Automatic text-based speech overlap classification: A novel approach using Large Language Models

Jan Domhof[1]

Supervisor(s): Catholijn Jonker [1]**, Morita 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: Jan Domhof
Final project course: CSE3000 Research Project
Thesis committee: Catholijn Jonker, Morita Tarvirdians, Mathijs Molenaar

Abstract

Meetings are the keystone of a good company. They allow for quick decision making, multiple-perspective problem solving and effective communication. However, most employees and managers have a negative view on the efficiency and quality of their meetings. High quality meetings where every participant feels equally heard and respected is crucial for having positive meeting sentiment within a company. One of the most influential aspects of meetings are speech overlaps. Overlaps range from short utterances such as backchannels, to follow up questions and clarifications, to complete interruptions. In non-competitive cases, the overlapped speaker feels that the other participants are listening and actively engaging with them during the meeting. In competitive cases, the overlapped speaker can feel interrupted and unimportant. Therefore, competitive overlaps often have a negative impact on the course of the discussion and the overlappee's meeting sentiment. In problematic cases, these overlaps should be reduced to a minimum. In order to do this, overlaps must be classified as either competitive or non-competitive. This paper proposes a novel approach to overlap classification, namely that of text-based classification through Large Language Models. Four different prompt designs are used and tested on the two best performing and publicly available models, GPT-3.5-turbo and GPT-4. The results show that the in-context learning approach using the GPT-4 model results in the most accurate classifications. When comparing the results to previous work, it is observed that the text-based GPT-4 model matches carefully engineered neural networks that even adopt a multi-modular approach.

1 Introduction

As the amount of meetings continues to grow, so does the need to improve their quality. During meetings, employees can combine their expertise, opinions and ideas to come to a better conclusion than any individual could on their own [1]. Employees and managers at companies around the world have, on average, 3.2 meetings every week [2]. However, 71% of managers said meetings are unproductive and inefficient and as the quality of meetings improved, satisfaction with work/life balance rose from 62% to 92% [3]. Besides work/life balance, high quality meetings could be imagined to lead to (1) an increase in meeting efficiency and productivity and therefore a reduction of time spent in meetings, (2) an improvement of well-being and intercollegiate relationships leading to an improvement of the working environment and (3) an improvement of the general employee morale, since low-quality meetings result in negative and pessimistic perspective on meetings [4]. Competitive overlaps are among the features that have the most negative impact on meeting sentiment. They are deemed inappropriate social behavior, as they are disrespectful, rude, and confrontational [5]. While the overlapper might have good intentions, it unpleasantly interrupts the overlappee and also takes away the speaking opportunity for party's that are patiently awaiting their turn. Methods to achieve this improvement often require consultation of an expert with a deep understanding of group dynamics to pay close attention to patterns and dynamics during the meeting. The strategy proposed in this paper is the usage of an automated tool that uses a Large Language Model (LLM) to classify overlap occurrences as either competitive or non-competitive. The tool displays these overlaps is a basic dashboard, such that meeting moderators can use this dashboard to analyse their meetings and act on the insights gained from the analysis.

Recent advances in NLP have shown ground-breaking results in an incredibly wide variety of tasks, as Bang et al. [6] and Qin et al. [7] showcase in their analyses of the impressive capabilities of ChatGPT[1]. These constant improvements in performance are a result of the application of a concept called self-attention. Vaswani et al. [8] introduced self-attention into the world of NLP in 2017, which was quickly adopted into the state-of-the-art models to date. The reason for that attention is because it enables models to model an understanding of any language (e.g. spoken languages, programming languages). This ability to model the meaning of words, in itself, is not enough to create a powerful understanding. As Liu et al. [9] point out, it is rather the combination of self-attention and Reinforcement Learning from Human Feedback (RLHF) [10, 11] what really drove the surge in performance, success and popularity. With these combined techniques, LLMs are now able to create new information that they haven't been trained on. For example, a model using self-attention can successfully write a snippet of code in the Julia programming language which is very unlikely to exist anywhere on the internet [12]. While the public is still discovering the unfathomable capabilities of ChatGPT, new and improved models are being released on a monthly basis. During this year, the state-of-the-art models (e.g. GPT-4 [13], LLaMA [14], LaMDA [15], BLOOM [16]) are shown to already outperform the now seemingly inaccurate ChatGPT. These ongoing and rapid improvements widen the horizon for possible applications of LLMs on tasks other than text generation. While the technology is not quite there yet, as Ziems et al. [17] have shown, it might only be a matter of time before the state-of-the-art models can replace humans on certain time-consuming and labor-expensive tasks.

The research questions that is answered in this paper is the following: **What is the performance of the GPT3.5 and GPT4 models in overlap classification relative to human annotations?** The hypothesis guiding this research is that, given the recent surge in performance and inferring power, the models will exhibit high accuracy, while not performing as accurate as human annotators, given the limitation to solely text-based data. The intention of this work is to investigate the current state of these LLMs on generating automatic annotations on overlapping speech. Ideally, this work will act as a motivation for a bigger investigation into the classification

[1] https://openai.com/blog/chatgpt

performance of LLMs. Consequently, it could ultimately be used to automatically analyse and annotate meetings, driving up meeting quality and overall meeting sentiment within the industry.

The structure of this paper is as follows. Section 2 covers related work. Then, the chosen methods and features are described in section 3. Section 4 contains a description of the dataset and models that are used, followed by the results of the experiments. The ethical aspects of this work are discussed in section 5, as well as the reproducibility of the experiments. The paper finishes up with a discussion and recommendation for future work in section 6 followed by a conclusion in section 7.

