Zero-shot learning for (dis)agreement detection in meeting transcripts

Comparing latent topic models and large language models

D.F.P. de Weerd[1]

Supervisor(s): prof. dr. C.M. Jonker, M. Tarvirdians[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 25, 2023

Name of the student: D.F.P. de Weerd
Final project course: CSE3000 Research Project
Thesis committee: prof. dr. C.M. Jonker, M. Tarvirdians, M.L. Molenaar

Abstract

This paper presents a novel approach to detect agreement and disagreement moments between participants in meeting transcripts without relying on labeled data. We propose a model in which disagreement detection is defined as the process of first identifying argumentative theses relevant to a given corpus of text and then classifying all phrases in the text as being either in favor of, against or expressing no opinion on a given thesis. To identify relevant theses, we compare the performance of a latent Dirichlet allocation-based topic model against that of a diverse set of large language models. To classify the stance of a phrase with respect to a thesis, only large language models are used. We find that, while state-of-the-art large language models do not outperform topic modeling-based approaches in extracting semantically relevant content, they are capable of presenting such content in a more concise and grammatically correct manner. We also find that state-of-the-art large language models are not capable of accurately performing stance classification as described above.

1 Introduction

In business, academia, government and most other institutions, almost all key decisions are made during meetings. Although the overwhelming majority of meetings used to be in-person, the COVID-19 pandemic spurred a large increase in the number of meetings conducted virtually. As virtual meetings easily allow for full recording and transcription of everything said, this opens up a wealth of data to analyze. Identifying key subjects in such meeting transcripts, along with statements on which participants either agree or disagree, can provide valuable insights in the decision-making process.

A natural language processing approach that seems applicable to the problem described is that of *argument mining*. Argument mining is the process of identifying argumentative structures in text documents [1]. This approach however assumes the existence of explicit argumentative structures in the text documents studied, which often holds for written text but does not necessarily remain applicable when processing transcripts of recorded meetings. Additionally, argument mining approaches tend to depend on labeled training data [2; 3], which is often not available. We therefore investigate other approaches.

1.1 Topic modeling

The process of detecting key subjects in an unlabeled corpus of text is generally referred to as *topic modeling*. The foundational approach of topic modeling is latent Dirichlet allocation (LDA) [4]. With LDA, a document is represented as a mixture of multinomial probability distributions over words, where each probability distribution corresponds to a topic discussed in the document. Key advantages of latent Dirichlet allocation are its conceptual simplicity and its ability to work without labeled training data. However, it cannot distinguish between topics and viewpoints on topics and is therefore of limited use for (dis)agreement detection.

To detect (dis)agreement moments in a corpus of text, Vilares & He [5] propose the latent argument model, an extension of latent Dirichlet allocation that separately models topics and viewpoints. It categorizes the words in a document as either background words, topic words or argument words and then models the document as a set $(\phi^b, \theta^d, \omega_z, \psi_{z,a})$ where ϕ^b is the background word distribution, θ^d is the document-specific topic distribution from which a topic z can be sampled, ψ_z is the topic-word multinomial distribution, ω_z is the topic-specific viewpoint distribution from which an viewpoint a can be sampled and $\psi_{z,a}$ is the topic-specific viewpoint word multinomial distribution.

Although the latent argument model remedies LDA's lack of ability to distinguish between topics and viewpoints, several disadvantages inherent to latent Dirichlet allocation remain. Its key drawback is its limited linguistic modeling capability: the model does not account for word order, context or homonyms and cannot generate abstractive summaries. Additionally, it can only detect a pre-specified number of topics and viewpoints and does not work on small documents.

1.2 Large language models

In recent years, most if not all advances in natural language processing have been driven by large language models. A large language model is a Transformer-based [6] deep learning model with tens of millions to hundreds of billions of parameters that is trained to generate natural language. It is first pre-trained on a large corpus of unlabeled text and can additionally be fine-tuned using prompt-response pairs to perform a specific task or to perform instructions in general [7; 8]. Some models are additionally fine-tuned for instruction following using Reinforcement Learning from Human Feedback (RLHF) [9], a form of reinforcement learning in which the human-evaluated quality of model outputs is incorporated in the training process.

The key advantage of large language models compared to LDA-based topic models is their superior ability to model the intricacies of natural language. They can take into account the context of a phrase, deal with ambiguity and nuance and generate coherent and fluent text, achieving near human-level performance on tasks such as text summarization [10] and utterance classification [11]. After pre-training and fine-tuning, such tasks can be performed without relying on additional training data, which means that a large language model can be used to process small and unlabeled documents. Examples of large language models include BERT, T5, GPT-3, PaLM, LLaMA, Claude, GPT-4, Pythia and PaLM 2 [12; 13; 14; 15; 16; 17; 18; 19].

Drawbacks of large language models include their tendency to generate plausible-sounding but false statements, their prohibitive cost of training and their "black box" nature that makes their inner workings extremely difficult if not impossible to understand. Additionally, most state-of-the-art large language models are developed by commercial entities that do not disclose the source code or weights of their models. This inhibits transparency and reproducibility and imposes additional cost on researchers attempting to evaluate

their performance.

1.3 Contribution

This paper aims to answer the following questions:

1. How well do different large language models perform at identifying concise, semantically relevant and grammatically correct argumentative theses from a corpus of text, when compared to latent topic models?
2. How accurate are different large language models in classifying whether a given expression agrees with, disagrees with or is neutral with respect to a given argumentative thesis?

We show that while state-of-the-art large language models do not outperform the latent argument model by Vilares & He [5] in extracting semantically relevant content, they add the ability to present equally relevant content in a more concise and grammatically correct manner. We further show that state-of-the-art large language models are not (yet) capable of accurately classifying phrases from a meeting transcript as agreeing with, disagreeing with or being neutral with respect to an argumentative thesis.

The paper is structured as follows. Chapter 2 describes how latent topic models and large language models can used to identify relevant theses and classify whether an expression (dis)agrees with an argumentative thesis. Chapter 3 describes the data set, specifies the models and model settings used for the experiment and defines the proposed evaluation metrics for relevance, grammatical acceptability, conciseness and classification performance. Chapter 4 provides the results of the experiment, Chapter 5 further discusses the results, notes some limitations and proposes some approaches for future research. Chapter 7 outlines a brief conclusion.

