Leveraging Few-Shot Data Augmentation and Waterfall Prompting for Response Generation

Lea Krause
Vrije Universiteit Amsterdam
l.krause@vu.nl

Selene Báez Santamaría
Vrije Universiteit Amsterdam
s.baezsantamaria@vu.nl

Michiel van der Meer
Leiden University
m.t.van.der.meer@liacs.leidenuniv.nl

Urja Khurana
Vrije Universiteit Amsterdam
u.khurana@vu.nl

Abstract

This paper discusses our approaches for task-oriented conversational modelling using subjective knowledge, with a particular emphasis on response generation. Our methodology was shaped by an extensive data analysis that evaluated key factors such as response length, sentiment, and dialogue acts present in the provided dataset. We used few-shot learning to augment the data with newly generated subjective knowledge items and present three approaches for DSTC11: (1) task-specific model exploration, (2) incorporation of the most frequent question into all generated responses, and (3) a waterfall prompting technique using a combination of both GPT-3 and ChatGPT.

1 Introduction

Task-Oriented Dialogue (TOD) Systems are traditionally designed to facilitate users in achieving specific objectives, such as looking up train times or booking a flight in a dialogue setting. For these tasks, the models are often given access to a database of factual information to complete the task. However, other tasks necessitate not only factual but also subjective insights, which are derived from other users' opinions. Handling subjective knowledge and using it for generating dialogue responses is the core of the *Subjective-Knowledge-based Task-Oriented Dialogue (SK-TOD)* (Zhao et al., 2023) challenge. The challenge is set up as conversations between users and artificial assistants, inquiring about and potentially booking hotels or restaurants. The organisers provided dialogue snapshots and a knowledge base with subjective reviews and FAQs related to said hotels and restaurants.

The challenge consists of three interlinked subtasks: 1) *Knowledge Seeking Turn Detection*, where it is determined whether a turn needs knowledge to create an appropriate response; 2) *Knowledge Selection*, where relevant items are selected from the knowledge base; and 3) *Knowledge-*
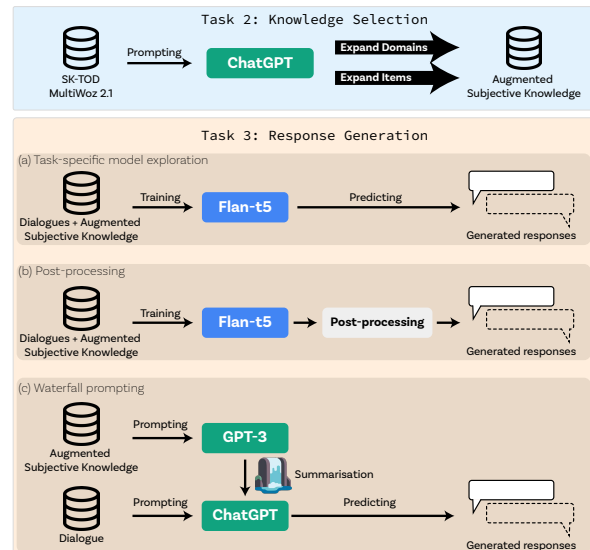

Figure 1: Overview of the approaches for *Knowledge Selection* and *Response Generation*. For approach 3a only the best-performing model is depicted.

Grounded Response Generation, where the dialogue history and the selected knowledge items must be aggregated to generate a concise response for the user.

Our primary interest was in sub-task 3, but due to the high correlation of performance with sub-task 2, we simultaneously worked on improving the knowledge selection (see Figure 1). Furthermore, we experimented with splitting sub-task 3 into two smaller tasks. First, we perform few-shot data augmentation, supported by a detailed data analysis, to substantially increase the knowledge base and address generalisation to new domains. Despite its marginal impact on the final test set, given no new domains were added, it forms a robust strategy for future generalisations. Second, we undertook the exploration of task-specific models, we found that using a larger model leads to slight improvements in task performance. In addition, we augmented the generated responses by integrating the most frequently asked question, aiming to

enhance conversation coherence and engagement. Lastly, we also employed a waterfall prompting technique that integrated a combination of Large Language Models. This strategy, though it scored lower in quantitative metrics compared to the baseline, possibly due to the abstractive nature of these models, is a promising avenue for further research[1].

2 Related work

Data Augmentation   In low-resource data scenarios, synthetic data augmentation has been shown to be a cost-efficient and effective way to increase dataset size (Anaby-Tavor et al., 2020). Traditional data augmentation techniques for natural language processing often include simple transformations like synonym substitution, random insertion, deletion, and swapping of words (Wei and Zou, 2019). A more complex approach is paraphrasing, where the meaning stays the same but words and syntax change, a popular example of this is back-translation in machine translation (Sennrich et al., 2016). Data augmentation was also used in successful approaches of last year's edition of DSTC (Tian et al., 2021; Thulke et al., 2022)

Recently, the emergence of large language models (LLMs) like GPT-3 (Brown et al., 2020) has opened up new avenues for data augmentation. One approach is to prompt these models to generate additional data that fits the distribution of the original dataset (Kumar et al., 2020; Xia et al., 2020; Lee et al., 2021). This leverages the capacity of LLMs to generate coherent and contextually appropriate sentences, thereby creating augmented data that closely mirrors real-world linguistic diversity.

