII summarizes the contributions of the paper. Lastly, Section VIII addresses the ethical considerations of the study.

II. Related Work

Automatic stance detection has been widely studied in the field of Natural Language Processing (NLP) [8]. Most literature focuses on arguments sourced from online debate forums and social media, most notably X (formerly Twitter) [7], [9]. Furthermore, most models used for this task are designed to determine a single user's stance towards a given topic [10]. In comparison, there is little research on stance detection using larger texts [3].

Initial work by Thomas et al. [11] attempts to detect stance in U.S. congressional floor debates by focusing on single-target, single-speaker scenarios. Each speech segment is treated as expressing either support or opposition to the bill under consideration.

More recently, a policy-focused approach was introduced by Abercrombie and Batista-Navarro [12], who performed stance detection across multiple speakers for UK parliamentary debates. However, each debate still revolves around a single policy. Thus, speakers are labelled with at most one stance per debate. A similar approach was used in [3], where the domain is Australian parliamentary debates.

A comprehensive survey by Abercrombie and Batista-Navarro [13] found that much of the existing work on stance detection in political settings focuses on parliamentary debates. In comparison, the domain of congressional hearings

[1]For this calculation, I assume that the approximately 2,000 hearings that took place in 2021-2022 have an average duration of 3 hours [1].

[2]These approximately 2,000 hearing transcripts span House, Senate, and joint congressional sessions.

remains largely unexplored within the stance detection literature, particularly in multi-topic, multi-speaker settings.

Consequently, annotated datasets for this NLP task in the domain of congressional hearings are scarce. For instance, recent literature introduced the CoCoHD dataset [14], which comprises over 32,000 U.S. congressional hearing transcripts. However, the labelled portion covers just 1,000 sentences and focuses exclusively on a single topic. Additionally, these annotations are made at the sentence level, without broader speaker- or topic-level aggregation. These limitations motivate the development of a new annotation framework suitable for the task of multi-speaker, multi-target stance detection.

III. Methods

This section is split into five subsections. The first describes the structure of hearings and the preprocessing steps. The second outlines the manual annotation process used to identify salient topics and produce speaker-topic stance matrices used as ground truth. The third presents the inter-annotator agreement analysis and adjudication process. The fourth details the prompting strategies and models used to generate predicted labels. Finally, the fifth subsection explains the scoring metrics and aggregation steps employed to evaluate the performance of each model-prompt combination.

A. Data and Preprocessing

Hearings follow a highly structured format that unfolds in a consistent and predictable manner, illustrated in Fig. 1. First, the Chair and Ranking Member give opening statements, followed by prepared statements from the invited witnesses. Then follows a question-and-answer segment, where each committee member, including the Chair and Ranking Member, is allocated five minutes to question the witnesses or make remarks. Once a committee member's time is up, they typically do not speak again during the hearing, except for the Chair and Ranking Member, who also deliver closing statements. Witnesses respond only when addressed, and their input is confined to the context of the questions they receive.

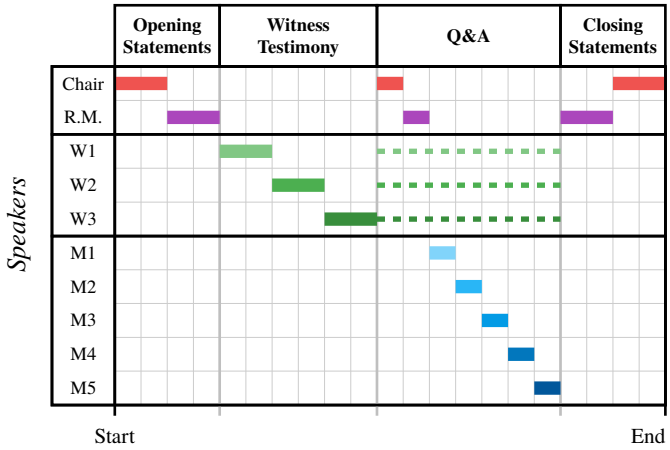

Fig. 1: Typical structure of U.S. congressional hearings. Colored bars represent speaker turns across three phases: opening statements by the Chair and Ranking Member (R.M.), witness testimony (W1–W3), a Q&A where committee members (M1–M5) question the witnesses, and finally closing statements. Witness and committee member counts vary by hearing.

The study uses a curated corpus of 10 hearing transcripts from the House Oversight and Government Reform Committee[3] of the 118th Congress (2023–2025). This particular committee was chosen because its hearings routinely cover high-profile, often contentious issues while generally avoiding deeply specialized jargon, making them both engaging and straightforward to annotate without extensive domain expertise.