2 Related Work

Automated meeting analysis often adopts a multi-modal approach, leveraging corpora such as the AMI corpus [18] and combining the usage of verbal, visual and acoustic cues to make the most accurate predictions. Logically, this approach yields the best results, usually because the accuracy of a model grows with the amount of data it has to its disposal. A question that remains unanswered is how much data we can extract only from the verbal cues. Gutierrez et al. [19] showed that the most common method for text-based behavior detection is Natural Language Processing (NLP), and that the main focus of this research field has been on emotion and empathy detection. The articles they include in their analysis [19] cover applications of NLP on tasks ranging from predicting suicide risks to customer profiling. However, not much text-based research has focused their attention on analyzing meeting transcripts, leaving this area as a promising research topic.

Chowdhury et al. [20] show the affect of overlaps on user-satisfaction in dialogue. They define user-satisfaction as the satisfaction of the customer after a call-center conversation. While this is not the same application as this work, they show both the positive effect of non-competitive overlaps as well as the negative effect of competitive overlaps. Chowdhury et al. inspired the main focus of this work to be on overlap classification.

Chowdhury is one of the main contributors to this research area in the last decade. In another one of her papers, [21], the authors investigated automatic classification of speech overlaps. They adopted both linear and non-linear approaches, using models such as Long Short-Term Memory networks [22], various neural networks (e.g. Convolutional Neural Networks, Feed-Forward Neural Networks) as well as Support Vector Machines. The architectures of these models can all be found in [21] and will not be further touched upon, since the focus of this work is on Large Language Models. An interesting conclusion from their work is that all aforementioned models are more precise in classifying non-competitive overlaps than competitive overlaps, making the non-competitive class the main driver of the F1 score that they report. Other works in the area of automatic overlap classification, [21, 23–26], use neural networks, SVMs or other classifiers and also display this characteristic. The performance of Large Language Models on these types of classifications still remains unexplored which defines a knowledge gap that this work intends to fill.

Dong et al. [27] present a broad overview of the related works around in-context learning (ICL). This overview covers everything from demonstration designing to instruction formatting. While the methods showcased in the works that are covered in the reviews are not used in this work, it should be noted that the presented results could certainly be improved by using the methods covered by Dong et al.

Finally, Indira Sultanic [28] surveyed 112 language interpreters, and explored the effects of remote meetings on turn-taking. From the survey, she concludes that remote meetings contain significantly more overlap. This overlap is result of many factors, including technical issues (e.g. delays, freezing screens), participants joining the meeting from distracting environments (e.g. a moving vehicle) making them more distracted and less focused on the meeting. Besides, the lack of video makes it significantly more difficult to anticipate each speaker's turn. Sultanic's findings [28] are included in this section to show that care has to be taken when comparing meetings to each other. Since especially in a remote setting, there are many factors which directly influence turn-taking patterns and behavior.

In conclusion, the recent developments create knowledge gaps in the widening capabilities of Large Language Models and in which fields they can be applied. While there is no keeping up with the speed of improvement of these models, this work will try to fill the aforementioned knowledge gaps in LLMs' capabilities in classifying overlaps in multi-party meetings.

3 Methodology

3.1 Overlap Classification

In order to classify an overlap as competitive or non-competitive, a formal definition is required to identify which segments are overlaps and which are not. In this work, the definition of an overlap as defined in the annotation guidelines in Chowdhury et al. [29] are used and listed below:

1. An overlap is an interval where multiple speakers are speaking simultaneously.
2. An overlapping segment may contain more than one overlap instance of the same category. Instances may be separated from each other with a gap less than 40ms.
3. If a speaker thinks aloud during another speaker's turn, that is considered an overlap instance.
4. Co-occurrences of "false start" by both the speakers are considered instances of speech overlap if and only if the segments contain complete words and the annotator can infer the speaker's intention on the basis of the perceived intonation of speech.
5. An overlap should not contain poor quality audio, unintelligible speech, background noise or human sounds like cough, sneezes and laughs.

Using the above guidelines, the overlap occurrences can be carefully identified. The fifth guideline above is due to the disturbance within the acoustic data. Since this work

is only focused on the lexical (text-based) features, we exclude overlaps where the markers (e.g."<vocalsound>", "<disfmarker>", <coughing>) heavily disrupt the overlapping text. Note that, if a participant laughs or coughs during a speakers turn, this does not disrupt the annotated text of the speaker, since the voices are recorded separately.

Now that occurrences of overlaps can be identified, they should classified as either competitive or non-competitive. In order to make the distinction between these classes, once more, the guidelines from Chowdhury et al. [29] are followed, which are quoted below. Competitive overlaps are scenarios where:

> 1) The intervening speaker starts prior to the completion of the current speaker, 2) both the speakers display interest in the turn for themselves, and 3) speakers perceive the overlap as problematic.

And non-competitive overlaps are scenarios where:

> 1) Another speaker starts in the middle of an ongoing turn, 2) both parties do not show any evidence for grabbing the turn for themselves, 3) speakers perceive the overlap as non-problematic and 4) speakers use it to signal the support for the current speaker's continuation of speech.

These guidelines form the base of this research, which should be in-line with previous work and help with data-annotation which is covered in section 4.2.

3.2 Prompt Engineering

One of the advantages of Large Language Models over other Artificial Intelligence models, is that they can be programmed by natural language. This enables any person with a basic understanding of their language to be able to provide instructions on how models should behave, making them very accessible. Through effective prompt engineering, a user can instruct the language model to perform a much wider variety of tasks. Since large language models have recently taken the world by storm, a lot of research[2] is being done on prompt engineering techniques. To stay in line with related works, the prompts that are presented in this section follow the best-practices guidelines used in Ziems et al. [17] with a slight modification; (1) classification options must be in order (numerical or alphabetical) and each option shall be on a new line, resembling the natural format of online multiple choice questions; (2) instructions and constraints are given before the context in order to lay emphasis on the context, since recent text has a greater effect on common attention patterns; (3) the expected output format should be clarified, especially in cases where there might be uncertainties in the response; (4) answers that should contain multiple pieces of information are requested to be responded in JSON format, to ensure that the answer is parsable. These guidelines help to ensure that the model generates consistent, machine-readable outputs for the different classification tasks. The final prompts used to generate the results are included in Appendix A and quickly explained below:

1. **simple_prompt**. A very short and quick explanation is given.
2. **explain_prompt**. The definitions of the overlap classes, as mentioned in section 3.1, are given in the prompt to assist in the classification process.
3. **one_shot_prompt**. A short instruction is given, followed by one example of each class.
4. **few_shot_prompt**. A short instruction is given, followed by five examples of each class.