Waterfall Prompting   Combining prompts or re-running the same prompt has shown increased performance and robustness. For example, Singhal et al. (2023) coin the term *ensemble refinement* which involves using a two-step prompting approach. In the first iteration, multiple answers are generated for a prompt, and in the second iteration, these answers are incorporated into the prompt to develop a final solution that is more robust for the given task. Similar efforts have been explored such as self-consistency (Wang et al., 2023), recitation-augmentation (Sun et al., 2023), and rational-augmentation (Wang et al., 2022). Pitis et al. (2023) expand on this framework by incorporating multiple rounds of prompting, each building

upon the improvements made in the previous round. However, these still focus on prompting a single model and differ from a typical ensemble where different models' output is aggregated (Wang et al., 2023). To the best of our knowledge, there is limited work on waterfall prompting or assembling the responses from different language models.

Previous DSTC editions   According to system reports from the DSTC9 and DSTC10 challenges, enhancing the performance of sub-task 2 significantly affects the overall task performance (Kim et al., 2020). In the DSTC9 iteration, it was found that ensemble model approaches are effective for knowledge selection (Kim et al., 2020). In DSTC10, successful approaches used a separate entity tracking component for knowledge selection to narrow down the search space before document ranking (Kim et al., 2022).

3 Data Analysis

Before developing our approach to the DSTC tasks, we analyse the available data. We 1. inspect the correlation between features, 2. investigate the sentiment of the available data, and 3. examine the distribution of dialogue act types.
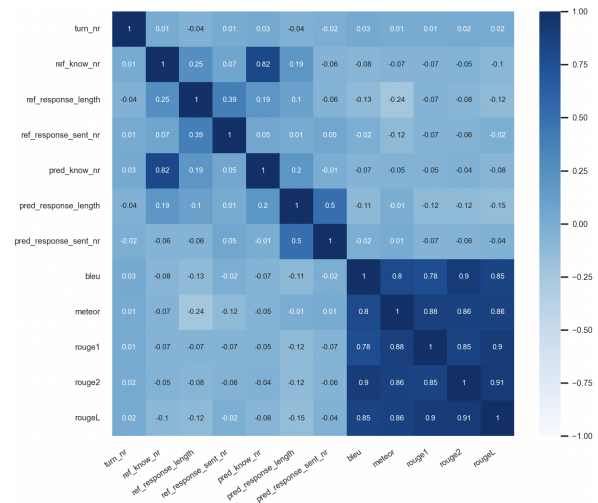

Figure 2: Correlation between reference data, baseline predictions, and automatic scores, on the validation set.

3.1 Feature correlation

To steer our approach, we investigate the potential influence of various factors on automatic metric scores. We examine correlations between the following variables: (1) the number of dialogue turns, (2) the number of selected knowledge items, and

---

(3) the lengths of both the reference and prediction in terms of characters and sentences (Figure 2). After not finding strong correlations among these parameters, we decide to split the response into two sub-tasks. Namely, we process system responses by separating them into two parts:

```
summary + [optional] question
```

These two parts are defined as follows:

- Summary: This part of the response focuses on summarising the selected knowledge items.

- Question: If included, this part of the response pertains to dialogue management, where the system can offer assistance in continuing with the current task, such as making a reservation.

3.2 Summary analysis

We enrich the dataset by adding two types of sentiment scores. Firstly, we calculate a sentiment score using the spaCy library[2] across all knowledge items selected for that dialogue. Secondly, we calculate a sentiment score specifically for the summary part of the ground truth responses. Our intuition is that the sentiment of the summary should align with the sentiment of the knowledge items used to generate the response. Our belief is supported by positive correlations observed in all data splits (train=0.41, val=0.53, test=0.54) and overall (all=0.45).
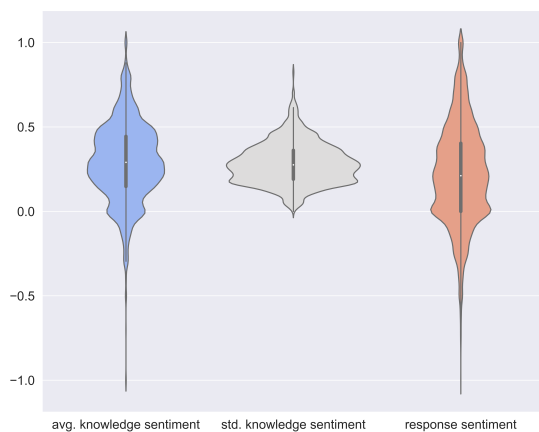

Figure 3: Sentiment distributions across dialogues for knowledge items (left) and responses (right). For knowledge items, we also show the distribution of standard deviation sentiment scores inside dialogues (centre).

Furthermore, we examined the data distributions of these new features, as depicted in Figure 3. The average sentiment of the knowledge items and the

sentiment of the responses have similar distributions. They both have positive medians, positive interquartile ranges, and a long tail extending into the negative range, indicating that the sentiment of the system's responses tends to be neutral to positive. The average sentiment of the knowledge items is slightly higher than the sentiment of the responses. The standard deviation of the spread of the knowledge sentiment scores inside dialogues is around 0.25, indicating that most knowledge items are not highly polarised, and their sentiment values are relatively close to each other.