First, each transcript is preprocessed to remove metadata such as the title, table of contents, and other non-dialogue elements, retaining only the spoken exchanges between participants. Then, since each speaker's name precedes their utterances, the cleaned transcript can be segmented into discrete speaker turns. Table I provides a minimal example of this segmented representation.

TABLE I
Hearing Transcript Segmented Representation Example

#	Speaker	Segment
1	Ms. Brown	Over the last three and a half years into the pandemic …
2	Mr. Smith	Thank you for your question, Ms. Brown. I'm glad …
3	Ms. Brown	I appreciate that, Mr. Smith. But how do you know …
⋮	⋮	⋮

By splitting transcripts into discrete speaker turns, each hearing can be treated as a dataset of utterances, aiding the manual annotation process described in Section III-B.

B. Manual Annotation Process

The manual annotation process is twofold. First, two non-expert annotators jointly compile a list of salient topics for a given hearing; then, they independently label each speaker's stance towards those topics based on their utterances.

1) *Topic Identification:* A topic refers to any word or short phrase, either stated directly or implied, to which a speaker's claim is directed [15]. Since stance detection applies only in the presence of an assertion, topics are limited to the issues about which such assertions are made. Based on this notion, a simple heuristic is proposed to aid the manual identification of topics: a candidate topic should plausibly elicit the evaluative question, *"Is speaker X expressing support for, or opposition to, topic Y?"* If this question cannot be meaningfully applied in context–such as in cases where a statement merely conveys a fact–then the candidate should likely not be considered a valid topic for the detection task.

Let us consider an example from a hearing on government measures to combat the COVID-19 pandemic. The candidate topic *"pandemic"* is not suitable for stance detection, as it describes a factual situation rather than an issue on which one can take a position. By contrast, *"vaccine mandates"* qualifies as a valid topic because speakers may express support for, or opposition to, vaccination policies.

Congressional hearings typically begin with the Chair and Ranking Member, each delivering opening statements that

[3]Information about the House Oversight and Government Reform Committee is available at congress.gov/committee/house-oversight-and-government-reform/hsgo00

frame the issues under discussion. The study assumes–as is generally observed–that no other participant introduces new themes. Therefore, topics are drawn exclusively from the Chair's and Ranking Member's opening statements. This approach is further motivated by the inherently subjective nature of stance detection, which relies on the presence of clearly defined topics [15]. Especially in the context of congressional hearings, it is imperative that the identified topics are both well-scoped and broadly relevant, such that speakers throughout the hearing are more likely to have expressed a stance towards them.

Furthermore, since this study focuses on multi-target stance detection, the identification process must yield at least one valid topic per hearing, but preferably two or more, to reflect the intended complexity of the task.

In practice, for each hearing in the corpus, the annotators read the Chair's and Ranking Member's opening statements in full and collaboratively compiled a list of candidate topics, i.e. phrases or terms that appeared to structure the argumentation. For each candidate, the aforementioned heuristic was applied to determine whether it could elicit a stance. Topics that met this criterion were shortlisted. In cases where the annotators disagreed on the inclusion of a candidate topic, the decision was resolved through discussion. The final list typically consisted of two to six topics per hearing, depending on the breadth and focus of the statements. These selected topics form the basis for the subsequent stance annotation process, where each speaker's stance is assigned with respect to each identified topic.

2) *Annotating Stance:* The proposed method of assigning stances leverages the structured nature of congressional hearings and builds upon the segmented transcript representation described in Section III-A. For a given hearing, each annotator reads every segment and assigns a stance to that segment's speaker towards every identified topic: $+1$ for support, -1 for opposition, or 0 for neutrality. A neutral stance indicates either the absence of any opinion towards the topic, or that the topic is not discussed, explicitly or implicitly. Table II shows the segmented representation example from Table I, updated to include stance labels towards topics.

TABLE II
HEARING TRANSCRIPT SEGMENTED REPRESENTATION EXAMPLE WITH STANCE LABELS TOWARDS TOPICS

#	Speaker	Segment	IRA	CHIPS Act	···
1	Ms. Brown	Over the last …	-1	0	···
2	Mr. Smith	Thank you for …	$+1$	$+1$	···
3	Ms. Brown	I appreciate …	-1	0	···
⋮	⋮	⋮	⋮	⋮	⋱

Annotating each segment individually offers two main advantages over assigning stances based on a full reading of the transcript. First, it allows annotators to focus on one utterance at a time, thereby reducing cognitive load. Second, it makes the source of each stance decision traceable, as it becomes clear which specific segments contribute to the final label.

At this point in the annotation process, each speaker typically receives multiple stance labels per topic across different segments. These need to be aggregated to obtain an overall stance for each speaker towards each topic.