A technique that, in combination with prompt engineering, further boost the performance of LLMs is called in-context learning. The one_shot_prompt and few_shot_prompt use this technique. The idea of in-context learning is to provide the model with a few examples of instances of the classification problem, as well as the desired target class. These examples are included in the prompt, and will provide the model with an idea on how to behave. Few-shot-learning (FSL) generally outperforms the two alternatives, zero-shot-learning (ZSL) and one-shot-learning (OSL), and Brown et al. [30] indicate that this increase in performance grows with the amount of model parameters. Thus, it can be concluded that larger models make increasingly efficient use of in-context information. This is a highly valuable characteristics for classification tasks, since classification depends solely on the in-context information. In this paper, we use the GPT-3.5 model, which has 175 billion parameters, as well as the GPT-4 model with a staggering amount of 170 trillion model parameters. The results section in this paper showcases the increase in performance caused by this massive increase in parameters, as well as the importance of well engineered prompts.

3.3 Evaluation

In order to evaluate the performance of the model, the model's output is compared to the annotations made by the author. Most relevant work use precision, recall and F1 score to report the prediction performance. The formula's for these metrics are found below and will be used to compare the findings of this paper to previous works. In the formula's below, TP, FP, TN and FN are abbreviations for true positive, false positive, true negative and false negative respectively.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

$$\text{F1} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

To analyse the influence of different prompt designs on the performance of the model, precision, recall and F1 score will be compared as well as the rates for TN, TP, TN and FN classifications to gain a more in-depth insight into which classification errors are being made.

[2] https://www.promptingguide.ai/papers

4 Experimental Setup and Results

4.1 Materials

The findings of this paper are incorporated into an application that takes a transcript as input, and generates an accurate summary and more analytical data, including an overlap analysis. This application will be applied mostly to company meetings, which are task oriented and focus on finding solutions to that company's problems. Therefore, it would be prudent to use a corpus with a similar nature.

Corpus

Large Language Models can generate an output without the need of similar training data. However, in order to test evaluate the accuracy of the output, we need to compare the model's output to the "correct" answer. Most high-quality related work use professionally annotated corpora. The ICC corpus was used in [20, 21, 23, 24, 29] or the Davero corpus in [25, 26], both consisting of two-party dialogues. These corpora would be excellent to test the classification accuracy of LLMs on conversations where two parties are involved, useful for example for dialogue systems or conversational agents. However, as mentioned before, the findings of this work are incorporated in tool which summarizes and analyses multi-party meetings. Multi-party meetings are even more abstract and unstructured than two-party conversations, since in multi-party meetings there can be more than two speakers overlapping each other at the same time. Because of this reason, the aforementioned corpora are deemed inappropriate for this work. Nevertheless, the author of this paper recommends future work in the research area to adopt these corpora. Instead, the AMI corpus [18] was selected in order to generate the results on multi-party meetings similar to the data that will be fed to the tool that incorporates the findings of this work.

The AMI corpus consist of meetings in which the participants discuss different aspects of the design of a new type of television remote control. During the meeting, each participant has an role, i.e. project manager (PM), marketing expert (ME), user interface designer (UI) or industrial designer (ID). Before the meeting, each participant received some information which they have to present to the other participants, as well as some hints and guidelines on how to do their "job". This enables the participants to do their "job" while lacking knowledge and experience. After each presentation, the participants share their opinions about the topics presented in the presentation and decide on those topics by means of a group discussion.

Format

Originally, the transcripts are formatted in an XML file. In order to feed the transcript to a language model, these XML files must be processed into a readable text format. Luckily, Guokan Shang[3] created a repository[4] which performs this preproccessing step for the AMI corpus. This process is as follows. The original corpus contains an XML file containing all spoken words, as well as an XML file containing the segments, where each segment contains pointers to the words that belong to that segment. These are then combined and saved in JSON objects, see Listing 1, that are easier to handle and iterate on for performing annotations and to use in prompts.

Listing 1: An example segment in JSON format from the preprocessed AMI corpus

```
{
    "id": "ES2002a.B.dialog-act.dharshi.3",
    "speaker": "B",
    "starttime": "55.415",
    "startwordid": "ES2002a.B.words4",
    "endtime": "60.35",
    "endwordid": "ES2002a.B.words16",
    "text": "<vocalsound> Um well this is the
        ↪ kick-off meeting for our our project
        ↪  .",
    "label": "inf",
    "attributes": {
      "reflexivity": "true",
      "role": "PM",
      "participant": "FEE005"
    }
}
```

Model and Parameters

The two models that are used in this work are the two best performing models which are publicly available; GPT-3.5 and GPT-4. The GPT-3.5 model can be used through OpenAI's API. This is useful, since the API enables the results to be generated automatically. At the time of writing this paper, there is limited access to the API of the GPT-4 model, meaning that the prompts had to be fed manually to the model through ChatGPT, which utilizes GPT-4. As mentioned in the API documentation[5], there is a list of parameters that can be changed when requesting an API completion. For this work, two of these parameters were changed, namely temperature and max_tokens. Temperature, a parameter for the amount of randomness in the models response, is set to a low 0.1, in order to get the most consistent and deterministic results, which should also enhance the reproducibility of the presented results. The max_tokens parameter is set to 1, since we want our model to return either "0" (non-competitive overlap) or "1" (competitive overlap). These two settings combines result in the least amount of hallucinations, which is ultimately what we want to reduce. A final parameter that was changed is the stop sequence, which is set to the newline character (i.e. "enter"). Setting this parameter to the newline character restricts the model from returning a multi-sentence response, since it automatically stops generating when entering the newline character.

