Can GPT models Follow Human Summarization Guidelines? A Study for Targeted Communication Goals

Yongxin Zhou **Fabien Ringeval** **François Portet**
Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG, 38000, Grenoble, France
firstname.lastname@univ-grenoble-alpes.fr

Abstract

This study investigates the ability of GPT models (ChatGPT, GPT-4 and GPT-4o) to generate dialogue summaries that adhere to human guidelines. Our evaluation involved experimenting with various prompts to guide the models in complying with guidelines on two datasets: DialogSum (English social conversations) and DECODA (French call center interactions). Human evaluation, based on summarization guidelines, served as the primary assessment method, complemented by extensive quantitative and qualitative analyses. Our findings reveal a preference for GPT-generated summaries over those from task-specific pretrained models and reference summaries, highlighting GPT models' ability to follow human guidelines despite occasionally producing longer outputs and exhibiting divergent lexical and structural alignment with references. The discrepancy between ROUGE, BERTScore, and human evaluation underscores the need for more reliable automatic evaluation metrics. [1]

1 Introduction

Although instruction-tuned Large Language Models (LLMs) have shown impressive performance on several benchmark datasets, they still struggle with various challenging tasks (Laskar et al., 2023; Qin et al., 2023). On automatic metrics like ROUGE, general-purpose models (e.g., ChatGPT) consistently trail behind task-specific, fine-tuned models (Bang et al., 2023; Zhang et al., 2023; Yang et al., 2023). In dialogue summarization specifically, ChatGPT demonstrates poorer performance than state-of-the-art fine-tuned models on datasets including SAMSum (Gliwa et al., 2019) and DialogSum (Chen et al., 2021). However, human evaluation studies reveal a more nuanced picture: instruction-tuned LLMs can demonstrate human-aligned summarization capabilities that automatic metrics may fail to capture (Zhang et al., 2024).

The underlying causes of GPT models' underperformance on automatic summarization metrics remain poorly understood. Human summarization is a complex cognitive process that involves understanding, condensing, and conveying essential information while adhering to specific guidelines. Given that human annotators have to follow these guidelines when creating reference summaries, we speculate that the performance discrepancy could be attributed to the lack of explicit summarization instructions tailored to specific communication goals during model prompting. This is particularly critical for dialogue summarization, where targeted communication goals play a significant role.

Even though dialogue summarization is a well-established task, it remains challenging due to the absence of a consensus for what constitutes an ideal dialogue summary (Guo et al., 2022). The approach to dialogue summarization is heavily influenced by specific communication objectives, which vary across different contexts such as meetings, or customer service interactions. A review of guidelines from major dialogue summarization datasets reveals that each corpus defines its own set of objectives for creating reference summaries, applying different summary criteria to meet specific needs (Zhou et al., 2024). For instance, in the context of customer service, while TWEETSUMM (Feigenblat et al., 2021) offers both extractive and abstractive summaries, CSDS (Lin et al., 2021) provides three distinct summaries for each dialogue: an overall summary and two role-oriented summaries (user and agent).

In this work, we evaluated the ability of GPT models to generate dialogue summaries that adhere to human guidelines, which provides the following contributions:

- Developing a range of prompts, from simpli-

[1] The generated summaries and human annotations are publicly available at https://github.com/yongxin2020/LLM-Sum-Guidelines

Proceedings of the 18th International Natural Language Generation Conference, pages 249–273
October 29–November 2, 2025.

fied instructions to detailed human-annotator guidelines, for summarization tasks;
- Evaluating three GPT models against task-specific pre-trained models on English and French datasets with distinct communication goals (DialogSum and DECODA);
- Assessing performance using automatic metrics (ROUGE, BERTScore, and LLM-as-judge) and a task-aligned human evaluation framework;
- Performing comprehensive analyses to assess and explain model performance, identifying both capabilities and limitations in meeting targeted communication objectives.

2 Related Work

2.1 Evaluation of Summarization Tasks

The dominant approach for automatic summarization evaluation employs similarity-based metrics. ROUGE (Lin, 2004) remains the standard, measuring n-gram overlap (e.g., ROUGE-1, ROUGE-2, ROUGE-L) between system outputs and human references. To capture semantic similarity beyond lexical overlap, embedding-based metrics like BERTScore (Zhang* et al., 2020) leverage contextual embeddings. Alternative paradigms include natural language inference (NLI) for evaluating factual consistency through entailment (Chen and Eger, 2023) and question answering (QA)-based methods that use question generation and answering (Wang et al., 2020). More recently, LLM-based evaluation has emerged, encompassing metrics derived from, prompted by, or fine-tuned on LLMs (Gao et al., 2025). For instance, G-Eval (Liu et al., 2023) operates without references and assesses dimensions like coherence, consistency, fluency, and relevance.

Beyond automatic metrics, human evaluation of summaries has evolved into a multi-dimensional framework assessing faithfulness, coherence, relevance, completeness, and conciseness (Fabbri et al., 2021). This framework is applied across diverse domains: from general news benchmarks evaluating faithfulness, coherence, and relevance (Zhang et al., 2024), to specialized tasks like topic-focused dialogue summarization assessed for completeness, relevance, and factual consistency (Tang et al., 2024), to domain-specific evaluations (e.g., bookings, interviews) that employ key fact validation to calculate scores for faithfulness, completeness, and conciseness (Min et al., 2025).

Nevertheless, a number of challenges remain. Automatic metrics show limited correlation with human judgment (Fabbri et al., 2021). LLM-based approaches face challenges such as prompt sensitivity, limited confidence calibration, the need for human oversight (restricting full automation), and limited explainability (Gao et al., 2025). Furthermore, the field lacks consensus on evaluation standards, with substantial variability in criteria and protocols across studies (Howcroft et al., 2020), complicating reliable and reproducible assessment.

2.2 Instruction Following in LLMs

Prompt engineering has emerged as an indispensable technique for guiding LLMs, enabling users to interact with generative AI systems through carefully designed instructions without modifying core model parameters (Sahoo et al., 2025; Schulhoff et al., 2025). Systematic surveys have categorized numerous distinct prompt engineering techniques based on their targeted functionalities (Sahoo et al., 2025), though challenges persist including biases, factual inaccuracies, and interpretability gaps.

Research demonstrates that prompt formatting impacts LLM performance through various approaches. For example, the *Chain-of-Thought (CoT)* technique prompts models to generate intermediate reasoning steps before delivering a final answer (Wei et al., 2022). *Optimization by PROmpting (OPRO)* employs an iterative process where the LLM generates new solutions from prompts containing previous solutions with their values, which are then evaluated and added to subsequent prompts (Yang et al., 2024). Studies also show that emotional intelligence can be leveraged through *EmotionPrompt*, which combines original prompts with emotional stimuli to enhance performance (Li et al., 2023). Furthermore, the *Rephrase and Respond* approach allows LLMs to rephrase and expand human-posed questions before providing responses, serving as an effective method for improving output quality (Deng et al., 2024). These techniques collectively demonstrate that prompt quality influences response quality, with even simple formatting adjustments proving effective for performance improvement.

3 Experimental Setup

Datasets. We used two datasets covering different communication goals, languages and contexts: DialogSum (Chen et al., 2021) and DE-

CODA (Favre et al., 2015).

DialogSum consists of social conversations in English and includes 12,460, 1,500 and 500 samples in its training, validation and test splits, respectively. It was used in a challenge (Chen et al., 2022), which favors comparison of results.

DECODA (Favre et al., 2015) is a call-center human-human spoken conversation corpus in French, collected from the RATP (Paris public transport authority) and mostly deals with customer inquiries and agent responses. The corpus was proposed as a pilot task at Multiling 2015 (Favre et al., 2015). The test set used in this study contains 100 conversations with 212 synopses, i.e. references.

Models and Parameters. We used OpenAI `ChatGPT (gpt-3.5-turbo)`,[2] `GPT-4`,[3] and `GPT-4o`[4] APIs for our experiments.

`gpt-3.5-turbo` is optimized for conversational tasks while maintaining strong performance on standard completion tasks. The model handles inputs up to 4096 tokens, compared to `GPT-4`'s 8192-token limit and `GPT-4o`'s 128,000-token context windows. To ensure a stable output, we configured the temperature parameter at 0 while maintaining the default values for all other parameters.

For comparison with task-specific pre-trained models, we also fine-tuned BART-large[5] on the DialogSum dataset and BARThez[6] on the DECODA dataset. The details of the training are presented in Appendix A.

Summarization Prompt Design. The different prompts used in our experiments are given in Table 1. The `WordLimit (WL)` prompt simply constrains the word length of the output to be less than X words (Laskar et al., 2023). Another prompt we tested was to provide a text only version of the full `Human Guideline (HG)`. For DialogSum, the annotation guideline was directly described in Chen et al. (2021). For DECODA, we contacted the authors (Favre et al., 2015) and obtained the guidelines used to write the synopses. In addition, to examine if giving the system a role helps, we proposed `Human Guideline with Role (HGR)`, which begins with *"You are an annotator ..."*. We have also investigated a two-step iterative approach, called `HG(R)`→`WL`, which first uses the summaries generated by `HG` or `HGR` as input, then re-injects them into the model to reduce the length of the output using the `WL` prompt.

Evaluation Metrics. We reported the F1 scores of ROUGE-1/2/L (Lin, 2004), which assess the similarity between candidates and references based on the overlap of unigrams, bigrams, and the longest common sequence. We used a publicly available implementation.[7]

We also reported the F1 scores of BERTScore (Zhang* et al., 2020), which measures the similarity between candidates and references at token level, using BERT's contextual embeddings.[8] For DialogSum, following Chen et al. (2022), we used `RoBERTa-large` (Liu et al., 2019) as the backbone. For DECODA in French, we used the default multilingual model: `bert-base-multilingual-cased` (Devlin et al., 2019).

However, few metrics excel across all dimensions, leading recent studies to explore LLMs for evaluating text quality (Liu et al., 2023). Following this trend, we adopted an LLMs-as-judge approach, using `DeepSeek-R1` as the backbone for automatic evaluation on a subset of model predictions.

4 Results

4.1 Quantitative Results

DialogSum Table 2 presents the automatic evaluation results on the test set. While the `ChatGPT` model with the `WordLimit` prompt underperforms state-of-the-art pre-trained models by approximately 15 (R1), 10 (R2), and 20 (RL) points, BERTScore indicates only a minor semantic discrepancy. The `GPT-4` model with the same prompt shows a slight overall improvement over `ChatGPT`, except on R2. Furthermore, `GPT-4o` performs comparably to `GPT-4`, with the `WL` prompt yielding superior results on most metrics, though not on BERTScore.