3.3 Dialogue management analysis

Dialogue acts We generate a dialogue act tag MIDAS (Yu and Yu, 2019) for each of the user's last utterances in the dialogue context and explore the baseline performance per act type (Table 1). Our findings reveal that questions are quite frequent, as users often rely on them to interact with the artificial assistant in task-oriented dialogues. It is worth noting that *yes/no questions* are the most frequent type and typically elicit short system responses, consisting of simple answers like "yes" or "no." Furthermore, we observe that some questions seek opinions, which may require subjective knowledge derived from reviews for an appropriate response, in contrast to factual questions that can potentially be answered using FAQs.

Optional questions We examined how frequently questions appear in the system's responses. Our findings indicate that 34.94% of the responses contain a question at the end.
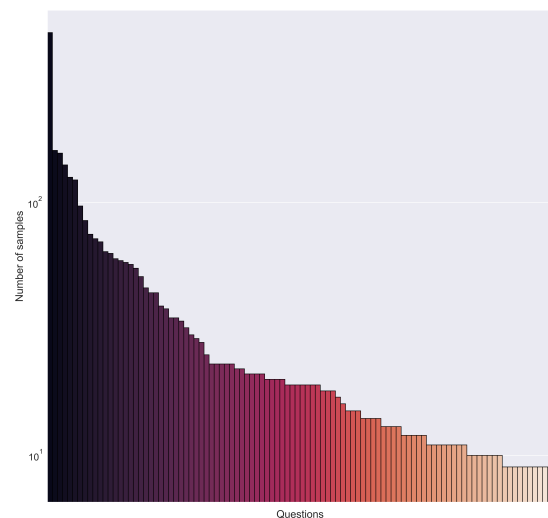

Figure 4: Question frequency distribution across all data splits. Only the first 100 elements are plotted.

---

[2]https://spacy.io/

195

Act	Freq.	Res. Len.	BLEU	METEOR	ROUGE-1	ROUGE-2	ROUGE-L
yes_no_question	1078	131.85	0.097	0.187	0.368	0.151	0.292
open_question_opinion	288	133.52	0.107	0.189	0.379	0.163	**0.301**
open_question_factual	203	134.51	0.094	0.179	0.358	0.148	0.285
statement	188	137.44	0.081	0.170	0.345	0.131	0.260
command	184	132.82	0.087	0.182	0.355	0.141	0.280
opinion	166	139.07	0.089	0.183	0.366	0.143	0.280
pos_answer	14	135.78	0.053	0.151	0.312	0.108	0.230
complaint	5	138.20	**0.110**	0.160	0.333	0.122	0.277
comment	1	181.00	0.038	0.084	0.196	0.081	0.117
neg_answer	1	**104.00**	0.069	**0.319**	**0.611**	**0.294**	0.5
nonsense	1	189.00	0.014	0.089	0.184	0.0	0.123

Table 1: Data statistics (frequency and average system response length) and baseline performance per dialogue act.

When analysing the questions, we observe that the data distribution exhibits a single peak and a long tail (Figure 4). Out of the 2,522 unique questions, 79% (n=1,994) appear only once. The most commonly occurring question is *"Would you like to know more about them?* accounting for 19% of the occurrences. Moreover, we notice that the most frequent questions are generic. The top five questions, which make up 42% of the data, include: *"Do you have any other questions?"*, *"Is there anything else I can help you with?"*, *"Would you like to make a reservation?"*, and *"Is there anything else you'd like to know about them?"*.

To evaluate the impact of optional questions, we conducted an ablation study where we remove the questions at the end of the responses (Table 2). This leads to slightly lower scores. Additionally, we experiment with adding the most frequent question (MFQ) to the responses that did not already contain a question. This improves the performance for longer responses.

4 Method

4.1 Sub-task 2: Knowledge Selection

Given the limited scope of reviews, traveller types and domains, we expand the knowledge base with few-shot data augmentation. Utilising ChatGPT, we synthesise artificial data, supplementing the MultiWoz 2.1 dataset's augmented version (Eric et al., 2020). This process involves creating new reviews for existing items within existing domains, in addition to the introduction of three new domains from the DSTC10 data set, all enriched with subjective knowledge. In total, our additional data contains 1,006 extra reviews. We combine these with the existing data for use in both the Knowledge

Seeking Turn Detection and Knowledge Selection tasks. For a data example see A.

Extension of item reviews The original knowledge data had 143 entities and 1,430 reviews (8,013 sentences) in the *hotel* and *restaurant* domains. We added 715 reviews to this by adding five different traveller types and producing five sentences for each type. This led to 25 extra sentences per entity, for a total of 2,145 reviews and 11,609 sentences. We significantly expanded and diversified the present traveller types. The original data had only 6 types, we increased this to 26, though some like *"Group of friends"* and *"Friends getaway"* could be merged into the same type. For a complete overview see Table 7. The average sentence length for the original reviews is 73.73 compared to 76.93 for the added reviews. The prompt to generate these reviews can be found in Prompt 4.1

Extension to new domains Inspired by Thulke et al. (2022), we expand our knowledge base by adding reviews from three different domains: *attraction, taxi,* and *train*.

Prompt 4.1: extension of new reviews

Please provide new reviews for the **{entity_name}** but for five different traveler_type.
Continue the counting of the reviews and make sure the new reviews are in a dict format like this:
"<id>": {{"traveler_type": "<traveler_type>", "sentences": {{"<id>": "<review>", "<id>": "<review>", "<id>": "<review>", "<id>": "<review>"}}}},

These are the existing reviews: **{reviews_all}**.

Take this start and continue. Use double quotes to comply with json format.