4.2 Annotation

Since the AMI corpus does not contain annotations on overlaps, this must be done manually. Due to the time constraints

[3] https://www.linkedin.com/in/guokan-shang

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

[5] https://platform.openai.com/docs/api-reference/completions/create

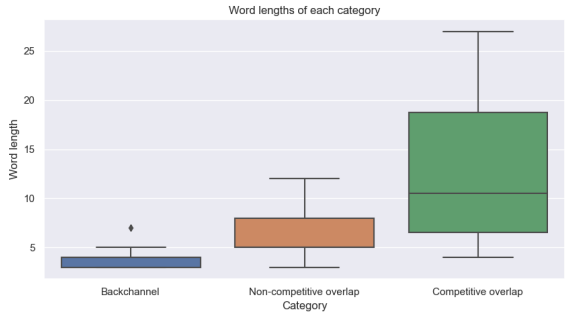

Figure 1: Boxplot of the word lengths for segments from the ES2012b meeting, for backchannels, non-competitive overlaps and competitive overlaps.

of this work, the decision is made that the author of this paper must perform the annotations. While this does not align with the standard procedures and guidelines of data annotation, this was the only feasible solution for this work. An analysis of the downsides and limitations of this decision, as well as the complete motivation and explanation can be found in section 5. To remain in-line with relevant work, the annotations that the author of this work made follow the guidelines specified in section 3.1. Prior to overlap classification, the overlaps need to be located within the transcript. As a first selection, all segments where that segment's start-time is before the end-time of the previous segment are analysed. Taking meeting ES2012b[6] as an example, this method results in 222 intervals of overlapping speech. Figure 1 shows a boxplot for the word lengths of the different categories. As can be seen, backchannels (e.g. "uhm", "uh-huh", "okay", etc.) almost always consist of 2 to 5 words. In order to speed up the annotation process, all instances of overlapping speech with less than 3 words were omitted from the annotation process. This insures that almost all the overlaps are processed, while a part of the backchannels and non-verbal vocalizations (e.g."<vocalsound>", "<disfmarker>", etc.). This reduces the number of overlapping speech occurrences in meeting ES2012b from 222 to 118, speeding up the annotation process significantly. During the annotation process, 4 meetings were annotated for overlaps, with each a duration of 30-40 minutes. These meetings contain a total of 110 overlaps, of which 53 are competitive and 57 are non-competitive. The annotated files can be found in this work's corresponding repository[7] in the folder: /corpora/ami/processed/.

4.3 Results

Taking a close look at table 1 reveals that, while the difference for the GPT3.5 model is smaller, the models score better on classifying non-competitive cases, which agrees with the results in [21]. It should be noted that in [21], they state that this is due to the imbalance in the training data. In their dataset, there are around 3 times as many non-competitive overlaps as there were competitive overlaps. Since their model is trained on the test set from their dataset, it might indeed explain this difference. LLMs, however, are not specifically trained on a dataset of overlap instances, meaning that the skewness in the results must have a different cause.

Model	Prompt	C	NC	Macro-F1	Std. Dev.
GPT3.5	simple	0.58	0.61	0.590	**0.024**
	explained	0.58	0.57	0.490	0.096
	one_shot	0.54	0.66	0.558	0.197
	few_shot	**0.70**	**0.69**	**0.678**	0.069
GPT4	simple	0.63	**0.94**	0.675	0.062
	explained	0.66	0.72	0.683	0.092
	one_shot	0.65	0.81	0.708	0.091
	few_shot	**0.66**	0.80	**0.716**	**0.035**

Table 1: Precision score for the competitive (C) and non-competitive (NC) classes, macro-averaged F1 score and standard deviation of each prompt design for the GPT3.5 and GPT4 models.

Figure 2 shows the influence of prompt designs on classification performance varies for the GPT3.5 and GPT4 model. While the **simple_prompt** design was intended to be used as a simple baseline design, the resulting scores are surprisingly high and consistent, with a standard deviation of 0.024 and 0.062, as can be seen in table 1.

One of the most interesting findings can be observed from the results of the **explain_prompt** design. From the plots in figure 2, it appears that GPT3.5 fails to use clear definitions of both overlap classes to infer the class of a given overlap. Even more surprising is the fact that it performs worse than the simple_prompt design, which contains a very short and primitive description of both classes. While it is the worst-performing prompt design for the GPT3.5 model, it can be seen that GPT4 is already a lot better at using the explanations to accurately infer the overlap class.

For the GPT3.5 model, the **one_shot_prompt** design heavily underperforms, with a worst F1 score of around 0.27, making it too inconsistent for real-life use. However, this prompt design works better with the GPT-4 model. Inconsistency is still an issue, but the average F1 score increases with 0.15 from 0.558 to 0.708.

Finally, the **few_shot_prompt** design is the best performing prompt design for both models for the overlap classification task. The combination with the GPT-4 model results in the highest average F1 score of 0.72, and low standard deviation of 0.035, making this the best solution for real-life usage of this technology.