Unlike debates, committee members generally enter a hearing with firm positions on the relevant issues. Even after listening to witness testimony, their positions tend to remain the same. Therefore, it is possible to aggregate their stances across all segments using summation, without concern for overlooking changes in viewpoint, where opposing stance values (e.g., $+1$ and -1) would cancel each other out, resulting in ambiguous labels. If the total is positive, then the overall stance is $+1$ (support); if negative, the stance is -1 (opposition); and if zero, the stance is 0 (neutral). Table III illustrates this speaker-topic matrix.

TABLE III
SPEAKER-TOPIC MATRIX OF STANCES

Speaker	Topic 1	Topic 2	Topic 3	···
Ms. <*Surname*>	$+1$	-1	-1	···
Mr. <*Surname*>	0	-1	-1	···
Dr. <*Surname*>	0	$+1$	$+1$	···
⋮	⋮	⋮	⋮	⋱

Majority voting was also considered as an alternative aggregation strategy, but most speaker-topic pairs receive a neutral stance (0), which would bias the outcome towards neutrality even when a clear stance is present. For example, the Chair often receives more neutral labels than any other category. This is not due to an absence of opinion, but because much of their speech consists of procedural remarks, such as introducing witnesses, transitioning between speakers, or handling administrative matters like allowing materials to be entered into the record. Summation, by contrast, allows non-neutral segments to accumulate and reflect the speaker's overall position.

C. Inter-annotator Agreement

The dataset was independently annotated by two annotators, each assigning stance labels at the segment level using the same set of identified topics and labelling guidelines. As a result, two speaker-topic matrices were produced for every hearing in the corpus. The reliability of these annotations is assessed using Cohen's κ, which yielded a score of 0.71, representing "substantial" agreement [16].

For each hearing, disagreements were identified by comparing the two stance matrices cell by cell: for each speaker-topic pair where the assigned stances differed, the cell was marked with an "X". These flagged entries were then reviewed collaboratively by the annotators, who revisited the relevant transcript segments and discussed their interpretations until a consensus label was reached. No third-party adjudicator was used.

Following this process, a single unified speaker-topic matrix was created for each transcript in the corpus. These matrices serve as the ground truth used for evaluation.

D. Experimental Setup

The aim of this study is to compare the performance of zero-shot and few-shot prompting strategies for multi-target stance detection across multiple speakers towards specific topics in U.S. congressional hearing transcripts, and to evaluate the effect of chain-of-thought prompting when applied to each of these approaches. Each prompting paradigm and the models used to perform this task are detailed as follows. All prompts were iteratively refined using two hearings not included in the evaluation set.

1) *Zero-shot:* Zero-shot prompting is a technique where an LLM is given a task without any prior examples or specific training for that task. The model relies solely on its pre-existing knowledge and general understanding of language to generate a response. This paradigm represents the most challenging setting, as it may pose significant difficulty even for humans [17].

The zero-shot prompt employed in this study first establishes the model's role as an *impartial congressional hearing annotator* and then defines its task: to determine the stance of various speakers towards a predefined list of topics. It includes an explanation for each stance label (support, oppose, and neutral) and specifies the output format. Finally, the entire cleaned hearing transcript is appended at the end of the prompt. Fig. 2 depicts the components of this prompt.

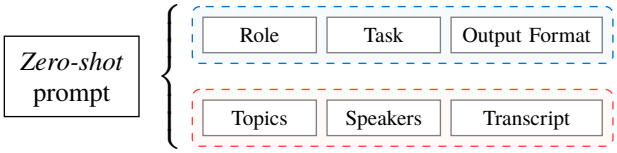

Fig. 2: Components of the zero-shot prompt. The blue section is constant across all hearings, while the red section contains hearing-specific elements.

2) *Few-shot:* Few-shot prompting builds on the foundation of the zero-shot method by including a small number of examples within the input to leverage the in-context learning capabilities of LLMs [18]. Therefore, in addition to the core instructions included in the zero-shot prompt, the few-shot prompt contains five manually annotated examples, preceding the hearing transcript. Three examples demonstrate subtle expressions of stance, while the remaining two are instances where a speaker appeared to be taking a stance but was in fact neutral. This few-shot approach aims to guide the model to avoid common pitfalls in interpretation and enhance its ability to discern complex or ambiguous cases. Fig. 3 illustrates the components that this prompt comprises.

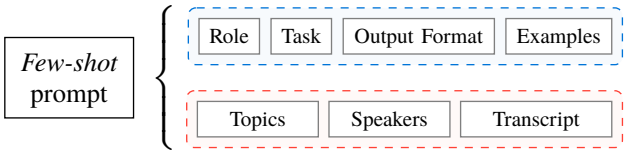

Fig. 3: Components of the few-shot prompt. Similarly to the zero-shot prompt, the blue section remains the same across all hearings and includes examples of the detection task, while the elements of the red section change for each hearing.