Although the DialogSum guideline states *"be brief (no longer than 20% of the conversation length)"*, the results of the `HG` prompt indicate that instructions tailored for human annotators may not be suitable when directly used as instructions to generate reference-like summaries with `ChatGPT`,

[2] OpenAI GPT-3.5: https://platform.openai.com/docs/models/gpt-3-5, version gpt-3.5-turbo-0613

[3] OpenAI GPT-4: https://platform.openai.com/docs/models/gpt-4, version gpt-4-0613

[4] gpt-4o-2024-08-06: https://platform.openai.com/docs/models#gpt-4o

[5] https://huggingface.co/facebook/bart-large

[6] https://huggingface.co/moussaKam/barthez

[7] https://github.com/google-research/google-research/tree/master/rouge

[8] https://huggingface.co/spaces/evaluate-metric/bertscore

Settings	Prompts
WordLimit (WL) (Laskar et al., 2023)	System: *Write a summary with not more than X words. User: [Test Dialogue]*
Human Guideline (HG) for DialogSum	*System:* ***Write a summary*** *based on the following criteria: the summary should (1) convey the most salient information of the dialogue and; (2) be brief (no longer than 20% of the conversation length) and; (3) preserve important named entities within the conversation and; (4) be written from an observer perspective and; (5) be written in formal language. In addition, you should pay extra attention to the following aspects: 1) Tense Consistency: take the moment that the conversation occurs as the present time, and choose a proper tense to describe events before and after the ongoing conversation. 2) Discourse Relation: If summarized events hold important discourse relations, particularly causal relation, you should preserve the relations if they are also in the summary. 3) Emotion: you should explicitly describe important emotions related to events in the summary. 4) Intent Identification: Rather than merely summarizing the consequences of dialogues, you should also describe speakers intents in summaries, if they can be clearly identified. In addition to the above, you should use person tags to refer to different speakers if real names cannot be detected from the conversation. User: [Test Dialogue]*
Human Guideline (HG) for DECODA	*System:* ***Write*** *a conversation-oriented summary in the form of a synopsis expressing both the customer's and the agent's points of view, and which should ideally report: 1. The main issues of the conversation: in call center conversations the main issues are the problems why the customer called; their identification constitutes the basis for classifying the call into several different classes of motivations for calling . 2. The sub-issues in the conversation: when in the conversation any sub-issue occurs, it may be there both because it is introduced by the customer or by the agents. 3. The resolution of the call: i.e. if the customers problem was solved in that call (first-call resolution) or not. User: [Test Dialogue]*
Human Guideline with Role (HGR)	The same as the **Human Guideline (HG)** for DECODA but we changed the begin of instruction to ***You are an annotator and are asked to write (dialogue summaries)*** ....

Table 1: Prompts used for experiments. X is the average length of the reference summaries/synopses (DialogSum: X=20, DECODA: X=25). Prompts for DECODA have been translated from French.

DialogSum	R1	R2	RL	BS	Avg. Len
Human Ref.	53.35	26.72	50.84	92.63	21.07
GoodBai	**47.61**	**21.66**	45.48	**92.72**	25.72
UoT	47.29	21.65	**45.92**	92.26	27.05
BART-large	47.36	21.23	44.88	91.42	27.20
3.5-WL	32.92	12.45	26.66	88.78	28.69
3.5-HG	24.15	8.60	18.70	87.68	104.89
3.5-HGR	26.43	9.27	20.50	88.16	88.85
3.5-HG→WL	35.10	11.52	27.39	89.72	40.02
3.5-HGR→WL	35.76	11.71	28.19	89.87	36.18
4-WL	34.94	12.38	28.99	89.15	18.63
4-HGR	26.25	8.01	19.86	88.44	86.78
4-HGR→WL	35.42	9.58	28.20	90.25	19.77
4o-WL	35.34	11.09	28.83	89.33	18.62
4o-HGR	25.53	7.92	19.34	88.21	89.67
4o-HGR→WL	33.80	9.18	26.96	89.83	20.23

Table 2: Comparison of ROUGE-1/2/L, BERTScore, and average length for human references and summaries generated by different systems, including the top-performing models from the DialogSum challenge: GoodBai (Chen et al., 2022) and UoT (Lundberg et al., 2022). Our results from ChatGPT (indicated as 3.5-), GPT-4 (indicated as 4-) and GPT-4o (indicated as 4o-) are presented alongside. The best results are in bold, and the top performances for ChatGPT, GPT-4 and GPT-4o are underlined. Results from the two-step approach are highlighted in gray. Abbreviations: WL (WordLimit), HG (Human Guideline), HGR (Human Guideline with Role), BS (BERTScore).

DECODA	R1	R2	RL	BS	Avg. Len
Reference	-	-	-	-	25.40
BARThez	**35.42**	**16.96**	**29.41**	**74.94**	18.53
3.5-WL	33.21	12.53	24.74	72.54	37.90
3.5-HG	18.42	7.18	13.68	67.72	162.83
3.5-HGR	16.72	6.73	12.66	66.93	190.87
3.5-HG→WL	31.61	11.00	23.34	71.77	46.30
3.5-HGR→WL	32.74	12.22	24.09	72.30	41.76
4-WL	35.23	13.54	27.73	73.23	23.81
4-HGR	20.71	8.29	15.12	69.04	136.53
4-HGR→WL	32.12	11.17	24.82	72.30	26.01
4o-WL	34.86	13.05	27.04	73.59	22.73
4o-HGR	19.35	7.73	13.94	68.57	160.73
4o-HGR→WL	31.43	10.40	23.67	72.42	24.17

Table 3: Comparison of ROUGE-1/2/L, BERTScore, and average length for references and the state-of-the-art fine-tuned model (Zhou et al., 2022) against ChatGPT (indicated as 3.5-), GPT-4 (indicated as 4-) and GPT-4o (indicated as 4o-) on the DECODA dataset. The best overall results are in bold, and the top performances for ChatGPT, GPT-4 and GPT-4o are underlined. Results from the two-step approach are highlighted in gray. Abbreviation: WL (WordLimit), HG (Human Guideline), HGR (Human Guideline with Role), BS (BERTScore).

as responses tend to be longer, leading to lower ROUGE and BERTScore values. This effect could have been reduced by specifying the role of the annotator in the prompt (HGR), which consistently improved performance with reference to HG across all measures.

Regarding the two-step prompt approach HGR→WL, which incorporates human annotation instructions as a first step, followed by applying a second prompt to limit word length, the performance aligned with those obtained with a single WordLimit prompt, with even higher R1 and RL scores, suggesting that the GPT model's word usage is more akin to the references, maintaining the primary content and logical order, albeit with some variability in language usage.

In terms of average length, GPT-4 and GPT-4o better followed the WordLimit prompt in obtaining a summary whose length is closest to the reference compared to ChatGPT.

DECODA Results in Table 3 show that the WordLimit prompt produces the best outcomes for both GPT models. However, their performance

still falls short of the pre-trained model across all metrics. The difference is less pronounced in the BERTScore than in the ROUGE scores.

When using human guidelines (HG(R)), results diverged from those of the other prompts, resulting in reduced ROUGE and BERTScore. For ChatGPT, HGR showed lower results than HG. The two-step prompting HG(R)→WL produced shorter final summaries than HG(R), resulting in higher scores across various measures, although they remain slightly lower than those obtained with the simple WordLimit prompt.

When comparing the GPT models, all three models performed better with the WordLimit prompt, while yielding similar results with the iterative prompt (HGR→WL). In terms of average length, the WordLimit prompt consistently gave the shortest summaries for all GPT models. GPT-4 and GPT-4o followed better the WordLimit prompt in obtaining a summary of less than 25 words than ChatGPT, which consistently exceeded the limit.

Using LLM-as-judge Using DeepSeek-R1[9] as our evaluation backbone, we evaluated the generated summaries against the human summarization guideline criteria (see Appendix B for details and prompts). We evaluated a subset of 20 samples from the DECODA dataset, with results shown in Table 4. To investigate the impact of length on quality, we include an additional column reporting summary length in our analysis.

However, we observed that even when the prompt explicitly provided guidelines for judging summaries and specified that only a score should be returned, many evaluation results contained not only scores but also explanatory justifications (sometimes in French and sometimes in English).

The GPT models' generated summaries achieved strong scores across all four criteria. The best results for *Faithfulness*, *Main Issues*, and *Resolution* came from 4-HGR→WL, 3.5-WL and 3.5-HGR→WL respectively, while the Reference performed best on *Sub-Issues*. Considering the average scores across all criteria, 4-HGR→WL achieved the highest values, while BARThez achieved the lowest.

4.2 Quantitative Results by Variation

For the DialogSum dataset, each dialogue is annotated by three annotators, resulting in three reference summaries per dialogue. Despite the shared annotation guideline, human annotators can exhibit variability in writing style. To evaluate whether GPT models follow human summarization guidelines, we compare the models' scores to the variance observed in the reference summaries for each metric (ROUGE-1/2/L and BERTScore). This approach determines whether the model-generated summaries fall within a reasonable range of human-annotated references, based on the mean and standard deviation of the reference scores.

We evaluated predictions from 4-WL, 4-HGR, and 4-HGR→WL, with results presented in Table 5. The first row – Mean (std) – represents the variance among reference summaries from three annotators. Our findings show that 4-WL predictions fall within the reference variance for R2 and RL, 4-HGR only for R2, and 4-HGR→WL for R2, RL, and BERTScore. These results demonstrate that, similar to how different annotators exhibit varying styles in writing references, the generated summaries align with the variance distribution of human annotators for certain metrics.

4.3 Manual Error Analysis

We examined the data points where GPT-generated summaries exhibit the greatest discrepancy, characterized by low ROUGE scores but high BERTScore values. For comparison, we selected predictions from two models: 4-WL (simple *WordLimit* prompt) and 4-HGR→WL, which incorporates human summarization guidelines as an intermediate step. For DialogSum, we show top discrepancies from 4-WL (Table 6) and 4-HGR→WL (Table 15). For DECODA, we present 4-WL results (Table 7) and 4-HGR→WL (Table 16), both translated from French.

4-WL We first analyze the summaries generated by 4-WL. In DialogSum examples with the greatest discrepancies, the references preserve specific mentions such as *#Person1#* and *#Person2#*, retaining named entities while summarizing the conversation. In contrast, GPT-4 predictions often provide a higher-level generalization, using generic terms like "two friends" or "employees" instead of specific tags. This indicates that GPT-generated summaries do not meet the following guideline: "**3) Preserve important named entities within the conversation**" when using a simplified prompt.