2 Methodology

For a given set of expressions, we aim to obtain a set of argumentative theses, each of which expresses an opinion that is relevant to the topics discussed. A thesis should be relevant to the set of expressions, grammatically correct and concisely formulated. Relevant theses are identified with the latent argument model and a diverse set of large language models.

After identifying relevant theses, we seek to classify expressions from the data set as agreeing with, disagreeing with or neutral with respect to the theses identified. This classification task is performed using large language models.

2.1 Problem statement

For a meeting transcript, we define the set of participants $\mathbf{P}$ and the set of spoken sentences $\mathbf{S}$. We then define a conversation as a set of expressions $\mathbf{E}$, where an expression is a 4-tuple:

$$(\mathbf{p} \in \mathbf{P}, \mathbf{s} \in \mathbf{S}, \mathbf{t_0}, \mathbf{t_1})$$

In this tuple, $\mathbf{p}$ is the identity of the participant speaking, $\mathbf{s}$ is the sentence spoken, $\mathbf{t_0}$ is the time at which the participant started speaking and $\mathbf{t_1}$ is the time at which the participant stopped speaking. For every conversation, we then seek to extract a list of 3-tuples:

$$(\mathbf{x}, \mathbf{E_a} \subseteq \mathbf{E}, \mathbf{E_d} \subseteq \mathbf{E})$$

where $\mathbf{x}$ is an argumentative thesis on a topic discussed in the conversation, $\mathbf{E_a}$ is the subset of expressions agreeing with the thesis, and $\mathbf{E_d}$ is the subset of expressions disagreeing with the thesis. This gives us two tasks. The first task is that of thesis identification: the process of creating the list $(x_1, x_2, ..., x_n)$. The second task is that of stance classification: the process of determining $\mathbf{E_{a,i}}$ and $\mathbf{E_{a,i}}$ for every x_i.

2.2 Latent argument model

To identify relevant arguments, the latent argument model is used to identify *viewpoint sentences*. Viewpoint sentences are sentences from the analyzed transcript that are scored as relevant to a given topic-viewpoint combination.

To score the relevance of a sentence to a topic-viewpoint combination, the latent argument model uses two distinct methods: *generative* scoring and *discriminative* scoring.[1] Recall that the latent argument model models a topic-viewpoint combination as a topic-specific argument word distribution $\psi_{z,a}$, where z represents a single topic and a represents a viewpoint on the given topic. We represent a sentence as a set W of size n consisting of words $\{w_0, w_1, ..., w_{n-1}\}$. When using generative scoring, the relevance of a sentence W to a topic-viewpoint combination (z, a) is expressed as the sum of the log-probabilities of its words given $\psi_{z,a}$:

$$s_{generative}(W) = \sum_{i=0}^{n} log(P(w_i|\psi_{z,a}))$$

When using discriminative scoring, the same computation is used but the probability $P(w_i|\psi_{z,a})$ is divided by the maximum probability of w_i appearing in other topic-viewpoint combinations.

To separate topics and viewpoints, the latent argument model distinguishes between background words, topic words and argument words. Several switch strategies are used to this end:

- `LAM`: Use only the statistical approach described in Chapter 1.1;
- `LAM_POS`: Use Part-of-Speech (PoS) tagging and assume that nouns are topic words, adjectives, adverbs and verbs are argument words, and words with other PoS-tags are background words;
- `LAM_LEX`: Incorporate PoS-tags as a prior rather than applying them directly and additionally use a pre-defined subjectivity lexicon.

[1] Vilares and He do not distinguish between generative and discriminate scoring in their paper, but the model implementation they offer allows for both. The description above is based on the Python implementation they provide on GitHub: https://github.com/aghie/lam/lam.py

2.3 Large language models

To identify relevant argumentative theses from a meeting transcript, a prompt is used that combines a description of the task with part of the data set. Theses are abstractively generated: the model analyses the data set and generates a new, natural-language sentence that re-states a position held by one of the meeting participants. The prompt used is shown in Figure 1.

```
From the following dialogue, extract an
argumentative thesis with which some
dialogue participants agree and others
disagree.  You should formulate the thesis
in a single sentence.  It should be very
concise, affirmatively formulated, and
contain a clear claim.  It should not
contain nuance or caveats.  Again, it
should be terse.  Output only the thesis
and nothing else.

Dialogue:  [EXPRESSIONS INSERTED HERE]
Thesis:
```

Figure 1: Thesis extraction prompt

To classify whether a given list of phrases agrees, disagrees or is neutral with regards to a given thesis, a prompt similar to that for thesis identification is used. The model takes a thesis and a list of phrases as input and returns a classification as output. The prompt used is shown in figure 2.

```
Task:  for the given argumentative thesis
and dialogue, indicate for every sentence
whether it agrees with, disagrees with, or
is unrelated to the thesis.  Give your
response as a Python list containing
one element per sentence, where 0 means
unrelated, 1 means agree, and -1 means
disagree.
Thesis:  [THESIS INSERTED HERE]
Dialogue:  [DIALOGUE INSERTED HERE]
```

Figure 2: Stance classification prompt

3 Experimental setup

3.1 Data set

The AMI corpus is a multimodal data set consisting of over 100 hours of meetings [20]. The corpus consists of a combination of naturally-occurring meetings and scenario-driven meetings specifically set up for the purpose of creating the data set.

Given that limited computational and financial resources were available for this research project, the experiment was limited to the transcripts `ES2002a`, `ES2002b`, `ES2002c` and `ES2002d`. These transcripts represent four one-hour sessions spread over one day in which the participants are tasked with designing a television remote control. This subset of the AMI corpus contains 3.720 lines of dialogue, which is equivalent to 25.162 words and 126.292 characters.

The AMI corpus was originally published in the XML format. For this experiment, a version pre-converted into the JSON format was used.[2]

For the task of stance classification, a list of argumentative theses based on the data set and relevant expressions from the data set was created. For every thesis, relevant expressions were manually labeled as either agreeing with, disagreeing with or being neutral with regards to the corresponding thesis. The list of argumentative theses was generated by `claude-v1`. Theses generated by this model tended to be most concise, which in turn allows for easier manual labeling.