"10": {{"traveler_type":

Approach	BLEU	METEOR	ROUGE-1	ROUGE-2	ROUGE-L
Baseline	**0.102**	0.179	**0.365**	0.149	**0.287**
Strip questions	0.083	0.165	0.359	0.148	0.284
Append MFQ	0.094	**0.181**	0.363	**0.150**	0.286

Table 2: Results of investigation regarding questions on the validation set.

We use the validation data from DSTC10[3] which contains entities and respective FAQs from these different domains. The data did not contain subjective reviews for any of the domains. Hence, we generate three reviews for each entity by prompting ChatGPT with their entity name and FAQs. Due to the limited context width of the ChatGPT API, we restrict the reviews to two sentences.

We generate the reviews with Prompt 4.2, which resulted in 291 reviews for the *attraction* domain.

Prompt 4.2: extension to new domains

Given this example:

{ **<entity_id>**: { "name": **<entity_name>**, "faqs": { **<faq_id>**: { "question": **<question>**, "answer": **<answer>** } } } },

can you generate three more reviews, not more than 2 sentences, as: traveler type: review?

Although we prepared the data for the *taxi* and *train* reviews, we excluded these two domains as there were no transportation-related domains present in the released test set.

4.2 Sub-task 3: Response Generation

4.2.1 Model exploration

The generation baseline provided by the organisers uses BART (Lewis et al., 2020), specifically bart-base[4], fine-tuned for ten epochs using Adam with a learning rate of 3e-5 and epsilon of 1e-8.

Instead of sticking to the base BART model, we opted to browse publicly available trained models on different tasks and investigate their performance after fine-tuning on the DSTC data. This approach is inspired by *model recycling* (Choshen et al., 2022), which suggests that there exist models that are generally better at being adapted to perform domain-specific tasks. We browsed the HuggingFace Hub[5] for models with a better fit (in

the task they were trained for or in the data they used for pre-training). Ultimately, we retrained the baseline architecture, as well as Flan-T5 (Chung et al., 2022) in its small, base and large variants, and a trained BART model, trained on the SAM-Sum dataset (Gliwa et al., 2019). See Table 3 for the results. For training the larger architectures, including flan-t5-large, bart-large and flan-t5-base, we modified the optimisation parameters (learning rate 4e-05, a warm-up ratio of 0.2). We found that using a larger model has small benefits over the original baseline model in most evaluation metrics. We did not find that training on intermediate tasks that are related to our objective improved performance substantially.

4.2.2 Prompting

We used OpenAI's APIs[6] to generate system responses given the baseline output for sub-task 1 and 2. To ensure consistent results, we set the temperature value to 0 for all configurations.

Model comparison To compare GPT-3 and ChatGPT, we conducted experiments using simple prompts while adhering to the respective API requirements. The prompts consisted of two components: a) the dialogue history and b) the selected subjective knowledge. The subjective knowledge was formatted as a list, explicitly specifying whether each item was a **Review** or an **FAQ**.

```
{knowledge_type}: {text}
```

The initial prompt for GPT-3 is shown in Prompt 4.3

Prompt 4.3: GPT-3, initial prompt

DIALOGUE:
{dialogue_context}

KNOWLEDGE:
{knowledge}

RESPONSE:

Model name	BLEU	METEOR	ROUGE-1	ROUGE-2	ROUGE-L
bart-large-cnn-samsum	0.095	**0.184**	**0.363**	0.144	0.278
bart-base	0.096	0.173	0.347	0.139	0.271
flan-t5-base	0.096	0.173	0.345	0.140	0.274
flan-t5-large	**0.104**	0.179	0.360	**0.147**	**0.281**
flan-t5-small	0.098	0.164	0.337	0.136	0.272

Table 3: Models explored, shown performance is on the test set.

Model name	BLEU	METEOR	ROUGE-1	ROUGE-2	ROUGE-L
ChatGPT	0.041	0.147	0.280	0.085	0.207
GPT-3	0.051	0.137	0.302	0.105	0.237
ChatGPT + CoT	0.050	0.155	0.305	0.099	0.228
ChatGPT + CoT + 3 examples	**0.069**	**0.161**	**0.325**	**0.117**	**0.243**
Waterfall	0.056	0.152	0.312	0.100	0.236
Waterfall + 3 examples	0.057	0.160	0.311	0.103	0.225

Table 4: Prompting settings explored, shown performance is on the validation set.

ChatGPT requires API requests to be given as a series of messages with a "role" that can have three values: *System*, *User* or *Assistant*. As they explain in their documentation, *System* messages can be used to give instructions to the model, while *User* and *Assistant* messages can be used to carry on a dialogue with the model. As such, we use the *System* messages to give the subjective knowledge to the model and then send each of the dialogue history messages as either *User* or *Assistant* messages. The initial prompt for ChatGPT is shown in Prompt 4.4.

Prompt 4.4: ChatGPT, initial prompt

```
{"role": "system",
"content": "You are a helpful assistant with access to
the following:
{knowledge}" },
{"role": "user",
"content": "{dialogue_context}" },
{"role": "assistant",
"content": "{dialogue_context}" },
{"role": "user",
"content": "{dialogue_context}" }
```

Although ChatGPT is optimised for dialogue and we expected it to perform better, it scored lower than GPT-3 in the automatic metric evaluation, as seen in Table 4. Upon further examination of the generated responses, we observed that ChatGPT tends to produce much longer responses. In fact, 367 API requests were truncated due to their length. On the other hand, GPT-3 generates more concise

responses compared to both the baseline and the ground truth (Table 5). This difference in response length has a negative impact on most natural language generation metrics, as they tend to penalise longer and wordier responses.