3) *Chain-of-thought:* Chain-of-thought (CoT) prompting is a technique that breaks down complex questions or tasks into smaller, logical steps, encouraging a language model to reason through the problem-solving process like a human [19]. This approach can be applied in tandem with both zero-shot and few-shot methodologies. When combined with CoT prompting, zero-shot has been found to significantly outperform zero-shot approaches without CoT in symbolic reasoning tasks [20].

In this study, CoT was applied to both zero- and few-shot prompts, resulting in *CoT-zero-shot* and *CoT-few-shot* variants. These new prompts guide the LLM through the stance detection problem by providing logical step-by-step instructions. Additionally, the model is required to justify the labels it assigns by referencing specific parts of the transcript. This second layer of "thought" aims to allow the LLM to verify its own classification decisions, potentially improving the accuracy and robustness of the output.

4) *Model Choice:* Two models were chosen to carry out the stance detection task: OpenAI's GPT-4o and o3.

GPT-4o is 50% cheaper than previous GPT-4 versions while matching or exceeding them in reasoning and language tasks, making it both efficient and economical for large-scale processing [21]. Furthermore, it has a context window of 128,000 tokens, meaning prompts can accommodate an entire congressional hearing transcript without truncation.

The second model used in the study, OpenAI's o3, is one of the company's most advanced reasoning models [22]. Compared to GPT-4o, it has an even larger context window of 200,000 tokens.

Reasoning models, like o3, are trained to adopt an "internal chain of thought" during inference, devoting additional internal deliberation to tasks requiring step-by-step logical reasoning [23]. Benchmarks reveal that o3 makes 20% fewer major errors than o1, o3′s previous version, on complex real-world tasks [22]. Furthermore, o1 itself significantly outperforms GPT-4o on demanding reasoning tasks [24]. Therefore, o3 is more powerful than GPT-4o. Given this inherent difference in capacity, comparing the results of the two models will demonstrate whether greater model capability yields improved stance detection performance.

E. Evaluation

The study evaluates four prompting strategies, namely zero-shot, few-shot, CoT-zero-shot, and CoT-few-shot, on two LLMs (GPT-4o and o3). For each model-prompt pairing, every hearing transcript in the curated corpus is processed to produce a speaker-topic stance matrix of predicted labels.

For each matrix, per-topic performance is quantified using macro-averaged accuracy, recall, precision, and F1-score. These metrics are then averaged across all topics to yield a single set of scores for each hearing. Finally, scores are averaged over all transcripts to obtain overall accuracy, recall, precision and F1-score for each of the eight model-prompt combinations.

Uniform averaging across topics is a deliberate choice. Naturally, some topics may be discussed extensively while others are mentioned only briefly. Weighting topics by frequency was

considered but ultimately rejected, as it would undervalue the model's ability to correctly identify neutral stances for less-discussed topics. By treating each topic equally, the evaluation better reflects the model's ability to handle infrequently mentioned topics, which often result in a majority of neutral stance labels. Correctly identifying these cases is just as important as detecting strong opinions, as it demonstrates the model's ability to withhold judgement when no stance is expressed.

IV. Results and Discussion

This section presents the results of the evaluation conducted on the curated hearing corpus. It first compares the performance of zero-shot prompting with that of few-shot. Then, it discusses the influence of chain-of-thought on the results. Finally, the section offers an error analysis of model outputs.

A. Presentation of Results

Table IV summarizes the results of the evaluation on the test set across four classification metrics: accuracy, precision, recall and F1-score. The reported scores represent the overall performance for each of the eight model-prompt combinations introduced in Section III-D for the multi-target stance detection task.

TABLE IV
Experimental Results

Model	Prompt	F1-score	Accuracy	Precision	Recall
GPT-4o	zero-shot	63.7	68.7	66.1	67.6
	few-shot	58.4	64.2	59.9	64.0
	CoT-zero-shot	68.9	70.8	71.7	72.6
	CoT-few-shot	66.6	71.4	70.8	68.3
o3	zero-shot	72.3	74.9	73.2	76.9
	few-shot	75.6	77.1	77.7	**78.8**
	CoT-zero-shot	73.1	74.9	76.9	76.1
	CoT-few-shot	**75.8**	**77.9**	**79.9**	78.1

Note—All values are reported as percentages (%), rounded to one decimal place.

The highest F1-score in the study is achieved with o3 using the CoT-few-shot prompt (75.8%), which is nearly identical to its plain few-shot baseline (75.6%). This model-prompt combination outperforms GPT-4o's best result (CoT-zero-shot at 68.9% F1-score) by 10%. Moreover, o3 outperforms GPT-4o under every prompting configuration. The advantage ranges from 6% in CoT-zero-shot to nearly 30% in plain few-shot.