For every thesis, relevant expressions from the data set were selected using the following algorithm. Given an argumentative thesis t, the ordered set of sentences $S = \{s_0, s_1, ..., s_{n-1}\}$ sorted in chronological order and a semantic similarity function $sim(x_0, x_1)$ defined in Section 3.4:

1. Define the set $R = \{r_0, r_1, ..., r_{n-1}\}$ where $r_i = sim(s_i, t)$, $0 \leq r_i \leq 1$ and $r_i \in \mathbb{R}$ for all $i < n$;
2. Smooth the values in R using a simple moving average;
3. Find the index j of the largest value in R ;
4. Take $S_r = \{s_{j-k}, ..., sj + k\}$ as the most relevant continuous subset of S.

All labeling work was done by the author personally. Although it is generally desirable that such an evaluation is conducted by multiple evaluators, obtaining the required approval from the Human Research Ethics Committee (HREC) was not feasible given the limited time frame in which this research project took place.

3.2 Model details

Latent argument model

The latent argument model was trained for 1000 epochs with 5 topics and 2 viewpoints per topic. The number of topics was chosen by running a preliminary evaluation with a large number of topics, in which the model extracted five topics that were manually judged to be meaningful and the other topics extracted were either filled with noise or empty. The number of viewpoints was chosen to model either agreement or disagreement with a position on a certain topic.

For every topic-viewpoint combination, the top-2 relevant viewpoint sentences were extracted, resulting in a total of 20 sentences. The model was trained with the `LEX`, `POS` and `LEX_POS` switch strategies with both generative and discriminative sentence scoring. Training took approximately 1 hour per session on an Intel Core i7-9750H with 16GB of available RAM.

Large language models

As the computational resources required to locally perform inference were not available, large language models were accessed through either an API or a web application powered by a given large language model. For API-based access, the API's provided by Replicate and OpenAI were used.[3] For

[2] See: https://github.com/guokan-shang/ami-and-icsi-corpora

[3] For details on access, see: https://replicate.com/docs and https://openai.com/product

application-based access, the OpenPlayground, Google Bard and ChatGPT applications were used.[4] The large language models evaluated are listed along with their respective access method, fine-tuning method used and number of parameters in Table 1.

The size of the inputs to a large language model is limited by its maximum context window size. For thesis identification, the transcript was split into chunks of 1500 tokens. Each chunk was appended to a prompt, as shown in Figure 1. Splitting the transcript resulted in 25 chunks, each of which was used to generate an argumentative thesis.

3.3 Baseline approaches

To serve as a reference point for comparison, two simple baseline approaches are used:

1. Generate a random unrelated sentence using `claude-v1`;
2. Pick a random sentence from the transcript analyzed.

For the first baseline, the prompt *"Generate a random sentence"* is used. The set of sentences obtained mostly consists of descriptive phrases like "The old rusty car stood alone in the empty field" or "She looked out the window at the pouring rain". A full list of generated phrases is provided in Appendix A.

3.4 Evaluation metrics

The models tested are evaluated using a combination of novel and established metrics. To evaluate the thesis extraction capabilities of both the latent argument model and large language models, we propose several metrics for relevance, grammatical acceptability and conciseness. To evaluate the stance classification capabilities of large language models, the traditional classification metrics of accuracy, precision, and recall are used.

Relevance

To evaluate whether an argumentative thesis is relevant to a given transcript, theses are scored by their mean semantic similarity to a k-sized subset of the transcript. Recall that a conversation consists of spoken sentences $s \in S$. Given a thesis t and a semantic similarity function $sim(s_1, s_2)$, let $S_{t,k} = \{s_0, s_1, ..., s_{k-1}\}$ be the set of k most similar sentences to t in S. We then define the top-k semantic similarity of t as follows:

$$sim_{top\_k}(t, S_{t,k}) = \frac{1}{k} \sum_{i=0}^{k-1} sim\,(t, s_i)$$

For a set of n generated theses $T = \{t_0, t_1, ..., t_{n-1}\}$ the mean semantic similarity (MSS) of T to S is then defined as the mean top-k similarity of the elements of T to S:

$$mss(T, S) = \frac{1}{n} \sum_{i=0}^{n-1} sim_{top\_k}(t_i, S_{t_i,k})$$

The similarity function $sim(s_1, s_2)$ is defined as the cosine similarity of the sentence embeddings of s_1 and s_2. Sentence embeddings are computed using the `all-MiniLM-L6-v2` model, an embedding model based on the MiniLM [21] base model that performs well on the Massive Text Embedding Benchmark (MTEB) [22] and runs on consumer-grade hardware.[5] It encodes semantic properties of sentences and short phrases by mapping them to a 384-dimensional vector space.

For this experiment, a value of $k = 10$ is used. For the latent argument model, which extracts sentences from the transcript rather than generating a new sentence, the sentence extracted is left out of the set S_{sim}.

Grammatical acceptability

To evaluate whether a generated thesis is grammatically acceptable, the `RoBERTa-base-CoLA` model is used.[6] `RoBERTa-base-CoLa` is a version of the `RoBERTa` language model [23] fine-tuned on the Corpus of Linguistic Acceptability data set [24] to classify whether a given phrase is grammatically correct. The grammatical acceptability metric for this paper is defined as the fraction of theses generated by a model that is classified as grammatically correct:

$$GA = \frac{n_{corr}}{n_{total}} \tag{1}$$

`RoBERTa-base-CoLa` can classify phrases with a length of up to 512 tokens, which is equivalent to a length of 200 to 300 words. As argumentative theses should be concisely formulated, any thesis longer than 512 tokens is rejected as grammatically incorrect.

Conciseness

Theses generated should ideally be concise, as this makes them more intelligible to a human reader and makes it easier to evaluate the stance of an expression regarding a thesis. To evaluate the conciseness of theses generated by a model, the number of characters, words and sentences per thesis is observed and the average is taken over all theses generated.

4 Results

This section describes the results of the experiments described above. It first discusses the performance of the latent argument model and large language models on the task of thesis identification and subsequently discusses the performance of large language models on the task of stance classification.

4.1 Thesis identification

Full results are shown in Table 2. Figure 3 and Figure 4 compare the performance of base large language models in relevance and grammatical acceptability of theses generated.