Approach	#Sentences		#Characters	
	Avg.	Std.	Avg.	Std.
GPT-3	1.55	0.77	120.04	70.16
Ground truth (val)	1.69	0.66	133.55	32.05
Ground truth (test)	1.70	0.66	135.55	32.12
Ground truth (train)	1.71	0.65	136.61	33.49
Baseline (val)	1.91	0.53	129.66	29.07
ChatGPT	2.09	0.79	182.62	74.61

Table 5: Length statistics on responses, ordered by ascending average length

Chain Of Thought We focus on optimising the prompts for the best performance of a single model, specifically ChatGPT. To achieve this, we adopt recommended techniques from the OpenAI Cookbook[7] for prompting ChatGPT, such as Chain of Thought (CoT) (Wei et al., 2022). For this, we divide the task into two sub-tasks: summarisation and dialogue management (See Section 3). The resulting prompt is shown in Prompt 4.5.

To accommodate CoT answers, which tend to be longer, we double the value of the "max_tokens"

---

[7]https://github.com/openai/openai-cookbook/blob/main/techniques_to_improve_reliability.md

parameter. This modification leads to improved performance (refer to Table 4); however, it also increases the number of truncated API responses to 1,110.

Prompt 4.5: ChatGPT, Chain Of Thought

You are assisting a user. Create a response for the user, using the following procedure:
(1) First, summarise the available knowledge into a couple sentences.
(2) Then, create a short follow-up question given the dialogue history.
(3) Create the final response to the user as <summary><follow-up>

Knowledge:
    {knowledge}

Dialogue history:
    {dialogue_context}

Solution:
    (1) summary:

Few-shot learning   In order to further enhance the prompt, we incorporate few-shot learning (Brown et al., 2020). After initial exploration, we select three examples that pose the most difficulty. To identify these examples, we analyse instances with the lowest performance and examine their data trends. We observe that ChatGPT struggles with dialogues that involve a high number of items and have a predominantly negative sentiment. Consequently, we select dialogues from the training dataset that meet the following criteria:

1. Filter by the number of knowledge items that are "higher than usual", between [mean+std, mean+2std]

2. Filter by the average sentiment that is "lower than usual", between [mean-2std, mean-std]

3. Since around one-third of responses contain a question, we sample accordingly

By applying these selection criteria, we focus on dialogues that require better summarisation due to the increased number of knowledge items with slightly negative sentiment, and a mix of dialogues with and without a question. This setting yields the best performance and reduces the number of truncated API responses to 380.

Waterfall prompting   Additionally, we explore combining prompts and models by utilising the output of one prompted model as part of the input prompt for a subsequent model, in a waterfall

manner. We use GPT-3 for summarisation since it behaves better at this part of the task (Bhaskar et al., 2023). We also separate the FAQs, containing factual information, from the reviews, which have opinionated information. The prompt we use for this task is shown in Prompt 4.6.

Prompt 4.6: GPT-3, Summarisation

Summarize the following into one or two sentences max:

FAQs:
    {faqs}

Reviews:
    {reviews}

We then use ChatGPT to generate the system response in dialogue, using the output of GPT-3 as part of the prompt. The final prompt is shown in Prompt 4.7. Interestingly, we observed a decline in performance when using this waterfall approach compared to a single model setting.

Prompt 4.7: ChatGPT, Final prompt

You are assisting a user. Create a response for the user, using the following procedure:
(1) First, summarise the available knowledge into a couple sentences.
(2) Then, create a short follow-up question given the dialogue history.
(3) Create a final brief response to the user as <summary><follow-up>

Knowledge:
    {knowledge}

Dialogue history:
    {dialogue_context}

Solution:
    (1) summary:
        {GPT-3 summary}
    (2) follow-up:

---

5   Results

Table 6 shows the submitted approaches' results for all three sub-tasks. We ranked as the 8th best team; hence we did not get official human evaluation results on our submissions. Two of our approaches perform better than the baseline on automatic evaluation metrics, and our identifier in the official results is Team 10.

In the sub-task of **Knowledge Seeking Turn Detection**, our added knowledge induced a slight decrement in performance, with precision declining

from 0.998 to 0.995 and F1 score from 0.998 to 0.997, albeit recall experienced a marginal increase (refer to Table 6 for further details).

Conversely, **Knowledge Selection** saw an improvement with the augmentation, resulting in a small uplift across all metrics: precision from 0.790 to 0.796, recall from 0.788 to 0.794, F1 from 0.789 to 0.795, and Exact Match from 0.91 to 0.418.

We submitted three different approaches for the **Response Generation** task. (1) Our first submission is based on the model exploration mentioned in 4.2.1. After fine-tuning flan-t5-large this approach outperforms the baseline on all metrics, leading to a mean reciprocal rank (MRR) of 0.068 compared to 0.039 of the baseline. On BLEU this approach ranked 7th out of 49 submissions. (2) Based on our data analysis (3.3), we extended this approach in post-processing with the most frequently appearing question from the training set appended to the end of the review summaries. This approach, although underperforming relative to our first except in terms of METEOR, still surpassed the baseline on all metrics, apart from BLEU, with an MRR of 0.045. (3) Despite lower performance relative to the baseline in quantitative metrics, our waterfall prompting approach involving a combination of ChatGPT and GPT-3 gave qualitatively promising results. Presented with non-matching knowledge items it did not generate a summary, but flagged them as not relevant to the questions at hand. A full human evaluation would be an interesting step to investigate the perceived accuracy and appropriateness of such responses.