In addition, GPT-4 predictions do not match the references concerning the guideline emphasizing "**Intent Identification**".[10] For instance, in the sec-

[9] https://api-docs.deepseek.com/guides/reasoning_model

[10] "Rather than merely summarizing the consequences of

	Faith. (VN)	Main Iss. (VN)	Sub-Iss. (VN)	Resol. (VN)	Avg. (VN)	Length
Reference	3.65±1.09 (20)	4.35±1.09 (20)	**3.85±1.27** (20)	3.15±1.81 (20)	3.75±1.32 (20)	24.85±17.19
3.5-WL	4.00±0.86 (20)	**4.60±0.75** (20)	3.60±1.10 (20)	3.35±1.73 (20)	3.89±1.11 (20)	44.00±23.86
3.5-HGR→WL	3.20±1.47 (20)	4.00±1.30 (20)	3.75±1.21 (20)	**3.95±1.15** (20)	3.73±1.28 (20)	38.40±10.18
4-WL	4.00±1.12 (20)	4.40±0.82 (20)	3.42±1.39 (19)†	2.45±1.70 (20)	3.57±1.26 (20)	23.35± 3.08
4-HGR→WL	**4.15±0.93** (20)	4.45±0.76 (20)	3.80±0.83 (20)	3.50±1.54 (20)	**3.98±1.02** (20)	25.20± 3.30
BARThez	1.70±1.13 (20)	2.00±1.65 (20)	1.42±0.84 (19)‡	1.60±1.27 (20)	1.68±1.22 (20)	18.05± 5.04

†One sample in 4-WL was excluded due to "N/A".

‡One sample in BARThez was excluded due to an error in returning the score.

Table 4: Results of using `DeepSeek-R1` for LLM-as-judge evaluation. Results are presented as Mean ± Std. (VN) indicates the "validated number" of samples (out of 20) used to calculate the score for that criterion. *Length* indicates summary word count.

	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
Human Variance	53.37 (16.75)	26.73 (20.57)	44.99 (18.11)	92.86 (2.68)
4-WL	34.96	**12.35**	**28.99**	89.15
4-HGR	26.24	**8.01**	19.86	88.44
4-HGR→WL	35.43	**9.58**	**28.18**	**90.25**

Table 5: Automatic metric results on the DialogSum test set. The *Human Variance* row reports the mean (std) of *pairwise comparisons* between all three human reference summaries. Model scores in **bold** fall within one standard deviation of this human variance mean.

ID	References	Prediction	RL	BS	Diff.
16	["#Person2# at first thinks #Person1#'s behaviour cruel but finally joins #Person1#.", "#Person1# plans on playing a trick to others. #Person2# thinks it's cruel at first but then joins.", "#Person1# is about to make a prank. #Person2# thinks it's cruel at first but then joins."]	Two friends engage in a potentially offensive prank, causing a disagreement about their choice of fun.	6.00	86.09	80.09
24	["#Person1# is interviewing #Person2# about #Person2#'s ability and previous experience.", "#Person1# asks #Person2# about #Person2#'s capacities and past working experience during an interview.", "#Person1# asks #Person2# about #Person2#'s qualification for the job."]	Person 2, an Automobile Designing graduate, has experience in understanding mechanical strength and corrosion resistance.	4.86	83.67	78.81
255	["#Person1# and #Person2# are conjecturing who will get sacked based on other employees' performance and find that it might be themselves. #Person1# is so anxious as #Person1# needs to raise family but #Person2# will be happy to leave.", "#Person1# and #Person2# talk about the company's upcoming downsizing and discuss who will get sacked. #Person1# is worried but #Person2# doesn't care.", "#Person1# and #Person2# have rounds of conjectures of whom shall be fired due to the company's cutback and comes to the conclusion that they both have the biggest possibility. It drives #Person1# crazy but doesn't worry #Person2#."]	Employees speculate about potential layoffs at their company, considering various colleagues' relationships with the boss.	7.15	85.90	78.76
48	['#Person1# and #Person2# are talking about the low temperature at night, although spring has come.', '#Person1# and #Person2# agree that it still felt very cold in spring.', '#Person1# and #Person2# talk about the weather and how to keep warm.']	Despite spring's arrival, cold nights persist, causing discomfort indoors and outdoors.	8.02	86.53	78.50
67	['#Person1# is driving #Person2# to an inn. They talk about their careers, ages, and where they was born.', '#Person1# drives #Person2# to an inn and they have a talk. #Person2# is 26 and had a business trip to China. #Person1# is 40 years old American.', '#Person1# drives #Person2# from the airport to an inn and they have a casual talk about themselves.']	A Mexican businesswoman, returning from China, converses with her American-Colombian taxi driver.	7.22	84.96	77.74

Table 6: Data points where GPT-generated summaries (`4-WL`) show the greatest discrepancy in the DialogSum corpus, with low ROUGE-L but high BERTScore values. RL and BS are the average scores of three references.

ond example, the prediction references only *Person 2*, omitting the intent of *Person 1*, who is interviewing the other individual. In the fourth example, neither *Person 1* nor *Person 2* is mentioned, and their intents are absent, as the prediction reduces the dialogue to a general remark about the weather. In the fifth example, the prediction fails to follow the "**use person tags**" rule, summarizing the conversation by identifying the interlocutors as *"A Mexican businesswoman"* and *"her American-Colombian taxi driver"*. These variations probably influenced the quantitative results, as evidenced by the disparity between high BERTScore values, which indicate strong semantic similarity, and low ROUGE-L scores, which reflect minimal lexical

dialogues, you should also describe speakers intents in summaries, if they can be clearly identified. In addition to the above, you should use person tags to refer to different speakers if real names cannot be detected from the conversation."

ID	Reference	Prediction	RL	BS	Diff.
10	Request for information about a large account order. Transferred to the relevant service.	A seminar organizer is seeking transport cards for their international guests from the RATP.	0	71.13	71.13
110	Request for information following the procedure after receiving a fine notice. Transferred to the relevant service.	A customer calls customer service to understand why she has to pay five euros after contesting and justifying a bus fine.	8.33	70.86	62.53
107	request for a refund procedure following a ticket purchase due to the late delivery of the ImagineR pass, with a waiting period of over 3 weeks. Refund possible, contact ImagineR by phone or email.	A customer calls to request a refund for an Orange card purchased while waiting for her Imagine+R card, which arrived late.	8.00	69.35	61.35
32	T2 circulation	A user inquires about the launch of the T+two line from Porte de Versailles to Val d'Issy.	8.70	69.85	61.15
29	Request for information on purchasing tickets for school groups. Transferred to the relevant service.	A representative of a departmental association is looking to organize transport for more than 100 classes using public transportation.	10.26	71.05	60.80

Table 7: Data points (translated version) where GPT-generated summaries (4-WL) show the greatest discrepancy in the DECODA corpus, with low ROUGE-L but high BERTScore values.

overlap. This contrast highlights the tendency of GPT-generated summaries using a simple length-constrained prompt to capture the essential meaning and key information of the conversation while diverging from the reference summaries in their specific wording and structural composition.

For the DECODA dataset, the discrepancy may stem from the distinct characteristics of the reference synopses. These summaries focus strictly on issues and resolutions, whereas GPT-generated predictions often adopt a more **individual-centered perspective**, using terms such as *"un organisateur (an organizer)"*, *"une cliente (a client)"*, and *"un représentant (a representative)"*, etc. This difference results in relatively high BERTScore values, reflecting strong semantic similarity, but low ROUGE-L scores due to minimal word overlap with the references. Additionally, reference synopses are often extremely brief and sometimes incomplete sentences, as seen in the fourth example: *"circulation du T2"* (T2 circulation). Furthermore, all samples except the fourth fail to conclude the resolution in the generated summaries.

4-HGR→WL Next, we examined the generated summaries of 4-HGR→WL. On the DialogSum dataset, we observe better retention of named entities and personal tags (e.g., *Person 1*) compared to 4-WL, indicating that the intermediate HGR step helps the model to follow specific rules from the human summarization guidelines. However, the issue of overlooking "**Intent Identification**" persists, as prediction often focus on one individual's intent while omitting the other's, whereas references typically indicate both interlocutors' intents.

For DECODA, similar to the 4-WL output, most generated summaries employ an individual-centered perspective that creates divergence in n-gram overlap and semantic similarity. Although reference summaries maintain conciseness, model predictions tend toward detail while frequently omitting critical "**Resolution**" elements. For example, the first sample omits "*Transfer to the relevant department*" despite this being an explicit requirement in the summarization guidelines.

In summary, comparing cases with the greatest discrepancy (low ROUGE-L but high BERTScore) between 4-WL and 4-HGR→WL, the latter adheres better to guidelines for named entities and personal tags. However, both models share similar shortcomings: in DialogSum, they overlook "Intent Identification"; in DECODA, they use an individual-centered perspective and omit "Resolution".

4.4 Human Evaluation

Due to the misalignment between automatic measures and human judgment of LLM outputs, as evidenced in news summarization experiments with GPT-3 (Goyal et al., 2023), we set up a human evaluation. To determine if the generated summaries meet predefined instructions, we conducted a human evaluation using criteria derived from the annotation guidelines provided to human annotators for writing reference summaries. We started with the DECODA dataset, as its task-oriented summarization guidelines are more easily adaptable into evaluation criteria. We selected the 10 shortest and 10 longest dialogues, assessing six summaries for each dialogue, including references and predictions from various models: BARThez, 3.5-WL, 3.5-HGR→WL, 4-WL, and 4-HGR→WL.

Annotation interface For our human evaluation task, we adapted the POTATO annotation tool (Pei et al., 2022). The evaluation interface, shown in Figure 1, features a dialogue at the top, accompa-

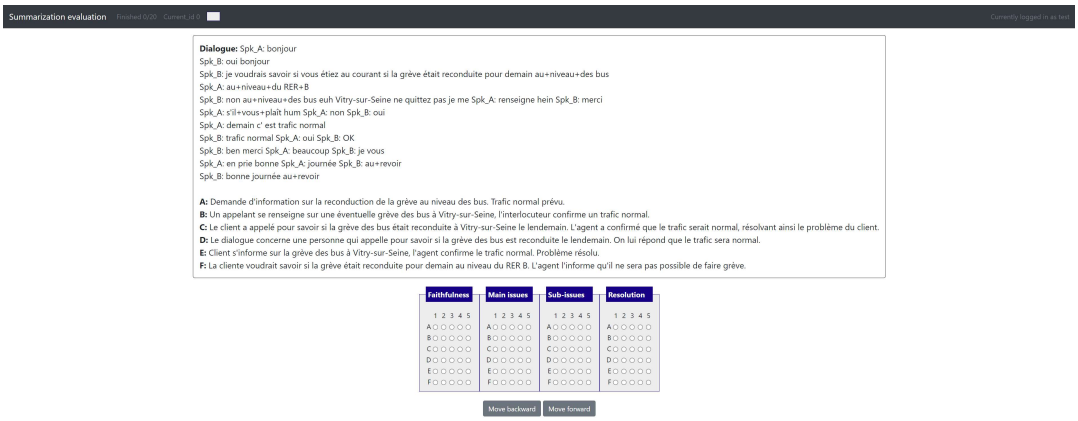

Figure 1: Screenshot of the interface used for evaluating summaries in the DECODA summarization task.

nied by six different summaries, including references and model outputs from BARThez, 3.5-WL, 3.5-HGR→WL, 4-WL, and 4-HGR→WL. To minimize bias, we randomized the order of summaries, assigning them names A, B, C, D, E, and F.