With respect to conciseness, we observe that the claude-v1, claude-v1-instant, palm-2 and flan-t5-xxl models perform best, reliably providing single-sentence outputs. The text-babbage-001, text-curie-001, text-davinci-003 and gpt-3.5-turbo models and all versions of the latent argument model

[4] See: https://nat.dev, https://chat.openai.com & https://bard.google.com

[5] See: https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[6] See: https://huggingface.co/textattack/roberta-base-CoLA

Base model	Model name	Access method	Fine-tuning method	Number of parameters
Claude	claude-v1	OpenPlayground	RLHF	NPD
Claude	claude-v1-instant	OpenPlayground	RLHF	NPD
PaLM 2	palm-2	Google Bard	RLHF	NPD
GPT-4	gpt-4	ChatGPT	RLHF	NPD
GPT-J	dolly-gptj	Replicate	SIFT	6B
LLaMA	vicuna-13b	Replicate	SIFT	13B
Pythia	dolly-v2-12b	Replicate	SIFT	12B
Pythia	oasst-pythia-12b	Replicate	SIFT	12B
StableLM	stablelm-tuned-alpha-7b	Replicate	SIFT	7B
T5	flan-t5-small	Replicate	SIFT	80M
T5	flan-t5-xxl	Replicate	SIFT	11B
GPT-3	text-ada-001	OpenAI	RLHF	175B
GPT-3	text-babbage-001	OpenAI	RLHF	175B
GPT-3	text-curie-001	OpenAI	RLHF	175B
GPT-3	text-davinci-003	OpenAI	RLHF	175B
GPT-3	gpt-3.5-turbo	OpenAI	RLHF	175B

Table 1: Overview of large language models evaluated with their respective base models, access methods, and fine-tuning methods used. Fine-tuning methods are either supervised instruction fine-tuning (SIFT) or supervised instruction fine-tuning combined with reinforcement learning from human feedback (RLHF). The number of model parameters is listed in millions (M) or billions (B). For some models, the number of parameters is not publicly disclosed (NPD).

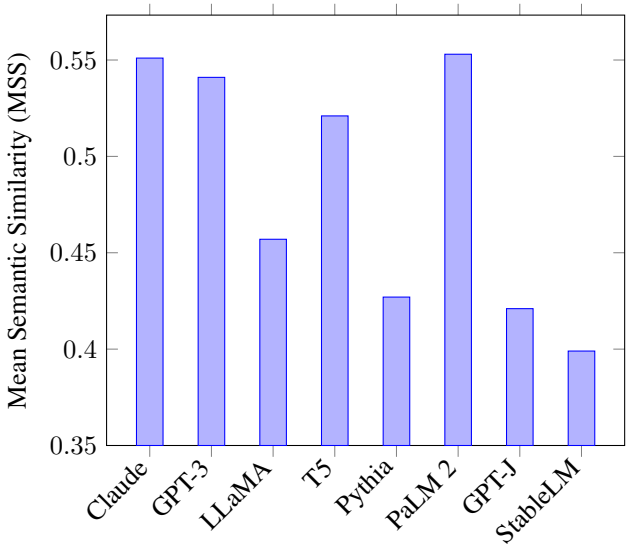

Figure 3: Comparison of mean semantic similarity (MSS) for the best-performing models based on different base models for the task of thesis identification

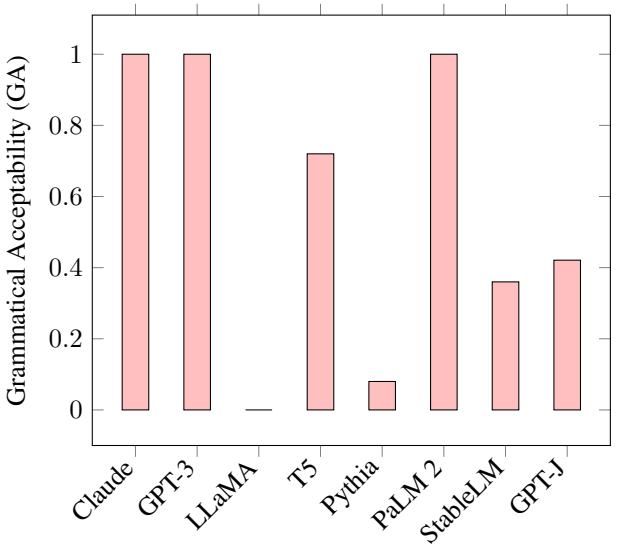

Figure 4: Comparison of grammatical acceptability (GA) for the best-performing models based on different base models for the task of thesis identification

produce output that is sometimes longer than a single sentence, but always remains of reasonable length. The models text-ada-001, vicuna-13b, stablelm-7b and flan-t5-small tend to produce short paragraphs rather than sentences and the dolly-12b and oasst-pythia-12b models consistently produce excessively long output.

With respect to relevance and grammaticality, we observe that the more advanced GPT-3 models, GPT-4, PaLM 2 and both Claude models are capable of writing argumentative theses that are both relevant to the meeting transcript and consistently grammatically correct. Next, we see that the T5-based models and lesser GPT-3 models tend to output theses that are semantically relevant, but not always grammatically correct. The latent argument model produces output that is semantically relevant but rarely correct grammatically and the LLaMA-, Pythia- or StableLM-based models produce output

Method	#CH	#WD	#ST	MSS	GA
Baselines					
random-unrelated	58	11	1.0	0.250	1.00
random-from-conversation	33	6	1.0	0.355	0.21
Latent argument model					
lam-lex-discriminative	143	27	1.5	0.490	0.15
lam-lex-generative	151	28	1.5	0.502	0.25
lam-lex_pos-discriminative	151	29	1.7	0.521	0.20
lam-lex_pos-generative	121	24	1.6	0.494	0.31
lam-pos-discriminative	129	26	1.5	0.481	0.19
lam-pos-generative	132	25	1.8	**0.541**	0.13
Large language models					
claude-v1	67	**10**	**1.0**	0.525	**1.00**
claude-v1-instant	105	15	**1.0**	**0.551**	**1.00**
text-ada-001	254	42	3.6	0.424	0.60
text-babbage-001	148	24	1.4	0.512	0.96
text-curie-001	165	27	1.3	0.514	0.88
text-davinci-003	123	20	1.1	0.536	**1.00**
gpt-3.5-turbo	187	29	1.2	0.541	**1.00**
gpt-4	130	19	**1.0**	0.507	**1.00**
vicuna-13b	1531	243	7.2	0.457	0.00
palm-2	80	14	**1.0**	**0.553**	**1.00**
dolly-12b	1761	261	14.7	0.421	0.12
oasst-pythia-12b	6310	940	36.0	0.427	0.08
stablelm-7b	734	98	3.9	0.399	0.36
flan-t5-small	177	36	5.8	0.486	0.80
flan-t5-xxl	**60**	**10**	**1.0**	0.521	0.72

Table 2: Comparison of lexical, semantic, and syntactic properties of identified theses by average number of characters (#CH), number of words (#WD), number of sentences (#ST), mean semantic similarity (MSS) and grammatical acceptability (GA)

that is significantly less relevant and lacks proper grammatical structure.