5.1 Submissions

CLTeamL-0 Baseline with augmented data for Knowledge-seeking Turn Detection and Knowledge Selection, flan-t5-large for Response Generation

CLTeamL-1 Baseline with augmented data for Knowledge-seeking Turn Detection and Knowledge Selection, flan-t5-large for Response Generation. Additionally, we do post-processing on the output, adding the most frequent question from the training set as a follow-up question to each review summary.

CLTeamL-2 Baseline for Knowledge-seeking Turn Detection and Knowledge Selection. For the Response Generation, we prompted ChatGPT with Chain Of Thought instructions, to first generate a summary of the knowledge and then append a follow-up question given the dialogue history. We provided three examples, selected from the most challenging cases in the training data according to the data trends on the validation set. Unfortunately, due to server problems, the data samples after 4500, are lacking a review summary. A full run (see Approach 2b in Table 6) after submission showed slight improvements across all metrics but still did not outperform the baseline or our other approaches.

6 Conclusion

In conclusion, our work on the Task-Oriented Conversational Modeling with Subjective Knowledge task produced three key contributions: Firstly, our detailed data analysis could serve as a basis for future dataset scoping within this domain. Secondly, we significantly expanded the knowledge dataset size utilising few-shot data augmentation. Lastly, our most successful model was a fusion of the baseline with augmented data and flan-t5-large. Our waterfall prompting approach incorporating a blend of Large Language Models demonstrated lower metrics compared to the baseline, but upon a qualitative assessment, the results were deemed satisfactory, albeit the absence of official human evaluation impedes a definitive judgement.

A preliminary qualitative analysis shows that ChatGPT can spot mistakes during knowledge selection. Thus, future work could explore incorporating an initial step where ChatGPT is employed to evaluate the relevance of the selected knowledge items. Additionally, we plan to experiment with summarising positive and negative reviews separately before creating a consolidated summary. This method may enhance performance when dealing with polarised reviews.

Limitations

The APIs provided by OpenAI come with particular length parameters which pose certain constraints in terms of flexibility and generalisability to larger datasets and longer sequences which are prevalent in knowledge-grounded dialogue.

There's also a financial aspect to consider, as the use of these APIs involves a monetary cost. We chose to proceed with ChatGPT instead of GPT-3 for parts of our research, even though preliminary explorations indicated higher performance by the latter, due to the additional cost associated with the use of GPT-3. Hence, there might be potential

Approach	sub-task 1			sub-task 2				sub-task 3					Total
	P	R	F1	P	R	F1	EM	BLEU	METEOR	ROUGE-1	ROUGE-2	ROUGE-L	MRR
Baseline	0.998	0.998	0.998	0.790	0.788	0.789	0.391	0.100	0.175	0.352	0.143	0.275	0.039
CLTeamL-0	0.995	0.999	0.997	0.796	0.794	0.795	0.418	0.104	0.179	0.360	0.147	0.281	0.068
CLTeamL-1	0.995	0.999	0.997	0.796	0.794	0.795	0.418	0.094	0.181	0.356	0.146	0.279	0.045
CLTeamL-2	0.996	0.998	0.997	0.772	0.818	0.794	0.423	0.060	0.121	0.241	0.087	0.186	0.025
CLTeamL-2b	0.996	0.998	0.997	0.772	0.818	0.794	0.423	0.068	0.157	0.311	0.112	0.232	NA

Table 6: Results over the test set. The final ranking is determined by the mean reciprocal rank (MRR) of all scores.

performance enhancements our approaches were not able to capture due to this cost-based decision.

Ethics Statement

We artificially increased the size of our knowledge data set with few-shot data augmentation to cover a broader range of traveller types and domains. The synthetic reviews were created using GPT-3 and ChatGPT. There exists a potential for the introduction of biases in this newly generated data due to inherent model biases that might have been propagated into the reviews. This is especially concerning for commercial LLMs, like the ones used, since their training data is not made public.

Acknowledgements

References

Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. 2020. Do not have enough data? deep learning to the rescue! *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7383–7390.

Adithya Bhaskar, Alexander R. Fabbri, and Greg Durrett. 2023. Prompted opinion summarization with gpt-3.5.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Leshem Choshen, Elad Venezian, Shachar Don-Yehia, Noam Slonim, and Yoav Katz. 2022. Where to start? analyzing the potential value of intermediate models. *arXiv preprint arXiv:2211.00107*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj

Goyal, Peter Ku, and Dilek Hakkani-Tur. 2020. MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 422–428, Marseille, France. European Language Resources Association.

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.

Seokhwan Kim, Mihail Eric, Karthik Gopalakrishnan, Behnam Hedayatnia, Yang Liu, and Dilek Hakkani-Tur. 2020. Beyond domain APIs: Task-oriented conversational modeling with unstructured knowledge access. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 278–289, 1st virtual meeting. Association for Computational Linguistics.