There are four blocks representing the different evaluation criteria. In each block, annotators interact with the interface by clicking on each of the six summaries, assigning a score from 1 to 5 to each summary based on its compliance with the specific criterion, 5 being the best.

Three native French speakers took part in the evaluation, all graduate students. We presented them with a PDF evaluation guide and explained that the annotations they made would be used for analysis.

Criteria Annotators assessed each summary based on four criteria: *Faithfulness*, *Main Issues*, *Sub-Issues*, and *Resolution*. The later three criteria are based on human summarization guidelines, while *Faithfulness* was added to evaluate the accuracy of the facts presented in relation to the dialogues. *Sub-Issues* was only evaluated if at least one sub-issue was present in the dialogue.

The evaluation criteria are described below. In the evaluation guide, we also provide detailed explanations for each score of criterion.

- **Faithfulness**: the summary must respect the dialogue at the level of factual information.
- **Main issues**: in call center conversations the main issues are the problems why the customer called; their identification constitutes the basis for classifying the call into several different classes of motivations for calling.
- **Sub-issues**: when in the conversation any sub-issue occurs, it may be there both because it is introduced by the customer or by the agents.
- **Resolution**: i.e. if the customer's problem was solved in that call or not.

To provide context for human evaluation scores, we include a *Length* column indicating the word count of summaries.

Results Table 8 presents the human evaluation results for the 20 evaluated dialogues across four criteria and their overall scores. A comprehensive comparison, including results for all dialogues, the 10 shortest, and the 10 longest dialogues, is provided in Table 10 in the Appendix. Overall, 3.5-HGR→WL achieved the highest score of 4.19, though 4-WL slightly excelled in *Faithfulness*. For the 10 shortest dialogues, 3.5-HGR→WL led with an overall score of 4.44, while 3.5-WL had the highest overall score at 4.08 for the 10 longest dialogues.

The *Sub-issues* criterion proved challenging to analyze for two primary reasons: sub-issues were not present in all dialogues, and evaluators demonstrated significant variability in their assessments. Specifically, one annotator identified sub-issues in five dialogues, another in three, and a third in only one dialogue. In addition, annotators generally preferred the outputs from GPT models over the reference summaries, with BARThez being the least favored overall. However, 4-WL received the lowest scores specifically for the *Resolution* criterion among the 10 longest dialogues.

The annotators' preference for GPT-generated summaries may stem from the models' use of more natural and fluid language. In contrast, the reference synopses in the DECODA dataset are highly specific and concise. For example, they are often short and to the point, such as: *"Request for train timetables from Maux station to Gare de l'Est*

	Faithfulness	Main Issues	Sub-issues	Resolution	Average	Length
Reference	3.72±1.12	4.15±1.07	3.22±1.86	3.25±1.60	3.68±1.36	24.85±17.19
3.5-WL	4.02±1.17	4.30±1.18	2.67±1.73	4.18±1.28	4.10±1.28	44.00±23.86
3.5-HGR→WL	3.97±1.15	**4.47±0.89**	**3.67±1.73**	**4.20±1.13**	**4.19±1.12**	38.40±10.18
4-WL	**4.13±0.96**	4.37±0.80	1.44±1.33	2.70±1.79	3.62±1.53	23.35± 3.08
4-HGR→WL	3.93±1.02	4.30±0.96	2.67±1.66	4.05±1.27	4.03±1.16	25.20± 3.30
BARThez	2.17±1.43	2.32±1.57	1.00±0.0	2.22±1.54	2.17±1.49	**18.05± 5.04**

Table 8: Mean (± std) scores for *Faithfulness*, *Main issues*, *Sub-issues*, *Resolution*, and *Overall* of each summary assessed by three native human evaluators on the 10 shortest and 10 longest dialogues in the DECODA dataset. The best results for each criterion are shown in bold.

station at a given time". In some cases, the reference synopses are not even complete sentences. This stylistic contrast likely explains the annotators' tendency to favor the more cohesive and natural language of the GPT-generated outputs.

	A1-A2	A1-A3	A2-A3
Faithfulness	0.422	0.273	0.315
Main issues	0.363	0.290	0.398
Sub-issues	0.140	0.445	-0.175
Resolution	0.546	0.503	0.463

Table 9: Inter-annotator agreement (IAA) scores among three annotators for dialogue summary evaluation on the DECODA dataset.

Inter-Annotator Agreement (IAA) We calculated inter-annotator agreement (IAA) scores for the three annotators using Cohen's kappa, the results are presented in Table 9. The analysis reveals stronger agreement on the *Resolution* criterion, whereas annotations for *Sub-issues* showed considerable variation.

5 Discussion and Conclusion

Human evaluation results show that GPT models (`ChatGPT`, `GPT-4` and `GPT-4o`) can follow human summarization guidelines to some extent. Their summaries were preferred over task-specific pretrained models and even outperformed reference summaries. The longer outputs of GPT models, compared to BARThez, likely contribute to more comprehensive coverage of issues and resolutions, enhancing faithfulness to dialogues. However, GPT models underperform task-specific pretrained models on both ROUGE and BERTScores metrics. This discrepancy between automatic metrics and human judgments aligns with previous findings, reinforcing (1) the continued necessity of human evaluation for text generation tasks, and (2) the critical need for developing better-aligned automatic evaluation metrics.

The results also indicate that GPT models demonstrate some proficiency in adhering to human guidelines. In the overall evaluation using human guidelines, the `HGR→WL` approach tends to produces better results than the simpler `WordLimit` prompt. Interestingly, `GPT-4` does not consistently outperform `ChatGPT`; for instance, `4-WL` only surpasses `3.5-WL` on the 10 shortest dialogues. Nevertheless, `GPT-4` exhibits better adherence to the specified word length constraints.

Our findings also highlight the inherent subjectivity of summary quality assessment, as evidenced by the human evaluators' preference for GPT-generated summaries over reference texts, which is partly attributable to stylistic differences.

By analyzing data points where GPT-generated summaries (`4-WL` and `4-HGR→WL`) exhibit significant discrepancies – characterized by low ROUGE scores but high BERTScore values – we found that the models effectively summarize key dialogue information and are semantically similar to the references. However, they sometimes neglect specific rules: in DialogSum, they miss named entities, intent identification, and personal tags; in DECODA, some summaries focus on individual locutors' perspectives and fail to present resolutions. The intermediate step of using `HGR` helped the model better adhere to the specific rules of human summarization guidelines, such as the use of named entities and personal tags in DialogSum.

For future work, we plan to develop LLM-based metrics that evaluate summaries against task-specific guidelines, with a focus on improving response stability and alignment with human judgment. Furthermore, we will investigate how LLMs follow different summarization guidelines, including those with lexical perturbations, to interpret their decision-making processes and operational mechanisms.

Limitations

In terms of model selection, our analysis was limited to GPT models: `ChatGPT`, `GPT-4` and `GPT-4o`. Future studies could benefit from a broader exploration of LLMs, encompassing models with varying parameter sizes and accessibility.

Supplementary Materials Availability Statement: Code Resources: The full experimental codebase is available at `https://github.com/yongxin2020/LLM-Sum-Guidelines`, including the LLM-as-judge framework in the `Guideline-Eval/` directory.

Datasets: The DialogSum dataset is publicly available at `https://github.com/cylnlp/dialogsum`, while the DECODA dataset can be obtained upon request from `https://pageperso.lis-lab.fr/benoit.favre/cccs/`.

Generated Outputs: All model predictions are stored in `decoda/output/` and `dialogsum/output/` directories, with two-step prompting results in `*/twoSteps/` subdirectories. BART baseline outputs are provided in `sota_barthez_predictions.txt` and `bart_large_summaries.txt`.

Human Evaluation: The annotation protocol is documented in `HumanEval_Guideline_DECODA.pdf`, with raw annotations available in `results/human_annotations_decoda/` and analysis scripts in `human_eval_scores.py`.

Acknowledgments

This research was supported by the Banque Publique dInvestissement (BPI) under grant agreement THERADIA and was partially supported by MIAI@Grenoble-Alpes (ANR-19-P3IA-0003 and ANR-23-IACL-0006). We thank the anonymous reviewers for their insightful comments, as well as the annotators who participated in the human evaluation process.

References

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali. Association for Computational Linguistics.

Yanran Chen and Steffen Eger. 2023. Menli: Robust evaluation metrics from natural language inference. *Transactions of the Association for Computational Linguistics*, 11:804–825.

Yulong Chen, Naihao Deng, Yang Liu, and Yue Zhang. 2022. DialogSum challenge: Results of the dialogue summarization shared task. In *Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges*, pages 94–103, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. DialogSum: A real-life scenario dialogue summarization dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5062–5074, Online. Association for Computational Linguistics.

Yihe Deng, Weitong Zhang, Zixiang Chen, and Quanquan Gu. 2024. Rephrase and respond: Let large language models ask better questions for themselves. *Preprint*, arXiv:2311.04205.

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

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Benoit Favre, Evgeny Stepanov, Jérémy Trione, Frédéric Béchet, and Giuseppe Riccardi. 2015. Call centre conversation summarization: A pilot task at multiling 2015. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 232–236, Prague, Czech Republic. Association for Computational Linguistics.

Guy Feigenblat, Chulaka Gunasekara, Benjamin Sznajder, Sachindra Joshi, David Konopnicki, and Ranit Aharonov. 2021. TWEETSUMM - a dialog summarization dataset for customer service. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 245–260, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mingqi Gao, Xinyu Hu, Xunjian Yin, Jie Ruan, Xiao Pu, and Xiaojun Wan. 2025. Llm-based nlg evaluation: Current status and challenges. *Computational Linguistics*, 51(2):661–687.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2023. News summarization and evaluation in the era of gpt-3. *Preprint*, arXiv:2209.12356.