5 Responsible Research

5.1 Annotation Quality

The annotations in this work are done by the author. It is noted that this is not a responsible approach due to many reasons. As Tseng et al. [31] state, before annotated data are submitted for downstream use, their quality, validity and reliability have to be assessed first. The annotations presented in this work would not pass the requirements for responsible

[6] Meeting ES2012b is the second scenario meeting from the 12^{th} group from the Edinburgh set and can be downloaded at https://groups.inf.ed.ac.uk/ami/download/

[7] https://gitlab.tudelft.nl/mtarvirdians/meeting-mastery/-/tree/main

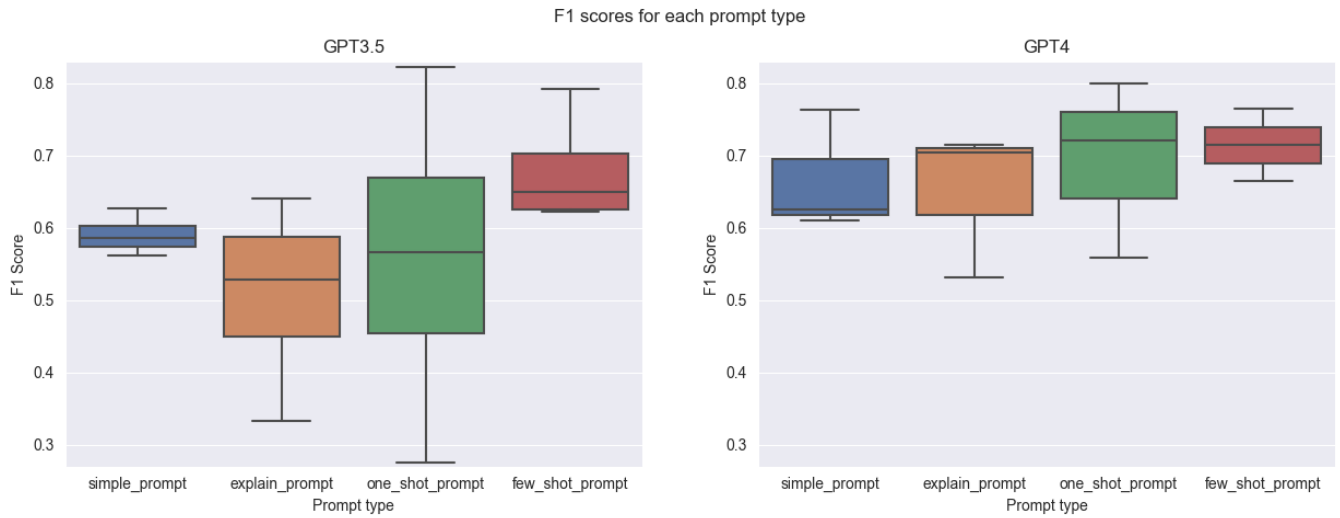

Figure 2: Boxplots that showcase the influence of prompt engineering techniques on the F1 score of the overlap class predictions. The left plot contains data points generated with the GPT3.5 language model. The right plot contains data points generated with the GPT4 language model. Each of the boxes in these plots contain the F1 scores of predictions of 4 meetings from the AMI corpus. These 4 meetings contain a cumulative of 110 overlap occurrences, of which 53 are competitive and 57 are non-competitive. The complete prompts that were used can be found in appendix A.

annotations, which has to be taken into account when reading this paper. While there are no well-defined guidelines on the amount of annotators needed to label data. It is good practice to do so in order to measure annotation agreements and to collect the most accurate labels. At the very least, the quality of annotations of any work should be similar to that of one expert annotator. To emulate expert level quality, an average of four non-expert annotators are required [32]. The author of this work is considered a non-expert annotator, meaning that this work does not conform to those standards. It also introduces a potential bias in the annotations. The author could, even subconsciously, be influenced to base the annotations on the question: What would the model return?

The decision that the author should perform the annotations was driven by the time constraints on this work. Since this work is a bachelor's thesis, a total of 10 weeks was available for the entire research cycle. To conform with the TU Delft standards[8], research where human participants are the source of the data should get approval from the Human Research Ethics Committee (HREC). This HREC application usually consists of the three documents; (1) An HREC checklist that has been signed by the responsible researcher; (2) Completed informed consent materials; (3) A complete data management plan. As mentioned on the TU Delft website, the duration of this approval process takes around 4-6 weeks. Given the fact that this project has a short duration of only 10 weeks, this would not make this a feasible solution for this work, which is why another solution was required. Because of the aforementioned motivation to use a multi-party meeting corpus, there was no other feasible option apart from the annotation process that was used in this work.

[8] https://www.tudelft.nl/en/about-tu-delft/strategy/integrity-policy/human-research-ethics

5.2 Responsible Use

Ideally, this technology will eventually be incorporated by actual companies or institutions to assist in real life situations. Since it will then affect the workflow and life of actual employees, ethical concerns arise of which a detailed analysis follows. Based on the data that the tool produces, moderators will form conclusions about the meeting participants participation throughout the meetings. While the average employee will not be affected as much by the results of this tool, above-average performing participants will stand out, as well as the below-average performing participants. Take as an extreme example a scenario where there is an employee that does not actively participate in meetings. The findings of this tool might motivate the company's management to reconsider a contract renewal, whereas this might not be the case if the company never adopted this tool to their workflow. Still, this might not be an issue if the tool output only numerical values (e.g. turn duration, amount of turns, etc). These values can be seen as factual and basically represent the events of the meeting. However, issues arise when there are uncertain predictions made which influence the data and consequently the employee. In order to create transparency and to minimize the negative impact of this tool, the user must be informed that the tool produces uncertain and, to some extent, random data and classifications. Further, the user should be instructed to confirm the findings of the tool before acting on the produced results. To account for these actions, the tool warns and informs the user when they perform an analysis and keeps track of all overlap instances, which can then easily be iterated and verified by the user.