Considering all metrics, the large language models evaluated can be divided in three categories by their performance. The Claude, advanced GPT-3, GPT-4 and PaLM 2 models produce high-quality output, the T5 and lesser GPT-3 models produce output of reasonable quality and the LLaMA, Pythia and StableLM models produce low-quality output. The latent argument model is unique in that its outputs are highly relevant and concise, but often grammatically incorrect.

4.2 Stance classification

We observe that the majority of models evaluated is not capable of completing the stance classification prompt, often returning a continuation of the data set, an unnecessarily verbose answer or an incorrectly formatted list. Models capable of completing the instruction show poor classification performance, with the GPT-4 model performing somewhat better than the other models tested. An overview of models capable of completing the instruction such that their performance can be evaluated is found in Table 3. The classification performance of models evaluated is shown in Table 4.

Completes instruction	Does not complete instruction
claude-v1	text-ada-001
claude-v1-instant	text-babbage-001
gpt-3.5-turbo	text-curie-001
gpt-4	vicuna-13b
text-davinci-003	palm-2
	dolly-12b
	dolly-gptj
	oasst-pythia-12b
	stablelm-7b
	flan-t5-small
	flan-t5-xxl

Table 3: Comparison of large language models baed on whether they are capable of completing the instruction described in Figure 2, such that the output satisfies all requirements listed in the prompt

Base model	Model	ACC	$\mathbf{P}_a$	$\mathbf{P}_d$	$\mathbf{R}_a$	$\mathbf{R}_d$
Claude	claude-v1	0.60	0.10	**0.77**	0.01	0.13
Claude	claude-v1-instant	0.78	0.09	0.42	0.11	0.25
GPT-3	gpt-3.5-turbo	0.56	0.08	0.40	0.11	0.01
GPT-3	text-davinci-003	0.57	0.07	0.45	0.05	0.13
GPT-4	gpt-4	**0.88**	**0.38**	0.67	**0.16**	**0.50**

Table 4: Comparison of accuracy (ACC), positive precision (P_a), negative precision (P_d), positive recall (R_a) and negative recall (R_d) for the task of stance classification

5 Discussion, Limitations & Future Work

This section discusses how the observed performance difference between evaluated models can be explained, notes limitations on the approach and evaluation methods described, and presents suggestions for future research.

5.1 Discussion

Thesis extraction

We see that the latent argument model performs well at the task of thesis extraction, with the sentences extracted being concise and semantically relevant. Although they are rarely grammatically correct, this is not a property of the model itself but of the underlying transcript as the latent argument model only performs extractive summarization. This altogether means that topic modelling approaches remain viable to detect semantically relevant content.

The large language models evaluated show significant differences in performance. While proprietary state-of-the-art large language models perform well at thesis identification, competing open-source models significantly lag behind in identifying theses that are relevant, grammatically correct and concise. Although the ”black box”-nature of large language models makes it difficult to precisely determine why some models outperform others, we believe that the performance disparities observed can largely be explained by differences in model scale and the fine-tuning process used.

The first factor we consider is model scale, i.e. the number of parameters of a large language model. The number of parameters of the worst-performing models ranges between 6 billion and 13 billion, while the best-performing models have hundreds of billions of parameters.[7] Given that scaling up large language models is well-known to result in better performance [25; 26], this major difference in scale could reasonably explain the performance gap described. Difference in scale can however not explain all performance disparities: the 80-million parameter `flan-t5-small` model outperforms GPT-J, LLaMA, Pythia and StableLM-based models that are bigger by two orders of magnitude.

A second factor that can explain the observed performance disparities is the fine-tuning process used. Models fine-tuned using Reinforcement Learning with Human Feedback (RLHF) significantly outperform models that only use supervised instruction fine-tuning (SIFT). We believe that the RLHF process meaningfully increases model performance for the task of thesis identification. Consider i.e. that the GPT-3 based `gpt-3.5-turbo` model outperforms LLaMA-based `vicuna-13b` on all metrics, despite the claim by LLaMA's developers that its 13B-parameter version outperforms GPT-3 on most tasks [16]. Additionally, we conclude that the FLAN fine-tuning process is better suited to the task of thesis identification than the other supervised fine-tuning processes used, as evidenced by `flan-t5-xxl` and `flan-t5-small` being the best-performing SIFT models even while `flan-t5-small` only has 80M parameters.

We conclude that the performance difference seen in evaluated large language models can be explained by both the scale of the model, measured in the number of parameters, and the fine-tuning process employed. Notwithstanding, the performance of both the latent argument model and the `flan-t5-small` model demonstrate that the task of thesis identification can be performed reasonably well without relying on large-scale models.

Stance classification

All large language models evaluated either fail to complete the instruction or return inaccurate classifications. Even GPT-4, the best-performing model for stance classification, shows a sufficiently large number of false positives and false negatives that the model is not usable in practice. Similar to the performance difference observed for thesis identification, the performance difference for stance classification can largely be explained by differences in model scale and fine-tuning process used. Other factors contributing to the poor observed classification performance could include the inherent ambiguity of the task and the relatively low quality of the validation labels.

5.2 Limitations

While the approach presented in this paper offers some potential for detecting (dis)agreements in multi-party conversations, there are several limitations we wish to note.