Yanzhu Guo, Chloé Clavel, Moussa Kamal Eddine, and Michalis Vazirgiannis. 2022. Questioning the validity of summarization datasets and improving their factual consistency. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5716–5727, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Huang. 2023. A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 431–469, Toronto, Canada. Association for Computational Linguistics.

Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. 2023. Large language models understand and can be enhanced by emotional stimuli. *Preprint*, arXiv:2307.11760.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Haitao Lin, Liqun Ma, Junnan Zhu, Lu Xiang, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2021. CSDS: A fine-grained Chinese dataset for customer service dialogue summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4436–4451, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Conrad Lundberg, Leyre Sánchez Viñuela, and Siena Biales. 2022. Dialogue summarization using BART. In *Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges*, pages 121–125, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

Hyangsuk Min, Yuho Lee, Minjeong Ban, Jiaqi Deng, Nicole Hee-Yeon Kim, Taewon Yun, Hang Su, Jason Cai, and Hwanjun Song. 2025. Towards multi-dimensional evaluation of LLM summarization across domains and languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14417–14450, Vienna, Austria. Association for Computational Linguistics.

Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Apostolos Dedeloudis, Jackson Sargent, and David Jurgens. 2022. POTATO: The portable text annotation tool. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 327–337, Abu Dhabi, UAE. Association for Computational Linguistics.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is ChatGPT a general-purpose natural language processing task solver? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1339–1384, Singapore. Association for Computational Linguistics.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2025. A systematic survey of prompt engineering in large language models: Techniques and applications. *Preprint*, arXiv:2402.07927.

Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, Pranav Sandeep Dulepet, Saurav Vidyadhara, Dayeon Ki, Sweta Agrawal, Chau Pham, Gerson Kroiz, Feileen Li, Hudson Tao, Ashay Srivastava, Hevander Da Costa, Saloni Gupta, Megan L. Rogers, Inna Goncearenco, Giuseppe Sarli, Igor Galynker, Denis Peskoff, Marine Carpuat, Jules White, Shyamal Anadkat, Alexander Hoyle, and Philip Resnik. 2025. The prompt report: A systematic survey of prompt engineering techniques. *Preprint*, arXiv:2406.06608.

Liyan Tang, Igor Shalyminov, Amy Wong, Jon Burnsky, Jake Vincent, Yu'an Yang, Siffi Singh, Song Feng, Hwanjun Song, Hang Su, Lijia Sun, Yi Zhang, Saab Mansour, and Kathleen McKeown. 2024. TofuEval: Evaluating hallucinations of LLMs on topic-focused dialogue summarization. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4455–4480, Mexico City, Mexico. Association for Computational Linguistics.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. 2024. Large language models as optimizers. In *The Twelfth International Conference on Learning Representations*.

Xianjun Yang, Yan Li, Xinlu Zhang, Haifeng Chen, and Wei Cheng. 2023. Exploring the limits of chatgpt for query or aspect-based text summarization. *Preprint*, arXiv:2302.08081.

Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023. Extractive summarization via ChatGPT for faithful summary generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3270–3278, Singapore. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Yongxin Zhou, François Portet, and Fabien Ringeval. 2022. Effectiveness of French language models on abstractive dialogue summarization task. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3571–3581, Marseille, France. European Language Resources Association.

Yongxin Zhou, Fabien Ringeval, and François Portet. 2024. PSentScore: Evaluating sentiment polarity in dialogue summarization. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13290–13302, Torino, Italia. ELRA and ICCL.

A Experiment details: BART-based models

DialogSum We fine-tuned the BART-Large model[11] on the DialogSum dataset using the following hyperparameters: a learning rate of 5×10^{-5} for 15 epochs, a training batch size of 2, and an evaluation batch size of 8. The maximum source and target lengths were set to 1024 and 128 tokens, respectively, with a length penalty of 1.0. The experiments were conducted on an NVIDIA Quadro RTX 6000 GPU, with each run taking approximately 2.5 hours.

DECODA We fine-tuned the BARThez model[12] on the preprocessed DECODA dataset. The model, which uses a base architecture with 6 encoder and 6 decoder layers, was trained for 10 epochs, and the checkpoint with the lowest loss on the development set was saved. We employed the default parameters for summarization tasks: an initial learning rate of 5×10^{-5}, a training and evaluation batch size of 4, and a seed of 42. The Adam optimizer with a linear learning rate scheduler was used. All experiments were run on an NVIDIA Quadro RTX 8000 48GB GPU, with each training run taking approximately 25 minutes.

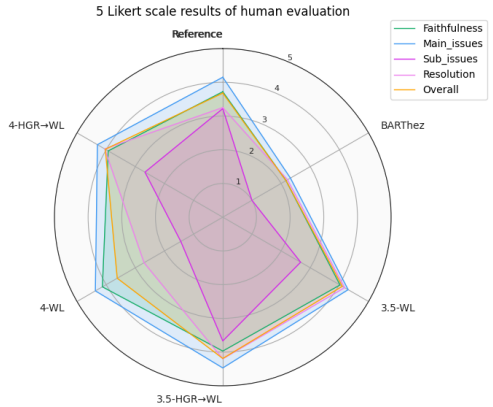

Figure 2: DECODA human evaluation results on a 5-point Likert scale. The radar chart compares six summaries (including reference summaries and those generated by various systems) across four criteria and the overall score.

[11] https://huggingface.co/facebook/bart-large
[12] https://huggingface.co/moussaKam/barthez

B LLM-as-judge

For the LLM-as-judge approach, we adapted the G-Eval framework (Liu et al., 2023), modifying its criteria to align with DECODA's dialogue summarization guidelines. Specifically, we evaluated summaries based on *Faithfulness* (Figure 3), *Main Issues* (Figure 4), *Sub-Issues* (Figure 5), and *Resolution* (Figure 6).

Each criterion was assessed via a distinct, specially designed prompt (see corresponding figures).

C Human Evaluation

C.1 Full results

The full human evaluation results for the DECODA dataset are detailed in Table 10, categorized into three groups: all 20 dialogues, the 10 shortest dialogues and the 10 longest dialogues.

C.2 Radar chart

The results for the 20 selected samples are presented in the radar chart in Figure 2. This visualization shows the evaluation of six summaries, including both reference summaries and those generated by different systems, across four criteria and the overall scores. The visual representation indicates that GPT-generated summaries are generally preferred by human annotators over those from the task-specific pre-trained model BARThez, and that they outperform the reference summaries on some criteria, such as *Faithfulness*.

D Example Analysis of Outputs

DialogSum We present three examples in the following tables. We observe that, when using prompts designated as `HG(R)`, GPT models tend to generate long summaries that are too precise and detailed, sometimes exceeding the length of the original conversation. Yet the instructions explicitly state that the summary should not exceed 20% of the original dialogue. Consequently, GPT models fail to fully adhere to prescribed human summarization guidelines.

In detail, Table 11 reveals that the summary generated by `ChatGPT` with the `WordLimit` prompt is of higher quality than that produced by `GPT-4`. It talks not only about the benefits of the job, but also about its demanding schedule. However, for `ChatGPT`, we found `HG`→`WL` to be quite good, whereas `HGR`→`WL` seems to tell the story from *Person1*'s point of view instead of *Person2*'s. As for `GPT-4`, the summary predicted by `HGR`→`WL` is not very good, as it only talks about the benefits, omitting the demanding schedule of this new job.

In Table 12, the `WordLimit` prompt yields well-sized summaries for both GPT models, whereas other prompts produce excessively long outputs – particularly notable given the short length of the source conversation. In Table 13, the predictions of `WordLimit` prompt are good for both GPT models, as does the `HG(R)`→`WL` prompt.

DECODA We present an example of the sample *FR_20091112_RATP_SCD_0826* in Table 14. We note that the synopses generated with `WordLimit` and `HG(R)`→`WL` prompts for both GPT models are good. The synopses generated with `HG(R)` prompt follow the structure of the given instructions: first talk about the main issue, then the sub-issues of the conversation, and finally the resolution of the call, even the predictions are long, but make the two-step prompt approach – `HG(R)`→`WL` generating quite good synopses.

E Output Length

Figure 7 and Figure 8 present summary length statistics for different prompts on the DialogSum and DECODA datasets, respectively, compared to reference summaries.

The summaries generated by `ChatGPT` are consistently longer than the reference summaries. Applying the `WordLimit` prompt reduced the average output length, and using it as a second step after the `HG(R)` prompt produced more concise summaries (`HGR`→`WL`).

Regarding `GPT-4`, it demonstrated better adherence to length instructions. The distributions for `4-WL` and `4-HGR`→`WL` are more constrained in both figures, while `4-HGR` exceeds the reference summary length.

F Manual Error Analysis

F.1 DialogSum Discrepancy Data Points

Table 15 presents DialogSum cases with the most significant scoring discrepancies (low ROUGE-L but high BERTScore) for `4-HGR`→`WL` summaries.

F.2 DECODA Discrepancy Data Points

Table 16 displays English translations of DECODA cases showing the most significant scoring discrepancies (low ROUGE-L but high BERTScore) for `4-HGR`→`WL` summaries.

You will receive a dialogue. You will then receive a written summary of this dialogue.
Your task is to evaluate this summary against a single criterion.
Please read these instructions carefully and ensure you understand them. Keep this document open during evaluation and refer to it as needed.

Evaluation Criterion:
Faithfulness (1-5) - The summary must be faithful to the dialogue in terms of factual information.

The 5-point Likert scale is as follows:

1. **Very unfaithful**: The summary contains many factual errors and omits important information from the dialogue.
2. **Somewhat unfaithful**: The summary contains some factual errors and/or omits some important information from the dialogue.
3. **Moderately faithful**: The summary is generally consistent with the dialogue but contains some minor errors or inaccuracies.
4. **Largely faithful**: The summary is very close to the dialogue, with very few factual errors or minor inaccuracies.
5. **Very faithful**: The summary is perfectly consistent with the dialogue, with no factual errors or significant omissions.

Example:

Source text:
`{{Document}}`
Summary:
`{{Summary}}`

Evaluation form (ONLY ratings):

– Faithfulness:

Figure 3: Prompt used to evaluate the *Faithfulness* dimension.