5.3 Reproducibility

This work also introduces concerns about reproducibility. There are two aspects of this work which raise these concerns

Citation	Corpus	Features	Method	F1 Score
Chowdhury et al. [23]	ICC	Acoustic, lexical, part-of-speech, psycholinguistic	SVM	0.66
Egorow & Wendemuth [25]	Davero	Acoustic, emotional	SVM	0.72
Chowdhury & Riccardi [24]	ICC	Acoustic, lexical	DNN	0.68
Chowdhury et al. [21]	ICC	Acoustic , lexical	FFNN	0.70
Egoro & Wendemuth [26]	Davero	Emotional	NBC	0.70
This work	AMI	Lexical	LLM	0.72

Table 2: A comparison of previous work. For each work, the used corpus, features, method and F1 score are given. The abbreviations for the methods are, from top to bottom, Support Vector Machine (SVM), Deep Neural Network (DNN), Feed-Forward Neural Network (FFNN), Naive Bayes Classifier (NBC) and Large Language Model (LLM).

around reproducibility; the state of the models which generate the results, and the annotations against which the model's output is compared. The former concern is raised because of the rapidly changing industry around Large Language Models. In order to reproduce the findings of this paper, the exact same models should be used. To assist in the reproduction, an analysis, explanation and motivation of the prompts were given in section 3.2 as well as a detailed overview of the models and parameters in section 4.1. It is important to note, however, that the output of these models is influenced by the continuous Reinforcement Learning from Human Feedback (RLFH). Meaning that as people continue using these models and labelling outputs as good or bad, the models are refined continuously and might be different when reproducing this work. To adress the latter concern, regarding the annotations, a description of the retrievable location of these annotated files is given at the end of section 4.2. Lastly, the complete codebase can be found in this work's corresponding repository repository[9].

6 Discussion

6.1 Results

Overlap classification from text itself faces some difficulties. One aspect that makes this task especially hard is the structure of the data. For instance, it is not completely clear from only the text, when exactly the overlapper starts speaking. The start and end time of each segment are given, however it would be more clear if the data included an indication of exactly where in the overlappee's sentence the overlapper starts speaking. A human could infer this by looking at the start and end times of the segments, but a language model has great difficulty extracting this information. When prompting the model with the question: "How many seconds do the sentences overlap?", the model often replies with the duration of either one of the sentences from the overlap. Furthermore, while sometimes the overlappee stops speaking during a competitive overlap, more often than not the overlappee completes their sentence, resulting in the complete sentence being transcribed before the start of the overlapping sentence. Without hearing the audio recording, this makes it extremely hard for the model to make a distinction between a competitive overlap and an event where the overlapper starts speaking near or at the end of the overlappee's sentence. These instances are the most difficult for LLM's to classify, and will always be a challenge for text-based overlap classification on data with a similar structure to the corpus used in this work.

It should be noted that, especially for few-shot-learning, prompt design is of utmost importance. Before choosing the final design, as shown in appendix A.4, many other designs, variants of support sets, and instructions were tested, most of which only confused the model and caused it to return the same class for all overlap occurrences. By varying the types of competitive and non-competitive overlaps as well as their order and choosing an appropriate support set size, the few-shot-learning approach can consistently generate very accurate results.

6.2 Comparison with previous work

Due to the novelty of this research area, there is no work, at the time of writing, that used large language models for overlap classification. To still place the results in a broader context, they are compared with work that uses different classification techniques. Egoro & Wendemuth [26] include a neat overview of the results of the related works. For each paper, they specify the employed features, corpus, method, performance metric and score. We focus our comparison on the works that use the F1 score as metric, similar to this work, taking the papers from [26] into account as well as [26] itself and another paper that was not included. A complete overview is given in table 2.

While different datasets were used in the compared works making a direct comparison difficult, the approach used in this work matches the performance with a macro-averaged F1 score of **0.72**. This result is surprising because it was obtained from only lexical features, showcasing the inferring powers of current LLMs. Furthermore, the little amount of engineering effort needed to instruct the model to perform the classifications is also impressive, compared to carefully engineered support vector machines and neural networks.

A week before the completion of this paper, Ding et al. [33] published a paper reporting on the annotation capabilities of GPT-3. Their results show that while the data annotated by GPT-3 is of high quality, there is room for improvement when comparing them to human annotations. The results from this paper align with those from [33], namely that they are good, but somewhat inconsistent even with a carefully engineered prompt. Ding et al. also conclude that individuals and low-budget organizations can rely on LLMs to deliver high quality annotations at a very low cost.

[9] https://gitlab.tudelft.nl/mtarvirdians/meeting-mastery/-/tree/main

6.3 Future Work

LLMs have already a massive impact on the world, affecting education, the work floor, research and so on. The sheer fact that Large Language Models are currently evolving at a rapid pace, creates the need for continuous research to track the capabilities of these models. Future work on the area of overlap classification could adopt different corpora such as the aforementioned LUNA corpus. The question: How well do LLMs perform on certain tasks on certain datasets? should be ammended to: How do we structure our data such that LLMs work best? By looking at this problem from both sides (LLMs and data structure) we can expect to extract even more performance from these incredibly intelligent models.

Another direction for future work is overlap detection with LLMs. Before overlap classification can take place, the overlaps must be extracted from the meeting. Overlap detection plays a very big role in bringing this technology to the real-life work floor, since a combination of overlap detection and classification can enable an automatic analysis of overlap dynamics from a complete meeting transcript. Due to the short duration of this project, there was insufficient time to also investigate overlap detection. However, if there had been more time allocated, overlap detection would have been the first thing to be included in this work.

Finally, it would be interesting to see how these models perform on 2-party dialogue dataset, such as the LUNA corpus. Multi-party meetings tend to contain a lot of backchannels and unnecessary overlapping speech such as think aloud. In a 2-party dialogue, the conversational dynamics of the meeting are less complicated since the overlap cannot be directed to a third party, and false starts occur less often. Because of this, the overlaps would be easier to classify for humans and LLMs alike.

6.4 Limitations

Clearly, this work has some limitations. Firstly, the time constraint. The 10-week period resulted in the need for irresponsible manual annotation, which has been clearly explained in section 5. Furthermore, at the time of writing, there is limited access to the GPT-4 API. Consequently, the temperature, max_tokens and stop_sequence parameters for the GPT-4 models could not be modified from the default values, as ChatGPT does not allow this. These limitations call for the need to reproduce and validate this work at a later moment, when access to the GPT-4 API is granted to all users, and there is a longer period of time allocated to the project to responsibly collect overlap annotations.