B. Zero-shot vs Few-shot

When paired with GPT-4o, zero-shot outperforms few-shot across all evaluation metrics. F1-score increases by 8.3%, accuracy by 2.9%, precision by 6.1%, and recall by 2.5%. These results suggest that providing exemplars does not aid GPT-4o in this context; on the contrary, performance degrades.

Committee members may express their opinions in various ways. Some express their views outright, while others embed them subtly in the premise of their questions to the witnesses. Any fixed set of exemplars is almost certain to under-represent some of these linguistic cues. Therefore, if the few-shot examples do not span the full spectrum of overt and covert stance indicators, they will inadvertently bias the model's predictions.

By contrast, o3 benefits from few-shot prompting. F1-score improves by 4.6%, accuracy by 2.9%, precision by 6.1%, and recall by 2.5%. These gains suggest that o3 does not overly focus on the provided examples. Rather, it is able to incorporate them into its reasoning process, thereby enhancing its ability to detect both subtle and explicit stances.

Thus, the two models exhibit opposite trends: few-shot prompting hinders GPT-4o but consistently improves o3. This divergence indicates that, in the context of congressional hearings, the effectiveness of few-shot prompting is model-dependent, rather than universally beneficial.

C. Impact of Chain-of-Thought

Adding chain-of-thought reasoning generally improves performance across both zero-shot and few-shot settings. For GPT-4o, CoT leads to gains across all metrics, with both CoT-zero-shot and CoT-few-shot outperforming their non-CoT counterparts. Unsurprisingly, o3′s performance improvements are modest by comparison. This implies that the model is able to realize many of the benefits that CoT offers without the need for explicit reasoning steps.

1) *CoT-zero-shot vs zero-shot:* In the case of GPT-4o, incorporating CoT leads to consistent improvements across all evaluation metrics. Compared to zero-shot, F1-score increases by 8.2%, accuracy by 3.1%, precision by 8.5%, and recall by 7.4%. It appears that the model makes fewer spurious predictions when asked to base its decisions on evidence in the transcript.

In contrast, o3 shows a mixed response. While CoT increases F1-score by 1.1% and precision by 5.0%, accuracy remains unchanged, and recall decreases by 1.0%. These results indicate that CoT boosts the model's ability to avoid false positives, but makes it more prone to missing valid stance expressions.

2) *CoT-few-shot vs few-shot:* The trend is similar in the few-shot setting. Again, GPT-4o benefits from the addition of CoT across the board with F1-score rising by 14.0%, accuracy by 11.2%, precision by 18.2%, and recall by 6.7%. These gains imply that CoT is able to counteract some of the excessive reliance on the examples that few-shot alone can introduce. Thus, the model manages to generalize more effectively.

For o3, CoT-few-shot yields a mere 0.3% gain in F1-score, a 1.0% improvement in accuracy, and a 2.8% increase in precision. However, recall drops by 0.9%. As in the zero-shot setting, applying CoT makes the model more conservative, hindering its ability to identify some obvious stances. Moreover, the marginal overall performance improvement further demonstrates that the benefits of CoT are already internalized by this model.

3) *CoT-zero-shot vs CoT-few-shot:* The performance trends observed when comparing the zero- and few-shot prompts reemerge when CoT is introduced. GPT-4o performs slightly worse with CoT-few-shot than with CoT-zero-shot, while o3

benefits from the addition of examples to the CoT-zero-shot prompt.

Using CoT-zero-shot, GPT-4o achieves a 2.3% higher F1-score and a 4.3% higher recall than with CoT-few-shot, although the latter shows a minor edge in accuracy (+0.6%). Even when the model is asked to reason through the task, examples hinder the its ability to generalize.

Conversely, o3 with CoT-few-shot performs better in all but one metric. Compared to CoT-zero-shot, F1-score increases by 3.7%, accuracy by 4.0%, and precision by 3.9%. Recall, however, drops by 2.6%. While the model's predictions appear to be more accurate with the aid of examples, the decline in recall indicates that it may also overlook subtler expressions of stance. Nonetheless, the overall improvement across the other metrics means that the combination of exemplars and CoT remains advantageous for this model.

D. Error Analysis

A qualitative review of model outputs reveals three key limitations. First, due to the vast length of transcripts[4], both GPT-4o and o3 are observed to overlook brief expressions of stance on less-prominent topics, defaulting instead to neutral.

Second, subtle argumentation tactics are often misinterpreted by both models. For instance, committee members sometimes bring up facts as implicit endorsements or critiques. The models tend to treat these utterances as mere reporting rather than position statements. The human annotators found these types of implicit stances ambiguous as well.