A fundamental limitation of the approach used is that the (dis)agreement model described in Section 2.1 does not model all dynamics of a multi-party conversation. The model used takes a logic-based view on (dis)agreement that is solely based on the linguistic meaning of the phrases from the meeting transcript. This ignores many aspects of multi-party conversations, such as tone of voice, non-verbal communication and meeting dynamics like interruptions or speaker dominance. An example of a (dis)agreement moment that would not be detected by the modeling approach used is the common situation in which meeting participants agree on the substance discussed, but keep arguing their own viewpoint because they do not pay attention to what other participants are saying.

Another key limitation is that several aspects of the approach described, such as the algorithm to find a relevant subset described in Section 3.2 or the top-k mean semantic similarity metric described in Section 3.4, depend on a well-performing language embedding model to work properly. Future research could attempt to evaluate whether using a different embedding model results in similar or different results.

A final limitation is that the quality of the labels used for evaluation the stance classification task does not meet academic standards, as they were created by a single evaluator without using pre-defined guidelines. Any future research should incorporate a proper evaluation process, especially given the fact that the labels encode subjective opinions on which reasonable people can often disagree.

5.3 Future work

This section describes approaches that could be explored in future work regarding thesis extraction and stance classification.

Thesis extraction

To improve the thesis extraction performance of the open-source models named, a number of different methods could be used. First of all, it could be investigated whether a different method of prompting the models yields better results. This could be accomplished both by adjusting the phrasing of the instruction given or by providing one or more examples of a successfully completed instruction in the prompt. The former method is generally referred to as *prompt engineering*, while the latter is referred to as *few-shot learning*.

A second approach could be to specifically fine-tune a base large language model for the task of thesis extraction, perhaps while using the outputs of well-performing models like GPT-4 or Claude as reference output. Although this approach could yield significantly better results, it also would require considerable computational resources. Additionally, comparing the performance of a set of identical base models fine-tuned using different approaches could lead to a better understanding of how differences in the fine-tuning process affect task performance.

Further improving the performance of the proprietary models would first require a more comprehensive evaluation method. Theses generated are concise, relevant and grammatically correct by the metrics used in this paper and appear indistinguishable from human-written theses on manual inspection. Refining model performance for thesis generation

[7] Although this number is not publicly disclosed for the Claude, GPT-4 and PaLM 2 models, we can reasonably assume them to have between 100 billion and 1 trillion parameters.

would require a detailed evaluation framework that either incorporates more refined evaluation metrics or relies on expert human evaluators.

An approach that could be considered is that of combining the latent argument model with a well-performing large language model. The latent argument model can identify semantically relevant content on par with state-of-the-art large language models, but does not have the linguistic modeling capabilities required to rephrase it into a grammatically correct sentence. A large language model could be used to augment the described abilities of the latent argument model by rephrasing the content it identifies, rather than replacing it entirely.

Stance classification
To improve the process of classifying the stance of a phrase with respect to a thesis, we propose re-evaluating the problem formalization described in Section 2.1. The approach used in this paper reduces (dis)agreement to a ternary variable: agree, disagree or neutral. An approach capable of capturing partial (dis)agreement, perhaps by modeling (dis)agreement as a continuous rather than discrete variable, could be more suitable to model the problem.

We additionally recommend improving the algorithm used to select the most relevant subset of a conversation. The algorithm described is provisional and lacks a strong theoretical basis and was mainly introduced to reduce the manual labeling work required. In future work, either an improved selection algorithm could be used or more data could be labeled to reduce the need for a selection algorithm.

Furthermore, it should be investigated whether large language models can be prompted to perform stance classification with a simpler prompt than the one described in Figure 2. As most models evaluated were not capable of validly completing the instruction given, it would be worth investigating whether they perform better when provided with a differently phrased prompt.

Besides improving the method used, we also suggest applying the method described to different data sets, in particular to a data set in which speakers more explicitly articulate their viewpoints. Political debate transcripts would be particularly relevant: available data sets include the UN General Debate Corpus [27] or the Europarl Corpus [28].

Finally and most importantly, any future research on stance classification should incorporate a comprehensive data labeling process, that follows a pre-defined procedure and involves multiple human evaluators.

6 Responsible Research

6.1 Ethical considerations

As this paper deals with analyzing human ideas, perspectives and discussions, care should be taken to ensure that the methods described are not used in an irresponsible way. To ensure responsible use, it is important to remember that the model output is not perfect. It may ignore contentious issues or wrongly classify participants' opinion on them. Unwarranted trust in the model can lead to participants' opinions being misrepresented or valuable perspectives being ignored. Additionally, large language models are well-known to capture undesirable societal biases on a large number of subjects including race, gender, religion and profession [29; 30; 31]. To address these shortcomings, it is key to carefully check the analysis performed by a model rather than uncritically accepting it.

6.2 Reproducibility

All information necessary to reproduce the experiment is detailed in Section 3. The Python implementation of the experiment, including full results, will be made available on the TU Delft repository.

7 Conclusion

This paper proposes a novel approach for (dis)agreement detection, dividing it in the tasks of thesis identification and stance classification. For the task of thesis identification, the performance of the latent argument model by Vilares & He [5] is compared against that of a diverse set of large language models.

We find that, while state-of-the-art large language models do not outperform the latent argument model in extracting semantically relevant content, they are capable of presenting such content in a more concise and grammatically correct manner. Additionally, we observe major performance differences between large language models evaluated that can be explained by both differences in model scale and different fine-tuning approaches used. Finally, we show that all large language models evaluated perform poorly on the task of stance classification, with the best-performing models showing large numbers of false positives and false negatives and the worst-performing models failing to return valid output.

We propose several directions for future research. The performance of open source models on the task of thesis identification could be improved through prompt engineering, few-shot learning or specialized fine-tuning. Improving such performance for the best-performing models would first require a more comprehensive evaluation framework. With regards to stance classification, we recommend to focus on improving prompt quality and the evaluation framework used and to test the approach on a data set in which speakers take more explicit stances than in the data set used for this research project.

References

[1] K. Al Khatib, T. Ghosal, Y. Hou, A. de Waard, and D. Freitag, "Argument mining for scholarly document processing: Taking stock and looking ahead," in *Proceedings of the Second Workshop on Scholarly Document Processing*, (Online), pp. 56–65, Association for Computational Linguistics, June 2021.

[2] H. Xu, J. Šavelka, and K. D. Ashley, "Using argument mining for legal text summarization," *Frontiers in Artificial Intelligence and Applications*, vol. 334, p. 184 – 193, 2020.