	Faithfulness	Main Issues	Sub-issues	Resolution	Overall
All Dialogues					
Reference	3.72±1.12	4.15±1.07	3.22±1.86	3.25±1.60	3.68±1.36
BARThez	2.17±1.43	2.32±1.57	1.00±0.0	2.22±1.54	2.17±1.49
3.5-WL	4.02±1.17	4.30±1.18	2.67±1.73	4.18±1.28	4.10±1.28
3.5-HGR→WL	3.97±1.15	**4.47±0.89**	**3.67±1.73**	**4.20±1.13**	**4.19±1.12**
4-WL	**4.13±0.96**	4.37±0.80	1.44±1.33	2.70±1.79	3.62±1.53
4-HGR→WL	3.93±1.02	4.30±0.96	2.67±1.66	4.05±1.27	4.03±1.16
10 Short Dialogues					
Reference	3.53±1.07	4.07±1.11	5.00	3.17±1.70	3.60±1.37
BARThez	2.27±1.39	2.20±1.63	1.00	2.27±1.57	2.23±1.51
3.5-WL	3.83±1.32	4.13±1.33	1.00	**4.47±1.04**	4.11±1.29
3.5-HGR→WL	4.27±1.01	**4.67±0.80**	5.00	4.37±1.10	**4.44±0.98**
4-WL	**4.37±0.89**	4.47±0.78	5.00	4.17±1.21	4.34±0.97
4-HGR→WL	4.10±0.92	4.63±0.61	5.00	4.20±1.30	4.32±1.00
10 Longest Dialogues					
Reference	3.90±1.16	4.23±1.04	3.00±1.85	3.33±1.52	3.76±1.36
BARThez	2.07±1.48	2.43±1.52	1.00±0.0	2.17±1.53	2.12±1.48
3.5-WL	**4.20±1.00**	**4.47±1.01**	2.88±1.73	3.90±1.45	**4.08±1.27**
3.5-HGR→WL	3.67±1.21	4.27±0.94	**3.50±1.77**	**4.03±1.16**	3.95±1.19
4-WL	3.90±0.99	4.27±0.83	1.00±0.0	1.23±0.77	2.96±1.65
4-HGR→WL	3.77±1.10	3.97±1.13	2.38±1.51	3.90±1.24	3.76±1.24

Table 10: Mean (± std) scores for *Faithfulness*, *Main issues*, *Sub-issues*, *Resolution*, and *Overall* of each summary assessed by three native human evaluators on the 10 shortest and 10 longest dialogues in the DECODA dataset. The best results for each criterion are shown in bold.

Tables 17 and 18 present original French DECODA cases with the most significant scoring discrepancies (low ROUGE-L but high BERTScore) for `4-WL` and `4-HGR→WL` summaries, respectively, as analyzed in Section 4.3.

You will receive a dialogue. You will then receive a written summary of this dialogue.
Your task is to evaluate this summary against a single criterion.
Please read these instructions carefully and ensure you understand them. Keep this document open during evaluation and refer to it as needed.

Evaluation Criterion:
Main issues (1-5) - The summary should capture the main topics of the conversation. These are the problems for which the customer called. Identifying these issues forms the basis for classifying the call into different motivation categories. The main problem of a call should be prioritized for inclusion in the summary.

The 5-point Likert scale is as follows:

1. **Incorrect or missing**: The main issues of the call are presented incorrectly, are missing, or are completely different from those discussed during the call.
2. **Confusing**: The main issues of the call are presented in a confusing or ambiguous way, making it difficult to understand the problems addressed.
3. **Correct but incomplete**: The main issues are correctly identified, but the presentation lacks details or context needed for full understanding.
4. **Correct but could be improved**: The main issues are correctly identified and clearly presented, but some details or clarifications are missing for complete understanding.
5. **Precise and concise**: The main issues are presented clearly, concisely, and precisely, enabling full understanding of the problems addressed during the call.

Example:

Source text:
`{{Document}}`
Summary:
`{{Summary}}`

Evaluation form (ONLY ratings):

- Main issues:

Figure 4: Prompt used to evaluate the *Main Issues* dimension.

G Datasets: License or Terms of Use

The DialogSum corpus used in this study comprises resources that are freely available online without copyright restrictions for academic use. For the DECODA dataset, the training and validation sets can be downloaded from their website[13] upon acceptance of the corresponding usage and sharing terms, while the test set is available only upon request from the authors.

[13] https://pageperso.lis-lab.fr/benoit.favre/ccs/

You will receive a dialogue. You will then receive a written summary of this dialogue.
Your task is to evaluate this summary against a single criterion.
Please read these instructions carefully and ensure you understand them. Keep this document open during evaluation and refer to it as needed.

Evaluation Criterion:
Sub-issues (1-5) - The summary should also note the sub-issues of the conversation: when a sub-issue appears in the conversation, it may be introduced by the customer or by the agent(s).

The 5-point Likert scale is as follows:

1. **Incorrect or missing**: The sub-issues of the call are presented incorrectly, are missing, or are completely different from those discussed during the call.
2. **Confusing**: The sub-issues of the call are presented in a confusing or ambiguous way, making it difficult to understand the problems addressed.
3. **Incomplete**: The sub-issues are correctly identified, but the presentation lacks details or context needed for a full understanding of the problems addressed.
4. **Correct but could be improved**: The sub-issues are correctly identified and clearly presented, but some details or clarifications are missing for complete understanding.
5. **Precise and concise**: The sub-issues are presented clearly, concisely, and precisely, enabling a full understanding of the problems addressed during the call.

Note: If no sub-issues are present in the conversation, leave the "Sub-issues" field as "N/A" for all summaries.

Example:

Source text:
`{{Document}}`
Summary:
`{{Summary}}`

Evaluation form (ONLY ratings):

- Sub-issues:

Figure 5: Prompt used to evaluate the *Sub-issues* dimension.

You will receive a dialogue. You will then receive a written summary of this dialogue.
Your task is to evaluate this summary against a single criterion.
Please read these instructions carefully and ensure you understand them. Keep this document open during evaluation and refer to it as needed.

Evaluation Criterion:
Resolution (1-5) - The summary should also note the call resolution: that is, whether the customer's problem was resolved during this call (first-call resolution) or not.

The 5-point Likert scale is as follows:

1. **Missing**: The summary does not mention whether the customer's problem was resolved or not.
2. **Confusing**: The summary vaguely mentions the resolution, but it is difficult to understand whether the problem was resolved or not.
3. **Present but incomplete**: The summary mentions that the problem was resolved, but details are missing or the resolution is not presented clearly.
4. **Correct but could be improved**: The summary mentions that the problem was resolved and provides some details, but still lacks information for a complete understanding.
5. **Precise and concise**: The summary clearly and detailedly presents the resolution of the customer's problem, enabling an understanding of how the problem was resolved.

Example:

Source text:
`{{Document}}`
Summary:
`{{Summary}}`

Evaluation form (ONLY ratings):

- Resolution:

Figure 6: Prompt used to evaluate the *Resolution* dimension.

test_20
Dialogue: #Person1#: Hey, Frank. I heard you got a new job. #Person2#: Yeah, Judy. I will be working for the Post Office. It's not a bad job. #Person1#: Is it true that you have a heavy work schedule? #Person2#: That's right. I am supposed to work at 5am everyday, and I only get 45 minutes for lunch. #Person1#: So, why did you apply for such a demanding job? #Person2#: Well, the government offers its employees excellent health insurance benefits. #Person1#: Oh, I see. And can your family members utilize the health insurance, too?#Person2#: Yeah, that's the best part. All of my children can get free medical care. #Person1#: That's a great employment benefit! #Person2#: Now you know why I wanted to work for the Post Office! **Reference Summaries:** **Summary1:** Frank got a new job and is telling Judy not only the heavy schedule but also the benefits of this job. **Summary2:** Frank explains to Judy the reason why he took a job at the Post office is that the government offers excellent health insurance benefits for its employees. **Summary3:** Frank describes his new job's schedule and insurance at the Post Office to Judy.
`ChatGPT` Generated Summaries
3.5-WL: Person2 got a new job at the Post Office with a demanding work schedule, but the excellent health insurance benefits for family members made it worth it.
3.5-HG: Person1 informs Person2 that they heard about Person2's new job. Person2 confirms that they will be working for the Post Office and mentions that it is not a bad job. Person1 asks if it is true that Person2 has a heavy work schedule. Person2 confirms and explains that they are supposed to work at 5am every day and only have 45 minutes for lunch. Person1 asks why Person2 applied for such a demanding job. Person2 explains that the government offers excellent health insurance benefits to its employees. Person1 understands and asks if Person2's family members can also utilize the health insurance. Person2 confirms that all of their children can receive free medical care. Person1 acknowledges that it is a great employment benefit. Person2 concludes by stating that this is why they wanted to work for the Post Office.
3.5-HGR: Person1 informs Person2 that they heard about Person2's new job. Person2 confirms and mentions that they will be working for the Post Office. Person1 asks about the work schedule, and Person2 explains that they have to work at 5am every day and have a short lunch break. Person1 questions why Person2 applied for such a demanding job, and Person2 explains that the government offers excellent health insurance benefits. Person1 realizes that family members can also utilize the health insurance, and Person2 confirms. Person1 comments on the great employment benefit, and Person2 explains that this is why they wanted to work for the Post Office.
3.5-HG→WL: Person2 tells Person1 about their new job at the Post Office, mentioning the demanding work schedule but highlighting the excellent health insurance benefits for their family.
3.5-HGR→WL: Person1 learns about Person2's new job at the Post Office, which has demanding hours but offers excellent health insurance benefits for family members.
`GPT-4` Generated Summaries
4-WL: Frank got a new job at the Post Office for its excellent health insurance benefits for his family.
4-HGR: In the conversation, Judy learns that Frank has secured a new job at the Post Office, which demands an early start at 5am and allows only a short lunch break. Despite the rigorous schedule, Frank's decision to take up the job is driven by the excellent health insurance benefits offered by the government to its employees. This benefit extends to his children, providing them with free medical care, which Judy acknowledges as a significant employment advantage.
4-HGR→WL: Frank accepts a demanding Post Office job for its excellent health insurance benefits, covering his children's medical care.