Lastly, both models rely heavily on explicit mentions of topics. When a speaker referred to a topic indirectly, the models often disregarded the reference and marked it as irrelevant. This issue is especially pronounced for vaguer topics. This target type inherently provides little context. Therefore, without an explicit cue, the models have to guess relevance, causing them to ignore subtle allusions.

V. Limitations

This section addresses three main limitations of the study: (i) the significance of the findings, (ii) the quality of the annotations, and (iii) the extent to which the experiment is reproducible.

A. Significance of Results

The primary limitation of this study is its very small dataset, which contains just 10 hearings. Thus, any statistical test is necessarily under-powered and should be interpreted with caution. Nonetheless, significance testing was conducted separately for accuracy, precision, recall, and F1-score to assess the reliability of observed performance differences. A bootstrap approach with 5,000 resamples of paired differences was used. For each contrast, one-tailed p-values were estimated for the hypothesis that the mean difference was greater than zero.

Table V reports these bootstrap-derived p-values for each model independently. In the case of GPT-4o, only adding chain-of-thought reasoning to the few-shot prompt led to substantial and statistically significant gains in F1-score ($p = 0.0238$), accuracy ($p = 0.0214$), and precision ($p = 0.0000$). While the improvement in recall was positive, it was not significant. For o3, none of the tested contrasts reached statistical significance.

TABLE V
Bootstrap-derived One-tailed p-values for the Effect of Different Prompting Strategies on Model Performance

Model	Contrast	F1-score	Accuracy	Precision	Recall
GPT-4o	FS > ZS	0.9022	0.9084	0.9868	0.8206
	CZS > ZS	0.1150	0.1754	0.1316	0.1412
	CFS > FS	0.0238*	0.0214*	0.0000**	0.1700
o3	FS > ZS	0.0730	0.0802	0.0628	0.1912
	CZS > ZS	0.3744	0.4988	0.1580	0.6530
	CFS > FS	0.4470	0.2532	0.0632	0.7212

* $p < 0.05$ ** $p < 0.01$

Note—ZS stands for zero-shot, FS for few-shot, CZS for CoT-zero-shot, and CFS for CoT-few-shot. Effect sizes (mean Δ) for these contrasts ranged from -0.0526 to $+0.0915$ for F1-score, -0.0444 to $+0.1008$ for accuracy, -0.0618 to $+0.1274$ for precision, and -0.0361 to $+0.0498$ for recall.

The bootstrap test was repeated by pooling data from both models to determine whether the observed trends held overall. The results are shown in Table VI. Here, again, CoT-few-shot compared to few-shot produced statistically significant improvements in F1-score (p = 0.0322), accuracy (p = 0.0154), and precision (p = 0.0004), while recall remained non-significant.

TABLE VI
Pooled Bootstrap-derived One-tailed p-values for the Effect of Different Prompting Strategies Across GPT-4o and o3

Contrast	F1-score	Accuracy	Precision	Recall
FS > ZS	0.6002	0.6688	0.5772	0.5998
CZS > ZS	0.1122	0.2464	0.0612	0.2364
CFS > FS	0.0322*	0.0154*	0.0004**	0.2498

* $p < 0.05$ ** $p < 0.01$

Note—ZS stands for zero-shot, FS for few-shot, CZS for CoT-zero-shot, and CFS for CoT-few-shot. Effect sizes (mean Δ) for these contrasts ranged from -0.0075 to $+0.0392$ for F1-score, -0.0092 to $+0.0463$ for accuracy, -0.0053 to $+0.0651$ for precision, and -0.0073 to $+0.0192$ for recall.

Although these results support the qualitative trend that chain-of-thought improves performance in the few-shot setting, they fall short of confirming broader effects with statistical confidence. A substantially larger dataset is needed to draw robust, generalizable conclusions.

B. Dataset Quality

Due to the inherent subjectivity of stance detection, another limitation of the study concerns the quality of the dataset. As shown in Section III-C, the inter-annotator agreement was substantial. However, the annotations were produced by only two annotators. A larger number of annotators would likely increase the reliability of the ground truth.

[4]The average length of the transcripts in the corpus is around 30,000 tokens.

Furthermore, the use of topics as stance targets introduces additional ambiguity. Intrinsically, topics are often vague and, thus, open to interpretation. Greater annotator diversity could help average out individual biases and yield more reliable annotations.

Increasing the number of human annotators, however, presents a significant practical challenge. The annotation process was found to be time-consuming: each hearing took approximately 2 hours to annotate, meaning a total of 4 hours per hearing when both annotators are considered. For the 10 hearings in the test set, this amounted to 40 hours of annotation effort. Scaling up the dataset–by increasing both the number of samples and the number of annotators–would therefore require a substantial investment of time and cognitive effort.