[3] V. Chekalina and A. Panchenko, "Retrieving comparative arguments using deep language models," vol. 3180, pp. 3032–3040, 2022.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, no. 4-5, p. 993 – 1022, 2003.

[5] D. Vilares and Y. He, "Detecting perspectives in political debates," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, (Copenhagen, Denmark), pp. 1573–1582, Association for Computational Linguistics, Sept. 2017.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," vol. 2017-December, p. 5999 – 6009, 2017.

[7] B. Min, H. Ross, E. Sulem, A. P. B. Veyseh, T. H. Nguyen, O. Sainz, E. Agirre, I. Heinz, and D. Roth, "Recent advances in natural language processing via large pre-trained language models: A survey," 2021.

[8] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le, "Finetuned language models are zero-shot learners," 2022.

[9] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," 2022.

[10] T. Zhang, F. Ladhak, E. Durmus, P. Liang, K. McKeown, and T. B. Hashimoto, "Benchmarking large language models for news summarization," 2023.

[11] C. Ziems, W. Held, O. Shaikh, J. Chen, Z. Zhang, and D. Yang, "Can large language models transform computational social science?," 2023.

[12] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018.

[13] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," 10 2019.

[14] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," 5 2020.

[15] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, and N. Fiedel, "Palm: Scaling language modeling with pathways," 4 2022.

[16] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "Llama: Open and efficient foundation language models," 2023.

[17] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, N. Joseph, S. Kadavath, J. Kernion, T. Conerly, S. El-Showk, N. Elhage, Z. Hatfield-Dodds, D. Hernandez, T. Hume, S. Johnston, S. Kravec, L. Lovitt, N. Nanda, C. Olsson, D. Amodei, T. Brown, J. Clark, S. McCandlish, C. Olah, B. Mann, and J. Kaplan, "Training a helpful and harmless assistant with reinforcement learning from human feedback," 2022.

[18] OpenAI, "Gpt-4 technical report," 2023.

[19] R. Anil, A. M. Dai, O. Firat, M. Johnson, D. Lepikhin, A. Passos, S. Shakeri, E. Taropa, P. Bailey, Z. Chen, E. Chu, J. H. Clark, L. E. Shafey, Y. Huang, K. Meier-Hellstern, G. Mishra, E. Moreira, M. Omernick, K. Robinson, S. Ruder, Y. Tay, K. Xiao, Y. Xu, Y. Zhang, G. H. Abrego, J. Ahn, J. Austin, P. Barham, J. Botha, J. Bradbury, S. Brahma, K. Brooks, M. Catasta, Y. Cheng, C. Cherry, C. A. Choquette-Choo, A. Chowdhery, C. Crepy, S. Dave, M. Dehghani, S. Dev, J. Devlin, M. Díaz, N. Du, E. Dyer, V. Feinberg, F. Feng, V. Fienber, M. Freitag, X. Garcia, S. Gehrmann, L. Gonzalez, G. Gur-Ari, S. Hand, H. Hashemi, L. Hou, J. Howland, A. Hu, J. Hui, J. Hurwitz, M. Isard, A. Ittycheriah, M. Jagielski, W. Jia, K. Kenealy, M. Krikun, S. Kudugunta, C. Lan, K. Lee, B. Lee, E. Li, M. Li, W. Li, Y. Li, J. Li, H. Lim, H. Lin, Z. Liu, F. Liu, M. Maggioni, A. Mahendru, J. Maynez, V. Misra, M. Moussalem, Z. Nado, J. Nham, E. Ni, A. Nystrom, A. Parrish, M. Pellat, M. Polacek, A. Polozov, R. Pope, S. Qiao, E. Reif, B. Richter, P. Riley, A. C. Ros, A. Roy, B. Saeta, R. Samuel, R. Shelby, A. Slone, D. Smilkov, D. R. So, D. Sohn, S. Tokumine, D. Valter, V. Vasudevan, K. Vodrahalli, X. Wang, P. Wang, Z. Wang, T. Wang, J. Wieting, Y. Wu, K. Xu, Y. Xu, L. Xue, P. Yin, J. Yu, Q. Zhang, S. Zheng, C. Zheng, W. Zhou, D. Zhou, S. Petrov, and Y. Wu, "Palm 2 technical report," 2023.

[20] W. Kraaij, T. Hain, M. Lincoln, and W. Post, "The ami meeting corpus," 2005.

[21] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, "Minilm: Deep self-attention distillation for

task-agnostic compression of pre-trained transformers," 2020.

[22] N. Muennighoff, N. Tazi, L. Magne, and N. Reimers, "Mteb: Massive text embedding benchmark," 2023.

[23] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," 2019.

[24] A. Warstadt, A. Singh, and S. R. Bowman, "Neural network acceptability judgments," *arXiv preprint arXiv:1805.12471*, 2018.

[25] A. Maloney, D. A. Roberts, and J. Sully, "A solvable model of neural scaling laws," 2022.

[26] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, and W. Fedus, "Emergent abilities of large language models," 2022.

[27] A. Baturo, N. Dasandi, and S. J. Mikhaylov, "Understanding state preferences with text as data: Introducing the un general debate corpus," *Research & Politics*, vol. 4, no. 2, p. 2053168017712821, 2017.

[28] P. Koehn, "Europarl: A Parallel Corpus for Statistical Machine Translation," in *Conference Proceedings: the tenth Machine Translation Summit*, (Phuket, Thailand), pp. 79–86, AAMT, AAMT, 2005.

[29] M. Nadeem, A. Bethke, and S. Reddy, "Stereoset: Measuring stereotypical bias in pretrained language models," 2020.

[30] A. Abid, M. Farooqi, and J. Zou, "Persistent anti-muslim bias in large language models," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, (New York, NY, USA), p. 298–306, Association for Computing Machinery, 2021.

[31] J. Vig, S. Gehrmann, Y. Belinkov, S. Qian, D. Nevo, Y. Singer, and S. Shieber, "Investigating gender bias in language models using causal mediation analysis," in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds.), vol. 33, pp. 12388–12401, Curran Associates, Inc., 2020.