7 Conclusion

In this study, four different prompt designs have been used and tested in order to find out what prompt engineering techniques work best for text-based speech overlap classification. Experiments were done on four meetings, with a cumulative of 110 overlap occurrences. From the results it is concluded that this text-based LLM approach matches the performance of multi-modular neural networks from previous work. As expected, it was shown that the predictions of GPT-4 outperform those of GPT-3 in terms of F1 score and most importantly in terms of consistency. In-context learning, such as few-shot-learning, is a valuable tool which requires very little engineering knowledge and effort. This technique can be used to instruct an LLM to perform very specific tasks, like classification and data annotation. A final conclusion is that as LLMs continue to improve, they become a more and more attractive tool for automatic text annotation, which can be used in a wide range of use cases, such as businesses who want to improve their meeting quality by reducing the number of competitive overlaps.

References

[1] Simone Kauffeld. *Kompetenzen messen, bewerten, entwickeln.* Jan. 2006. ISBN: 978-3-7910-2508-7.

[2] Simone Kauffeld and Nale Lehmann-Willenbrock. "Meetings Matter Effects of Team Meetings on Team and Organizational Success". In: *Small Group Research* 43 (Apr. 2012), pp. 130–158. DOI: 10.1177/1046496411429599.

[3] Leslie Perlow, Constance Noonan Hadley, and Eunice Eun. "Stop the meeting madness". In: *Harvard Business Review* (July 2017), pp. 62–69. URL: https://hbr.org/2017/07/stop-the-meeting-madness.

[4] J. R Hackman. *Leading teams: Setting the stage for great performances*. Boston: Harvard Business School Press, 2002.

[5] Sally Farley. "Attaining Status at the Expense of Likeability: Pilfering Power Through Conversational Interruption". In: *Journal of Nonverbal Behavior* 32 (2008), pp. 241–260. DOI: 10.1007/s10919-008-0054-x.

[6] Yejin Bang et al. "A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity". In: (Feb. 2023). URL: http://arxiv.org/abs/2302.04023.

[7] Chengwei Qin et al. "Is ChatGPT a General-Purpose Natural Language Processing Task Solver?" In: (Feb. 2023). URL: http://arxiv.org/abs/2302.06476.

[8] Ashish Vaswani et al. "Attention Is All You Need". In: (June 2017). URL: http://arxiv.org/abs/1706.03762.

[9] Yiheng Liu et al. "Summary of ChatGPT/GPT-4 Research and Perspective Towards the Future of Large Language Models". In: (Apr. 2023). URL: http://arxiv.org/abs/2304.01852.

[10] Long Ouyang et al. *Training language models to follow instructions with human feedback*. 2022. arXiv: 2203.02155 [cs.CL].

[11] Paul F Christiano et al. "Deep Reinforcement Learning from Human Preferences". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf.

[12] Rik Huijzer and Yannick Hill. *Large Language Models Show Human Behavior*. 2023. URL: https://www.techopedia.com/definition/5967/artificial-.

[13] OpenAI. *GPT-4 Technical Report*. 2023. arXiv: 2303.08774 [cs.CL].

[14] Hugo Touvron et al. “LLaMA: Open and Efficient Foundation Language Models”. In: (Feb. 2023). URL: http://arxiv.org/abs/2302.13971.

[15] Google. “LaMDA: Language Models for Dialog Applications”. In: (Jan. 2022). URL: http://arxiv.org/abs/2201.08239.

[16] BigScience Workshop. “BLOOM: A 176B-Parameter Open-Access Multilingual Language Model”. In: (Nov. 2022). URL: http://arxiv.org/abs/2211.05100.

[17] Caleb Ziems et al. *Can Large Language Models Transform Computational Social Science?* 2023. arXiv: 2305.03514 [cs.CL].

[18] Jean Carletta et al. “The AMI Meeting Corpus: A pre-announcement”. In: vol. 3869 LNCS. 2006, pp. 28–39. ISBN: 3540325492. DOI: 10.1007/11677482_3.

[19] Edgar Gutierrez et al. “Analysis of human behavior by mining textual data: Current research topics and analytical techniques”. In: *Symmetry* 13 (7 July 2021). ISSN: 20738994. DOI: 10.3390/sym13071276.

[20] Shammur Absar Chowdhury, Evgeny A. Stepanov, and Giuseppe Riccardi. “Predicting user satisfaction from turn-taking in spoken conversations”. In: vol. 08-12-September-2016. International Speech and Communication Association, 2016, pp. 2910–2914. DOI: 10.21437/Interspeech.2016-859.

[21] Shammur Absar Chowdhury et al. “Automatic classification of speech overlaps: Feature representation and algorithms”. In: *Computer Speech and Language* 55 (May 2019), pp. 145–167. ISSN: 10958363. DOI: 10.1016/j.csl.2018.12.001.

[22] Sepp Hochreiter and Jürgen Schmidhuber. “Long Short-term Memory”. In: *Neural computation* 9 (Dec. 1997), pp. 1735–80. DOI: 10.1162/neco.1997.9.8.1735.

[23] Shammur Chowdhury, Morena Danieli, and Giuseppe Riccardi. “The Role of Speakers and Context in Classifying Competition in Overlapping Speech”. In: Sept. 2015. DOI: 10.21437/Interspeech.2015-68.

[24] Shammur Absar Chowdhury and Giuseppe Riccardi. “A Deep Learning approach to modeling competitiveness in spoken conversations”. In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017, pp. 5680–5684. DOI: 10.1109/ICASSP.2017.7953244.

[25] Olga Egorow and Andreas Wendemuth. “Emotional Features for Speech Overlaps Classification”. In: Aug. 2017, pp. 2356–2360. DOI: 10.21437/Interspeech.2017-87.

[26] Olga Egorow and Andreas Wendemuth. “On Emotions as Features for Speech Overlaps Classification”. In: *IEEE Transactions on Affective Computing* 13 (1 2022), pp. 175–186. ISSN: 19493045. DOI: 10.1109/TAFFC.2019.2925795.