Despite these challenges, larger datasets with broader annotator coverage are crucial for this type of NLP task. Its subjective nature demands high-quality annotations at scale in order to produce reliable evaluation data.

C. Experiment Reproducibility

Regarding the reproducibility of the experiment, a significant limitation is the topic selection step in the manual annotation process. In this study, topics were manually extracted from the transcripts, following predefined guidelines on how many to select, where to locate them, and how to phrase them. While these guidelines were designed to improve consistency, the process still involves subjective judgment.

As a result, it is possible that others attempting to reproduce the experiment may not identify exactly the same topics. Even if the same concepts are captured, differences in wording could influence the stance annotations and, by extension, affect the results of the evaluation.

Another limitation concerns the range of models evaluated. This paper focuses only on two large language models, both of which are proprietary and pre-trained by OpenAI on undisclosed data. As such, it is difficult to assess how much of the observed performance is due to the prompting strategies versus the training data or model architecture. Evaluating a broader set of models–particularly open-source alternatives–would not only offer a more comprehensive view of LLM performance on this task but also improve transparency and reproducibility.

VI. Future Work

Recent literature in the domain of social-media indicates that topic-based stance detection rarely exceeds an F1-score of approximately 76%, even with large annotated datasets and model fine-tuning [25]. The same ceiling appears to have been reached in the present study on congressional hearings (75.8% F1). This suggests that the limitation is methodological rather than domain-specific.

By contrast, treating Frames of Communication (FoCs) as stance targets supplies the model with the explicit reasons underlying an opinion [26]. When FoCs were used on the CoVaxFrames Twitter corpus, direct FoC prediction followed by simple rule-based aggregation lifted performance on topic-level evaluations by almost 12 F1 points [25]. Annotation reliability also improved, with inter-annotator agreement reaching 98% when mapping FoCs to topics.

Crucially, Weinzierl and Harabagiu [27] show that FoCs can be discovered and articulated automatically using LLMs. Compared to manual topic selection, automated FoC identification may offer greater consistency and improve reproducibility across studies.

These findings motivate an FoC-focused research trajectory for the domain of U.S. congressional hearings and potentially other forms of political discourse.

VII. Conclusion

This paper presents a novel framework for annotating topic-based stances in U.S. congressional hearings and demonstrates the potential of large language models to perform multi-target stance detection across multiple speakers in this domain. Experimental results using a small, manually annotated corpus suggest that adding chain-of-thought to few-shot prompting yields the best classification performance. However, a larger dataset is needed to validate these findings. Future work should also explore the use of Frames of Communication as stance targets, given their potential to improve both reproducibility and classification performance.

VIII. Ethical Considerations

The research described in this paper was conducted in accordance with the Netherlands Code of Conduct for Research Integrity [28].

All hearing transcripts analyzed in this study are publicly available records from the U.S. Congress, freely accessible via congress.gov. Furthermore, the data consist entirely of verbatim dialogue from official proceedings, therefore no private or sensitive personal information was collected or processed.

To facilitate full reproducibility, all materials including the preprocessing scripts and evaluation code are available under an open-source license on GitHub[5].

Acknowledgments

I would like to thank my Supervisor, Edgar Salas Gironés, and my Responsible Professor, Stephanie Tan, for their support and guidance throughout this study. I would also like to express my gratitude to the person who assisted me with annotating the dataset–this project would not have been possible without their help.

References

[1] G. Lee, "Ten Questions About Congressional Hearings." Aug. 2019.

[2] A. Irani, J. Y. Park, K. Esterling, and M. Faloutsos, "A Discourse Analysis Framework for Legislative and Social Media Debates," 2024, doi: 10.48550/ARXIV.2407.06149.

[3] S. Ng, J. Zhang, S. Yu, A. Bhatti, K. Backholer, and C. P. Lim, "Stance Classification: A Comparative Study and Use Case on Australian Parliamentary Debates," *Journal of Computational Social Science*, vol. 8, no. 2, p. 43, Mar. 2025, doi: 10.1007/s42001-025-00366-y.

[4] Y. Jiang, J. Gao, H. Shen, and X. Cheng, "Few-Shot Stance Detection via Target-Aware Prompt Distillation," in *Proceedings of the*

[5] github.com/tonynikolaidis/cse3000-research-project

45th International ACM SIGIR Conference on Research and Development in Information Retrieval, in SIGIR '22. New York, NY, USA: Association for Computing Machinery, Jul. 2022, pp. 837–847. doi: 10.1145/3477495.3531979.

[5] E. Allaway and K. McKeown, "Zero-Shot Stance Detection: A Dataset and Model Using Generalized Topic Representations," no. arXiv:2010.03640. arXiv, Oct. 2020. doi: 10.48550/arXiv.2010.03640.