A Phrases generated for random-unrelated-sentence baseline

1. The old rusty car stood alone in the empty field.
2. The memory of her smile still brightened his day.
3. He wondered if he would ever find his way again.
4. She looked out the window at the pouring rain.
5. The little tree stood bravely in the winter snow.
6. He checked his watch for the tenth time that hour.
7. Her heart fluttered at the sight of the tiny cottage.
8. The smell of fresh pie wafted through the kitchen.
9. They hiked the winding trail through the forest.
10. She curled up with her favorite book on the sofa.
11. The moon glowed brightly in the night sky.
12. Birds chirped merrily in the leafy trees.
13. A cool breeze swept through the open window.
14. Mist rose from the lake in the early dawn.
15. Thunder boomed loudly overhead.
16. Apples dotted the trees in the orchard.
17. Waves crashed onto the sandy beach.
18. A deer peeked out from the line of trees.
19. Wildflowers bloomed along the dusty road.
20. Stars twinkled like diamonds against a black velvet sky.
21. An old clock ticked rhythmically on the wall.
22. Pancakes sizzled on the griddle.
23. A butterfly landed gently on the daisy.
24. Smoke curled upwards from the chimney.

B Sample model outputs for thesis identification

claude-v1 — Remote controls should be inexpensive to produce.

claude-v1-instant — Original remote controls provide higher value and appeal to consumers.

dolly-12b — The thesis that some participant(s) agree with states states that dogs are a desirable pet; they can stay agree with states that dogs are a desirable pet; they can stay with the family and enjoy plenty of physical affection. a desirable pet; they can stay with the family and enjoy plenty of physical affection. Some participants disagree stating that they are allergic to most agree with states that dogs are a desirable pet; they can stay with the family and enjoy plenty of physical affection. Some participants disagree stating that they are allergic to most pets and prefer not to deal with the upkeep states that dogs are a desirable pet; they can stay with the family and enjoy plenty of physical affection. Some participants disagree stating that they are allergic to most pets and prefer not to deal with the upkeep and maintenance associated with those kinds of pets. Both agree with states that dogs are a desirable pet; they can stay with the family and enjoy plenty of physical affection. Some participants disagree

stating that they are allergic to most pets and prefer not to deal with the upkeep and maintenance associated with those kinds of pets. Both parties appear to agree that roosters belong in some participant(s) agree with states that dogs are a desirable pet; they can stay with the family and enjoy plenty of physical affection. Some participants disagree stating that they are allergic to most pets and prefer not to deal with the upkeep and maintenance associated with those kinds of pets. Both parties appear to agree that roosters belong in the background instead of taking centre stage during conference calls.

flan-t5-small — 1: The project is gonna be a few minutes of the project.

flan-t5-xxl — The project brief is to design a new remote control.

gpt-3.5-turbo — The project team aims to design a trendy and user-friendly remote control that can be sold internationally for 25 Euro with a production cost of no more than 12.50 Euro per unit, to generate 50 million Euro in revenue, which some members of the team agree with and others may disagree with.

gpt-4 — The production cost of the new remote control should not exceed fifty percent of its selling price.

latent argument model — I'm thinking the price might might appeal to a certain market in one region, whereas in another it'll be different, so

oasst-pythia-12b — Here is a short essay on my favorite animal, animal, a Beagle.

A Beagle favorite animal, a Beagle.

A Beagle is such a loyal friend and companion! Their my favorite animal, a Beagle.

A Beagle is such a loyal friend and companion! Their loyalty knows no bounds and they love to short essay on my favorite animal, a Beagle.

A Beagle is such a loyal friend and companion! Their loyalty knows no bounds and they love to please their owners unconditionally. Beagles animal, a Beagle.

A Beagle is such a loyal friend and companion! Their loyalty knows no bounds and they love to please their owners unconditionally. Beagles also possess a mischievous side that makes my favorite animal, a Beagle.

A Beagle is such a loyal friend and companion! Their loyalty knows no bounds and they love to please their owners unconditionally. Beagles also possess a mischievous side that makes them quite entertaining to watch. However, despite their playful behavior, beagles are generally gentle favorite animal, a Beagle.

A Beagle is such a loyal friend and companion! Their loyalty knows no bounds and they love to please their owners unconditionally. Beagles also possess a mischievous side that makes them quite entertaining to watch. However, despite their playful behavior, beagles are generally gentle and non-violent unless annoyed or scared. animal, a Beagle. *[REST OF RESULT OMITTED]*

PaLM 2 — The new remote control should be designed to be both user-friendly and trendy.

StableLM — The Argument: For many customers, agreeing on a common theme such as identifying personal preferences while having conversations through business meetings means on a common theme such as identifying personal preferences while having conversations through business meetings means sharing ideas readily and creating meaningful relationships between team members.

text-ada-001 — The thesis is that in order to be successful in project management, one's team should be knowledgeable of their project, its stages, and theirMemo's.

text-babbage-001 — Some dialogue participants agree with the thesis that a remote control should be designed for Twelve Fifty Euros, while others disagree.

text-curie-001 — The designers of a new remote control project agree that their favourite animal is a beagle, have strong opinions about fish, MONKEYS, and ROOSTERS, and disagree on the cost of their product.

text-davinci-003 — Our goal is to design and produce an original, trendy, and user-friendly remote control that is sold internationally for 25 Euros and costs no more than 12.50 Euros in production.

vicuna-13b — Thethe dialogue disagree about the selling price of the new remote the selling price of the new remote the selling price of the new remote control being discussed, with some believing it will be sold for twenty-five new remote control being discussed, with some believing it will be sold for twenty-five new remote control being discussed, with some believing it will be sold for twenty-five Euros and others believing it will be sold for wholesale or retail price.

C Theses used for stance classification

1. A curvy, ergonomic shape in vibrant yellow should be adopted for the remote.
2. A jog dial control would improve the remote control design.
3. Incorporating an LCD screen and scroll wheel in place of many buttons improves the remote control design.
4. Physical meetings are superior to virtual ones.
5. Play-Doh fosters creativity.
6. Rechargeable batteries should be included in the remote control.
7. Special colours are necessary for this multifunctional smart button.
8. The design of remote controls should prioritize aesthetics over function.
9. The structured nature of the leadership model restricts creativity.
10. The use of visual communication tools enhances collaborative work.
11. Voice recognition should be incorporated into remote controls.