Seokhwan Kim, Yang Liu, Di Jin, Alexandros Papangelis, Behnam Hedayatnia, Karthik Gopalakrishnan, and Dilek Hakkani-Tür. 2022. Knowledge-grounded task-oriented dialogue modeling on spoken conversations track at dstc10.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. In *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, pages 18–26, Suzhou, China. Association for Computational Linguistics.

Kenton Lee, Kelvin Guu, Luheng He, Tim Dozat, and Hyung Won Chung. 2021. Neural data augmentation via example extrapolation. *arXiv preprint arXiv:2102.01335*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Silviu Pitis, Michael R Zhang, Andrew Wang, and Jimmy Ba. 2023. Boosted prompt ensembles for large language models. *arXiv preprint arXiv:2304.05970*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. 2023. Towards expert-level medical question answering with large language models.

Zhiqing Sun, Xuezhi Wang, Yi Tay, Yiming Yang, and Denny Zhou. 2023. Recitation-augmented language models.

David Thulke, Nico Daheim, Christian Dugast, and Hermann Ney. 2022. Adapting Document-Grounded Dialog Systems to Spoken Conversations using Data Augmentation and a Noisy Channel Model. In *Workshop on DSTC10, AAAI*.

Xin Tian, Xinxian Huang, Dongfeng He, Yingzhan Lin, Siqi Bao, Huang He, Liankai Huang, Qiang Ju, Xiyuan Zhang, Jian Xie, Shuqi Sun, Fan Wang, Hua Wu, and Haifeng Wang. 2021. Tod-da: Towards boosting the robustness of task-oriented dialogue modeling on spoken conversations. *arXiv preprint arXiv:2112.12441*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. 2022. Rationale-augmented ensembles in language models. *arXiv preprint arXiv:2207.00747*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Congying Xia, Caiming Xiong, Philip Yu, and Richard Socher. 2020. Composed variational natural language generation for few-shot intents. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3379–3388, Online. Association for Computational Linguistics.

Dian Yu and Zhou Yu. 2019. MIDAS: A dialog act annotation scheme for open domain human machine spoken conversations. *CoRR*, abs/1908.10023.

Chao Zhao, Spandana Gella, Seokhwan Kim, Di Jin, Devamanyu Hazarika, Alexandros Papangelis, Behnam Hedayatnia, Mahdi Namazifar, Yang Liu, and Dilek Hakkani-Tur. 2023. " what do others think?": Task-oriented conversational modeling with subjective knowledge. *arXiv preprint arXiv:2305.12091*.

A Appendix

```
{
  "hotel": {
    "0": {
      "name": "A AND B GUEST HOUSE",
      "reviews": {
        "0": {
          "traveler_type": "Solo travelers",
          "sentences": {
            "0": "I was really happy with my recent stay at A and B Guest House.",
            "1": "I stayed on my own, and I'm a smoker, so I was super happy that there was a designated area especially for smokers.",
            "2": "I also thought that my room was very spacious, and I was pleased with the breakfast options that were available."
          }
        },
        "1": {
          "traveler_type": "Couples",
          "sentences": {
            "0": "My husband was pleased to be able to park on site for free.",
            "1": "We thought it was a bit noisy at A and B especially because it was just us and we had looked forward to quiet."
          }
        },
        "2": {
          "traveler_type": "Business travelers",
          "sentences": {
            "0": "I wasn't thrilled the with lack of conveniences at A and B Guest House on my recent business trip.",
            "1": "No luggage storage service was inconvenient, and I was surprised there was no safety deposit box in a room.",
            "2": "I do appreciate the good breakfast options and the comfort of the bed, though.",
            "3": "So overall, I would probably stay there again."
          }
        },
        "3": {
          "traveler_type": "Solo travelers",
          "sentences": {
            "0": "I stayed at this guesthouse by myself.",
            "1": "I was happy with the laundry facilities that were on site.",
            "2": "The free available Internet provided great speeds and a strong connection.",
            "3": "I was treated to a lovely filling breakfast in the morning.",
            "4": "However, when I went to use the bathroom I found that the bathroom was not clean.",
            "5": "Also, when I went to put away my luggage I found that there were no provided spaces in which to do so."
          }
        },
        "4": {
          "traveler_type": "Solo travelers",
          "sentences": {
            "0": "Not long ago, I stayed at A and B Guest House by myself.",
            "1": "The staff was really friendly when I checked in, and I was thrilled with how comfy the bed was.",
            "2": "I got a great night of sleep!",
            "3": "But I was sad that there was no elevator so I had to use the stairs, and overall I think the room was overpriced."
          }
        },
        "5": {
          "traveler_type": "Solo travelers",
          "sentences": {
            "0": "I stayed here by myself, and it was nice to be able to wash my clothes at the laundry.",
            "1": "Looking out of the window was nice, as the view was unique.",
            "2": "The location makes it a delight to get around on foot, and see things.",
            "3": "If only the room was a bit cleaner, it would have been perfect."
          }
        },
        "6": {
          "traveler_type": "Solo travelers",
          "sentences": {
            "0": "The A and B Guest House is a great place to stay.",
            "1": "It is a beautiful place in a good location.",
            "2": "Very quiet.",
            "3": "The staff went out of their way to make sure that our stay was outstanding.",
            "4": "The clerk even sent a hand drawn card to us when he found it was our anniversary.",
            "5": "I don't remember when I've slept in a bed more comfortable than the one here.",
            "6": "Fair warning if you travel with valuables; there is no safety box in the room.",
            "7": "Be careful."
          }
        },
        "7": {
          "traveler_type": "Families",
          "sentences": {
            "0": "My family and I decided to stay here while we were in town.",
            "1": "It was kid friendly, so my kids weren't treated like garbage.",
            "2": "It was also in a good location with access to most important things, and parking was completely free.",
            "3": "The staff, on the other hand, were really unprofessional and laid back.",
            "4": "They were not professional at all."
          }
        },
        "8": {
          "traveler_type": "Solo travelers",
          "sentences": {
```