[27] Qingxiu Dong et al. *A Survey on In-context Learning*. 2023. arXiv: 2301.00234 [cs.CL].

[28] Indira Sultanic. “Interpreting in pediatric therapy settings during the COVID-19 pandemic: benefits and limitations of remote communication technologies and their effect on turn-taking and role boundary”. In: *FITISPos International Journal* 1 (9 June 2022). DOI: 10.37536/fitispos-ij.2023.1.9.313.

[29] Shammur Chowdhury, Morena Danieli, and Giuseppe Riccardi. “Annotating and categorizing competition in overlap speech”. In: vol. 2015. Apr. 2015. DOI: 10.1109/ICASSP.2015.7178986.

[30] Tom B. Brown et al. “Language Models are Few-Shot Learners”. In: *CoRR* abs/2005.14165 (2020). arXiv: 2005.14165. URL: https://arxiv.org/abs/2005.14165.

[31] Tina Tseng, Amanda Stent, and Domenic Maida. *Best Practices for Managing Data Annotation Projects*. Sept. 2020.

[32] Rion Snow et al. “Cheap and Fast – But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks”. In: *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Honolulu, Hawaii: Association for Computational Linguistics, Oct. 2008, pp. 254–263. URL: https://aclanthology.org/D08-1027.

[33] Bosheng Ding et al. *Is GPT-3 a Good Data Annotator?* 2023. arXiv: 2212.10450 [cs.CL].

A Prompts

A.1 simple_prompt

```
Following are the 2 different categories for
    ↪ overlapping speech:

0: Non-competitive overlap (where the speaker
    ↪ encourages the current speaker to
    ↪ continue)
1: Competitive overlap (where the speaker
    ↪ tries to take over the conversation)

The following sentences are a part of a
    ↪ meeting transcript. The middle sentence
    ↪  that starts with "Overlap:" is the
    ↪ sentence that starts before the
    ↪ previous speaker has finished speaking.
    ↪  Classify the overlap as non-
    ↪ competitive or competitive. Only return
    ↪  either the number corresponding
    ↪ overlap category, nothing else.

Preceding sentence: <CONTEXT>
Overlap: <CONTEXT>
Category:
```

A.2 explain_prompt

```
Following are the explanations for the
    ↪ different categories for overlapping
    ↪ speech:

0: Non-competitive overlap. These are
    ↪ occurences where 1) another speaker
    ↪ starts in the middle of an ongoing turn
    ↪ , 2) both parties do not show any
    ↪ evidence for grabbing the turn for
    ↪ themselves, 3) speakers perceive the
    ↪ overlap as non-problematic and 4)
    ↪ speakers use it to signal the support
    ↪ for the current speaker's continuation
    ↪ of speech.
1: Competitive overlap. These are occurences
    ↪ where 1) the intervening speaker starts
    ↪  prior to the completion of the current
    ↪  speaker, 2) both the speakers display
    ↪ interest in the turn for themselves,
    ↪ and 3) speakers perceive the overlap as
    ↪  problematic.

The following sentences are a part of a
    ↪ meeting transcript. The middle sentence
    ↪  that starts with "Overlap:" is the
    ↪ sentence that starts before the
    ↪ previous speaker has finished speaking.
    ↪  Classify the overlap as non-
    ↪ competitive or competitive. Only return
    ↪  either the number corresponding
    ↪ overlap category, nothing else.

Preceding sentence: <CONTEXT>
Overlap: <CONTEXT>
Category:
```

A.3 one_shot_prompt

```
You are an overlap classification tool.

An overlap occurs when the intervening speaker
    ↪  starts speaking prior to the
    ↪ completion of the current speaker.
    ↪ Based on both sentences, classify the
    ↪ overlap as non-competitive (category 0)
    ↪  or competitive (category 1):

Preceding sentence: um might have been a good
    ↪ idea to all deliver our presentations
    ↪ and then discuss
Overlap: Yeah , that's a good idea
Category: 0

Preceding sentence: Um , so so far , just to
    ↪ recap you've got volume and channel
    ↪ control and
Overlap: There's um on and off , um volume and
    ↪  channel , and skip to certain channels
    ↪  with the numbers .
Category: 1

Preceding sentence: <CONTEXT>
Overlap: <CONTEXT>
Category:
```

A.4 few_shot_prompt

```
You are an overlap classification tool.

An overlap occurs when the intervening speaker
    ↪  starts speaking prior to the
    ↪ completion of the current speaker.
    ↪ Based on both sentences, classify the
    ↪ overlap as non-competitive (category 0)
    ↪  or competitive (category 1):

Preceding sentence: um might have been a good
    ↪ idea to all deliver our presentations
    ↪ and then discuss
Overlap: Yeah , that's a good idea
Category: 0

Preceding sentence: um the channels like the
    ↪ the numbers on thing , um <disfmarker>
Overlap: Up <disfmarker> the numbers , or the
    ↪ up down ?
Category: 0

Preceding sentence: Like if you want it on <
    ↪ disfmarker>
Overlap: Where you can activate it and
    ↪ deactivate it ?
Category: 1

Preceding sentence: uh any kind of like
    ↪ display controls at all do you think we
    ↪  need to worry about ,
Overlap: We don't ? No ?
Category: 0

[.................]

Preceding sentence: Try and sell it t sell it
    ↪ to them to supply with um <disfmarker>
Overlap: There is that possibility , yes .
Category: 0

Preceding sentence: Um , so so far , just to
    ↪ recap you've got volume and channel
    ↪ control and <disfmarker>
Overlap: There's um on and off , um volume and
    ↪  channel , and skip to certain channels
    ↪  with the numbers .
Category: 1

Preceding sentence: we could maybe incorporate
Overlap: those little key-rings have both ,
Category: 1

Preceding sentence: <CONTEXT>
Overlap: <CONTEXT>
Category:
```