Table 11: Input dialogue, reference summaries, and model-generated summaries for test sample *test_20*.

test_151
Dialogue: #Person1#: Have you enjoyed your weekend? #Person2#: Yes, it's been marvelous. It really was very kind of you to invite me. I hope it hasn't been too much trouble. #Person1#: Not at all. We've really enjoyed having you. I hope you'll come and stay again next time you're in Changsha. **Reference Summaries:** **Summary 1:** #Person2# tells #Person1# #Person2# enjoyed the weekend organized by #Person1#. **Summary 2:** #Person1# and #Person2# has enjoyed a nice weekend together. **Summary 3:** #Person1# invited #Person2# to spend the weekend and #Person2# has enjoyed it.
`ChatGPT` Generated Summaries
3.5-WL: Person2 had a marvelous weekend and thanked Person1 for the invitation, hoping to stay again in the future.
3.5-HG: Person1 asks Person2 if they have enjoyed their weekend. Person2 responds positively, saying it has been marvelous and thanking Person1 for the invitation. Person2 expresses concern that it may have been too much trouble, but Person1 reassures them that it was not and that they have enjoyed having Person2. Person1 also invites Person2 to come and stay again next time they are in Changsha.
3.5-HGR: Person1 asks Person2 if they have enjoyed their weekend. Person2 responds positively, saying it has been marvelous and thanking Person1 for the invitation. Person1 reassures Person2 that it was not trouble at all and expresses their enjoyment of having Person2. Person1 also invites Person2 to stay again in Changsha in the future.
3.5-HG→WL: Person2 had a marvelous weekend and thanked Person1 for the invitation. Person1 reassured them and invited them to visit again.
3.5-HGR→WL: Person2 had a marvelous weekend and thanked Person1 for the invitation. Person1 enjoyed having Person2 and invited them to visit again in Changsha.
`GPT-4` Generated Summaries
4-WL: Person1 invites Person2 for a weekend stay in Changsha, which Person2 thoroughly enjoys.
4-HGR: Person1 inquires if Person2 enjoyed their weekend, to which Person2 responds affirmatively, expressing gratitude for the invitation and concern about any potential inconvenience caused. Person1 dismisses any notion of trouble and extends another invitation for Person2 to visit again when they are next in Changsha.
4-HGR→WL: Person1 invites Person2 again after a pleasant weekend visit in Changsha.

Table 12: Input dialogue, reference summaries, and model-generated summaries for test sample *test_151*.

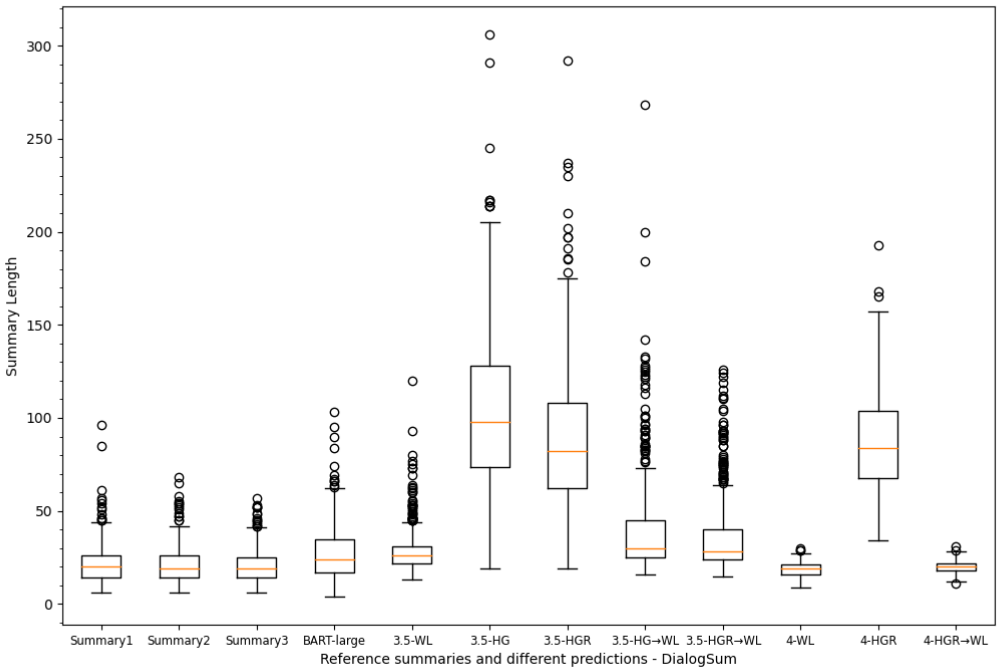

Figure 7: Distribution of generated summary lengths on the DialogSum test set, compared to the three reference summaries.

test_440
Dialogue: #Person1#: Man, I'm freaking out! You gotta help me! #Person2#: Whoa, whoa, take it easy, relax. Geez, you're sweating like a pig! What's going on? #Person1#: I can't go through with this! I just can't! I'm not ready for marriage! What was I thinking? I'm only thirty five years old! I've got my entire life ahead of me, adventures waiting! I can't settle down yet! #Person2#: What are you talking about? It wasn't more than a month ago that you were rambling on about how you are tired of living the life of a bachelor and how you envy your friends that have a family! #Person1#: I know. I know! #Person2#: Let's think this through. First of all, you cannot leave Amy at the altar. Not only will she be humiliated and kill you, but she is the woman that you love and the woman of your dreams! Why would you want to end that or jeopardize it like this? Second of all, you are just getting cold feet. You know deep down inside that you want to marry her, so cut the crap and do it! #Person1#: You're right. I'm being crazy! Ok, I'm ready, let's do this! #Person2#: Great! Phew! That was a close one! You had me going there for a minute I thought I was gonna have to slap some sense into you. **Reference Summaries:** **Summary 1:** #Person1# is not ready for marriage. #Person2# reminds him of what he said a month ago and what will happen if he leaves Amy. After listening to #Person2#, #Person1# is ready for marriage. **Summary 2:** #Person1# feels anxious about future marriage and sweats a lot. #Person2# helps #Person1# to think through the current situation and get ready for the marriage. **Summary 3:** #Person1# is freaking out before his marriage with Amy and #Person2# helps him get over it by helping him think through all the details.
`ChatGPT` Generated Summaries
3.5-WL: Person 1 is having second thoughts about getting married, but Person 2 convinces them to go through with it.
3.5-HG: Person1 is feeling anxious and unsure about getting married. They express their concerns about settling down at their age and missing out on adventures. Person2 tries to calm them down and reminds them of their previous desire to have a family. Person2 also emphasizes the importance of not leaving Amy at the altar and reminds Person1 of their love for her. Person1 realizes they are just getting cold feet and agrees to go through with the wedding. Person2 is relieved and jokes about having to slap some sense into Person1.
3.5-HGR: Person1 is panicking and expressing doubts about getting married. Person2 tries to calm Person1 down and reminds them of their previous desire to settle down. Person2 also emphasizes the importance of not leaving Amy at the altar and encourages Person1 to go through with the wedding. Person1 eventually agrees and decides to proceed with the marriage. Person2 expresses relief and jokes about having to intervene if Person1 didn't change their mind.
3.5-HG→WL: Person1 is anxious about getting married but Person2 reassures them and reminds them of their love for their partner.
3.5-HGR→WL: Person1 has doubts about getting married, but Person2 convinces them to go through with it, emphasizing the importance of not leaving Amy at the altar.
`GPT-4` Generated Summaries
4-WL: A man panics before his wedding, but his friend calms him down and convinces him to go through with it.
4-HGR: In a state of panic, Person1 expresses his fear of getting married to Person2, despite being thirty-five years old. He feels he's not ready to settle down and still has many adventures to experience. Person2 reminds him of his previous desire to leave bachelorhood and his envy of his friends with families. Person2 also emphasizes that Person1 cannot leave his fiancée, Amy, at the altar, as it would not only humiliate her but also jeopardize their relationship. Person2 convinces Person1 that his fear is just a case of cold feet and that he truly wants to marry Amy. Accepting this, Person1 calms down and decides to proceed with the wedding, much to Person2's relief.
4-HGR→WL: Person1, fearing marriage, is reassured by Person2 and decides to proceed with his wedding to Amy.

Table 13: Input dialogue, reference summaries, and model-generated summaries for test sample *test_440*.

Dialogue:
<Spk_A> bonjour
<Spk_B> oui bonjour madame
<Spk_B> je vous appelle pour avoir des horaires de train en+fait c' est pas pour le métro je sais pas si vous pouvez me les donner ou pas
<Spk_A> trains SNCF
<Spk_B> oui trains SNCF oui
<Spk_A> vous prenez quelle ligne monsieur
<Spk_B> euh la ligne euh enfin en+fait c' est pas SNCF enfin c' est Île-de-France quoi je sais pas comment ils appellent ça
<Spk_B> RER voilà c' est pour les RER
<Spk_B> voilà et euh je prends la ligne euh Meaux de Meaux pour aller à Paris je sais plus c' est laquelle c' est la
<Spk_B> E je crois
<Spk_A> d'accord et vous voulez donc les horaires euh pour quel jour et à quel moment monsieur
<Spk_B> euh là pour euh tout à l' heure euh pour euh aux environs de dix-sept heures en partant de la gare de Meaux
<Spk_B> euh vers la Gare+de+l'Est à Paris
<Spk_A> alors vous partez de Meaux et vous allez donc à la Gare+de+l'Est <Spk_A> et vous voudriez les horaires
<Spk_B> voilà
<Spk_A> vers euh dix-sept heures <Spk_A> alors moi je peux regarder ce+que <Spk_B> ouais dix-sept heures ouais
<Spk_A> j' ai comme horaires un instant monsieur <Spk_A> s'il+vous+plaît <Spk_B> d'accord il y a
<Spk_B> pas de souci
<Spk_A> monsieur s'il+vous+plaît
<Spk_B> oui
<Spk_A> donc voilà ce+que j' ai comme horaires moi vous avez donc un départ à la gare de Meaux donc à seize heures quarante-neuf
<Spk_B> seize heures quarante-neuf d'accord
<Spk_A> et après vous avez celui de dix-sept heures dix-neuf
<Spk_B> alors seize heures quarante-neuf dix-sept heures dix-neuf <Spk_A> OK d'accord <Spk_B> oui
<Spk_B> ben je vous remercie
<Spk_A> mais je vous en prie <Spk_B> bonne journée madame
<Spk_A> merci à vous aussi monsieur au+revoir
<Spk_B> au+revoir

Reference Synopses:
synopsis 1: Demande d'horaires de train de la gare de Maux à la gare de l'Est à une heure donnée.
synopsis 2: horaires RER E de Meaux à la Gare de l'Est

`ChatGPT` Generated Summaries

3.5-WL: Un homme appelle pour demander les horaires des trains RER entre Meaux et Paris. Il obtient les horaires pour un départ à 16h49 et à 17h19.

3.5-HG: Le client appelle pour obtenir des horaires de train pour la ligne RER de Meaux à Paris. Il demande les horaires pour un départ aux environs de 17 heures. L'agent recherche les horaires et lui donne deux options : un départ à 16h49 et un autre à 17h19. Le client remercie l'agent et l'appel se termine.

3.5-HGR: Le client appelle pour obtenir des horaires de train pour la ligne RER de Meaux à Paris. Il demande les horaires pour un départ aux environs de 17 heures. L'agent recherche les horaires et lui donne deux options : un départ à 16h49 et un autre à 17h19. Le client remercie l'agent et l'appel se termine.
Le problème principal de l'appel est la demande des horaires de train pour la ligne RER de Meaux à Paris. Les sous-problèmes incluent la clarification de la ligne de train (RER) et la spécification de l'heure de départ (aux environs de 17 heures).
L'appel est résolu au cours de cet appel, car l'agent fournit les horaires demandés au client.