[6] E. Allaway and K. McKeown, "Zero-Shot Stance Detection: Paradigms and Challenges," *Frontiers in Artificial Intelligence*, vol. 5, Jan. 2023, doi: 10.3389/frai.2022.1070429.

[7] D. Küçük and F. Can, "Stance Detection: A Survey," *ACM Comput. Surv.*, vol. 53, no. 1, pp. 1–37, Feb. 2020, doi: 10.1145/3369026.

[8] B. Schiller, J. Daxenberger, and I. Gurevych, "Stance Detection Benchmark: How Robust Is Your Stance Detection?," *KI - Künstliche Intelligenz*, vol. 35, no. 3, pp. 329–341, Nov. 2021, doi: 10.1007/s13218-021-00714-w.

[9] A. ALDayel and W. Magdy, "Stance Detection on Social Media: State of the Art and Trends," *Information Processing & Management*, vol. 58, no. 4, p. 102597, Jul. 2021, doi: 10.1016/j.ipm.2021.102597.

[10] P. Sobhani, D. Inkpen, and X. Zhu, "A Dataset for Multi-Target Stance Detection," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, M. Lapata, P. Blunsom, and A. Koller, Eds., Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 551–557.

[11] M. Thomas, B. Pang, and L. Lee, "Get out the Vote: Determining Support or Opposition from Congressional Floor-Debate Transcripts," in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, D. Jurafsky and E. Gaussier, Eds., Sydney, Australia: Association for Computational Linguistics, Jul. 2006, pp. 327–335.

[12] G. Abercrombie and R. Batista-Navarro, "Policy-Focused Stance Detection in Parliamentary Debate Speeches," *Northern European Journal of Language Technology*, vol. 8, no. 1, Jul. 2022, doi: 10.3384/nejlt.2000-1533.2022.3454.

[13] G. Abercrombie and R. Batista-Navarro, "Sentiment and Position-Taking Analysis of Parliamentary Debates: A Systematic Literature Review," *Journal of Computational Social Science*, vol. 3, no. 1, pp. 245–270, Apr. 2020, doi: 10.1007/s42001-019-00060-w.

[14] A. Hiray, Y. Liu, M. Song, A. Shah, and S. Chava, "CoCoHD: Congress Committee Hearing Dataset," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds., Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 15529–15542. doi: 10.18653/v1/2024.findings-emnlp.911.

[15] A. Irani, J. Y. Park, K. Esterling, and M. Faloutsos, "WIBA: What Is Being Argued? A Comprehensive Approach to Argument Mining." arXiv, 2024. doi: 10.48550/ARXIV.2405.00828.

[16] J. R. Landis and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977, doi: 10.2307/2529310.

[17] T. B. Brown *et al.*, "Language Models Are Few-Shot Learners," no. arXiv:2005.14165. arXiv, Jul. 2020. doi: 10.48550/arXiv.2005.14165.

[18] Q. Dong *et al.*, "A Survey on In-context Learning," no. arXiv:2301.00234. arXiv, Oct. 2024. doi: 10.48550/arXiv.2301.00234.

[19] J. Wei *et al.*, "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," no. arXiv:2201.11903. arXiv, Jan. 2023. doi: 10.48550/arXiv.2201.11903.

[20] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large Language Models Are Zero-Shot Reasoners," no. arXiv:2205.11916. arXiv, Jan. 2023. doi: 10.48550/arXiv.2205.11916.

[21] "Hello GPT-4o." May 2024.

[22] "Introducing OpenAI O3 and O4-Mini."

[23] F. Xu *et al.*, "Towards Large Reasoning Models: A Survey of Reinforced Reasoning with Large Language Models," no. arXiv:2501.09686. arXiv, Jan. 2025. doi: 10.48550/arXiv.2501.09686.

[24] "Learning to Reason with LLMs."

[25] M. A. Weinzierl and S. M. Harabagiu, "The Impact of Stance Object Type on the Quality of Stance Detection," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, Eds., Torino, Italia: ELRA and ICCL, May 2024, pp. 15942–15954.

[26] R. M. Entman, "Framing: Toward Clarification of a Fractured Paradigm," *Journal of Communication*, vol. 43, no. 4, pp. 51–58, Dec. 1993, doi: 10.1111/j.1460-2466.1993.tb01304.x.

[27] M. Weinzierl and S. Harabagiu, "Discovering and Articulating Frames of Communication from Social Media Using Chain-of-Thought Reasoning," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, Y. Graham and M. Purver, Eds., St. Julian's, Malta: Association for Computational Linguistics, Mar. 2024, pp. 1617–1631.

[28] "Netherlands Code of Conduct for Research Integrity | NWO." 2018.