3.5-HG→WL: Le client appelle pour obtenir les horaires de train de la ligne RER de Meaux à Paris. L'agent lui propose deux options de départ : 16h49 ou 17h19. L'appel est résolu.

3.5-HGR→WL: Le client appelle pour obtenir les horaires de train de la ligne RER de Meaux à Paris. L'agent lui propose deux options de départ : 16h49 ou 17h19. L'appel se termine avec satisfaction.

`GPT-4` Generated Summaries

4-WL: Un homme appelle pour connaître les horaires de train de la ligne RER E entre Meaux et Paris pour le jour même.

4-HGR: Le client a appelé pour obtenir des informations sur les horaires des trains RER de la ligne E, de Meaux à Paris Gare de l'Est, pour le jour même aux environs de 17 heures. L'agent a pu fournir les horaires demandés, indiquant un départ à 16h49 et un autre à 17h19. Le client a remercié l'agent pour les informations fournies et a terminé l'appel. Le problème principal de l'appel, à savoir la demande d'horaires de train, a été résolu lors de cet appel.

4-HGR→WL: Un client a demandé les horaires des trains RER E de Meaux à Paris, l'agent a fourni les informations nécessaires.

Table 14: Input dialogue, reference synopses, and model-generated summaries for DECODA test sample *FR_20091112_RATP_SCD_0826*.

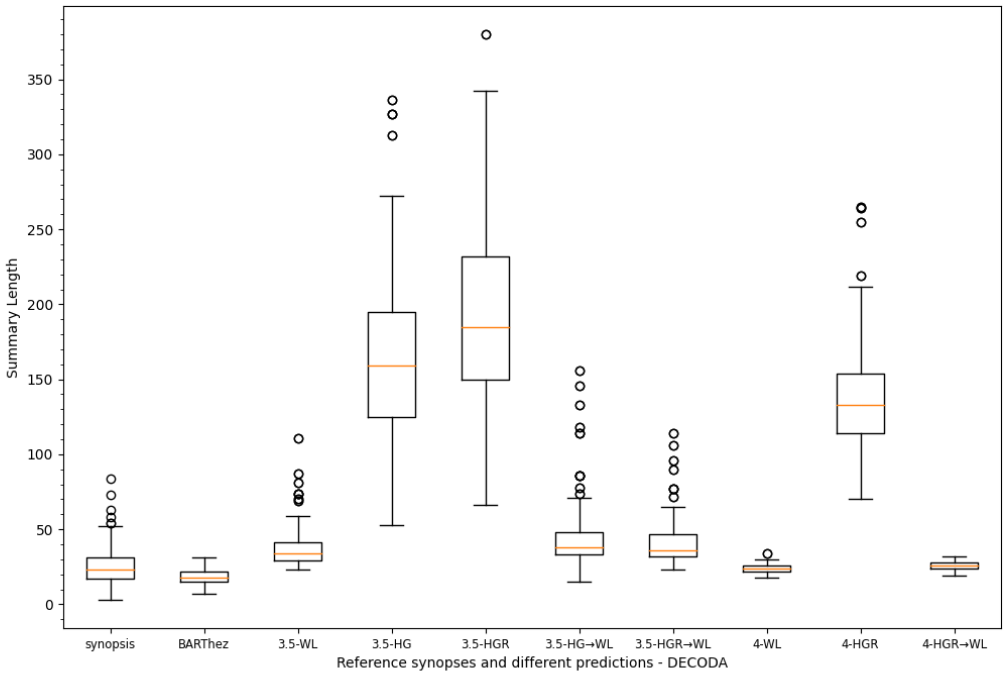

Figure 8: Distribution of generated summary lengths on the DECODA test set, compared to the reference synopses.

ID	References	Prediction	RL	BS	Diff.
246	["Tim"s project is about how to lead a greener life. #Person1# consults about his motivation and his findings at school. Tim finds there are many aspects that can be improved for his school.", "Tims is talking about his greener life. He now rides his bike to school and recommends teachers to send electrical paper rather than paper ones. And he thinks the school shouldn"t have machines with bottled drinks.", "Tim chooses a green-life-related project because he wants to see what he could do in his own life. He"s started being greener with himself and influences others."]	#Person2# proposes eco-friendly changes at school and personal life, inspired by environmental science lessons.	6.77	86.81	80.04
232	["#Person1# asks #Person2# about past experiences of living in different cities around the world.", "#Person1# is asking #Person2# about the places that #Person2# has stayed or lived.", "#Person1# and #Person2# talk about #Person2#"s experience. #Person2# moves a lot. #Person1# finds it interesting."]	#Person2# shares their life journey from Chicago to Los Angeles and Munich with #Person1#.	9.29	89.26	79.96
283	["#Person2# wants to buy some clothes but #Person1#"s recommendation doesn"t interest #Person2#.", "#Person1# gives #Person2# a few suggestions on clothes but none is taken.", "#Person2# is not happy with #Person1#"s recommendeds."]	Salesperson suggests brown, silk, and cotton dresses to a customer who prefers a white dress for the warm season.	4.17	83.96	79.79
433	["Amy is talking about her first job experience and later job with #Person1#.", "Amy tells #Person1# about her first job experience and her current job.", "Amy shares her former work experience and her current job with #Person1#."]	Amy, after starting as a secretary in 1998, became a Personnel Manager, then moved to Sales due to salary dissatisfaction.	6.19	85.7	79.51
379	["#Person1# interviews Vet, a high school girl who has become a mother. She shows her love to her daughter and talks about her future plans and gives advice to teens.", "Vet tells her own experience as a young mother and how her daughter changes her. Then she tells #Person1# her future goals and dreams. And she also gives teens advice.", "#Person1# is interviewing Vet and asking her about the ways she deals with stress as a young mother, feelings of having a baby, plans for the future, her favorite quote, and advice for teens."]	#Person2# discusses stress coping mechanisms, the impact of motherhood, future ambitions, and advises teenagers on self-love and independence.	11.83	89.75	77.92

Table 15: Data points where GPT-generated summaries (4-HGR→WL) show the greatest discrepancy in the DialogSum corpus, with low ROUGE-L but high BERTScore values. RL and BS are the average scores of three references.

ID	Reference	Prediction	RL	BS	Diff.
10	Inquiry about a corporate account order. Transfer to the relevant department.	A representative from the Ministry of Foreign Affairs is organizing a seminar in Paris and contacts RATP to obtain transport cards for participants.	0.00	70.14	70.14
135	Route request from Drancy station to Gare du Nord. Clarification of explanations following a previous call.	A customer confused about RER B connection to reach Drancy is reassured they can travel directly from Bourg-la-Reine.	4.55	70.29	65.75
29	Inquiry about purchasing tickets for school groups. Transfer to the relevant department.	A school association representative calls to understand public transport procedures for students. The agent proposes a solution and transfers the call.	8.51	72.83	64.32
68	Wrong number - caller meant to dial social security at 36 46 but misdialed	A customer mistakes RATP customer service for Social Security regarding a health card issue in the Antilles.	4.65	66.03	61.38
140	No ImaginR card subscription by mail, only at ticket counters in all stations	A customer requested an Imagine+R application by mail. The agent explained it's impossible but provided the service number for more information.	4.17	65.54	61.38

Table 16: Data points where GPT-generated summaries (4-HGR→WL) show the greatest discrepancy in the DECODA corpus, with low ROUGE-L but high BERTScore values. This is the translated version; the original texts are in Table 18 in the Appendix.

ID	Reference	Prediction	RL	BS	Diff.
10	Demande d'information sur une commande grands comptes. Transfert vers le service concerné.	Un organisateur de séminaire cherche à obtenir des cartes de transport pour ses invités internationaux auprès de la RATP.	0	71.13	71.13
110	Demande de renseignements suite à la procédure suivi après la réception d'un courrier d'amende. Transfert vers le service concerné.	Une cliente appelle le service clientèle pour comprendre pourquoi elle doit payer cinq euros suite à une amende de bus contestée et justifiée.	8.33	70.86	62.53
107	demande procédure pour remboursement suite à achat tickets car envoi tardif du passe ImaginR, attente au-delà de 3 semaines donc possible, appeler ImaginR ou par mail	Une cliente appelle pour demander le remboursement d'une carte Orange achetée en attendant sa carte Imagine+R, arrivée en retard.	8.00	69.35	61.35
32	circulation du T2	Un utilisateur se renseigne sur la mise en service du T+deux de la Porte de Versailles à Val d'Issy.	8.70	69.85	61.15
29	Demande de renseignements sur l'achat de ticket pour des groupes scolaires. Transfert vers le service concerné.	Un représentant d'une association départementale cherche à organiser le transport de plus de 100 classes via les transports en commun.	10.26	71.05	60.80

Table 17: Data points where GPT-generated summaries (4-WL) show the greatest discrepancy in the DECODA corpus, with low ROUGE-L but high BERTScore values.

ID	Reference	Prediction	RL	BS	Diff.
10	Demande d'information sur une commande grands comptes. Transfert vers le service concerné.	Un représentant du Ministère des Affaires étrangères organise un séminaire à Paris et contacte la RATP pour obtenir des cartes de transport pour les participants.	0	70.14	70.14
135	Demande d'itinéraire gare de Drancy gare du Nord. Précision des explications suite à un précédent appel.	Un client confus sur l'interconnexion du RER B pour se rendre à Drancy est rassuré qu'il peut voyager directement depuis Bourg-la-Reine.	4.55	70.29	65.75
29	Demande de renseignements sur l'achat de ticket pour des groupes scolaires. Transfert vers le service concerné.	Un représentant d'une association scolaire appelle pour comprendre la procédure de transport en commun pour les étudiants. L'agent propose une solution et transfère l'appel.	8.51	72.83	64.32
68	appel erroné l'appelant ayant voulu joindre l'assurance maladie au 36 46 et ayant mal tapé le numéro	Un client confond le service client de la RATP avec la Sécurité Sociale, concernant un problème de carte vitale aux Antilles.	4.65	66.03	61.38
140	aucun abonnement de la carte ImaginR par voie postale, seulement au guichet dans toutes les gares et stations	Un client a demandé un dossier Imagine+R par courrier. L'agent a expliqué que c'est impossible mais a fourni le numéro du service pour plus d'informations.	4.17	65.54	61.38

Table 18: Data points where GPT-generated summaries (4-HGR→WL) show the greatest discrepancy in the DECODA corpus, with low ROUGE-L but high BERTScore values.