          "0": "The room was very large and spacious.",
          "1": "Made it very easy to settle in and relax after a long trip, and even more relaxing with the excellent view out of my window!",
          "2": "The room included free Wi-Fi which was very much appreciated because its quite uncommon to get free internet service these days!",
          "3": "My only complaint was that there was no concierge service and I had to take care of this myself - a bit annoying but overall a great stay."
        }
      },
      "9": {
        "traveler_type": "Couples",
        "sentences": {
          "0": "The A and B could be a problem if you need an accessible room you may need to book elsewhere.",
          "1": "The room was not very clean but it was large.",
          "2": "My boyfriend and I both thought the bed had a good firmness for our liking.",
          "3": "There is a great view especially at sunset."
        }
      },
      "10": {
        "traveler_type": "Family travelers",
        "sentences": {
          "0": "We had a great time staying at A and B Guest House with our kids.",
          "1": "The staff was very accommodating and provided us with a baby crib and high chair upon request.",
          "2": "The breakfast options were great and the location was convenient for our family activities.",
          "3": "The only downside was that the room was a bit cramped for our family of four."
        }
      },
      "11": {
        "traveler_type": "Group travelers",
        "sentences": {
          "0": "Our group of friends stayed at A and B Guest House and we had a great time.",
          "1": "The staff was friendly and helpful, and the location was perfect for exploring the city.",
          "2": "The rooms were clean and comfortable, and the breakfast was delicious.",
          "3": "The only issue we had was that the Wi-Fi was a bit spotty in some areas of the guest house."
        }
      },
      "12": {
        "traveler_type": "Budget travelers",
        "sentences": {
          "0": "A and B Guest House was a great choice for our budget-friendly trip.",
          "1": "The room was clean and comfortable, and the breakfast options were good.",
          "2": "The location was also convenient for exploring the city on foot.",
          "3": "The only downside was that the room was a bit small and there was no elevator."
        }
      },
      "13": {
        "traveler_type": "Luxury travelers",
        "sentences": {
          "0": "As a luxury traveler, I was disappointed with my stay at A and B Guest House.",
          "1": "The room was small and lacked the amenities I expect from a luxury hotel.",
          "2": "The staff was friendly but not particularly attentive to my needs.",
          "3": "The only positive was the location, which was convenient for exploring the city."
        }
      },
      "14": {
        "traveler_type": "Pet owners",
        "sentences": {
          "0": "We were happy to find a pet-friendly option in A and B Guest House.",
          "1": "The staff was accommodating and provided us with a pet bed and bowls upon request.",
          "2": "The location was also convenient for walking our dog.",
          "3": "The only downside was that there was no designated pet area for our dog to use the bathroom."
        }
      }
    }
  },
  "faqs": {
    "0": {
      "question": "Are children welcomed at this location?",
      "answer": "Yes, you can stay with children at A and B Guest House."
    },
    "1": {
      "question": "Can I bring my pet to A and B Guest House?",
      "answer": "No, pets are not allowed at this property."
    },
    "2": {
      "question": "Do you have onsite parking for your guests?",
      "answer": "There is onsite parking at A and B Guest House but it costs extra."
    },
    "3": {
      "question": "What time is check-in there?",
      "answer": "Check-in time is from 3:30pm - 9:00pm."
    },
    "4": {
      "question": "Is smoking allowed on the property?",
      "answer": "There are designated smoking areas throughout"
    },
    "5": {
      "question": "What languages are spoken?",

```
      "answer": "English,  Italian ,  Lithuanian ,  Portuguese ,  and  Russian  are  spoken  here ."
    },
    "6": {
      "question": "Should  I  make  a  reservation   for  parking?",
      "answer": "You need to  make  a  reservation   at  A and B Guest House for  parking ."
    },
    "7": {
      "question": "Are  children  allowed  to  check  in  here?",
      "answer": "An  individual  has  to  be  18 and over  to  check  in  at  A and B Guest House."
    },
    "8": {
      "question": "what time  do  I  check  out?",
      "answer": "Check out  times  range  from 7:30 AM to 10:00 AM."
    },
    "9": {
      "question": "Can my small dog stay  with  me?",
      "answer": "Pets  are  not  allowed  at  the  A and B Guest House."
    }
  }
},
```

Knowledge data example: Slightly shortened example from the knowledge.json. Reviews 0-9 are original, and reviews 10-14 are augmented.

Traveler Type	Count Original Reviews	Count Augmented Reviews
Families	320	95
Solo travelers	315	101
Couples	279	109
Colleagues	230	11
Friends	210	84
Business travelers	76	111
Budget travelers	0	49
Luxury travelers	0	39
Families with children	0	22
Foodies	0	19
Vegetarians	0	10
Pet owners	0	9
Business	0	6
Groups	0	6
Backpackers	0	5
Friends getaway	0	5
Families with teenagers	0	3
Adventure travelers	0	3
Retirees	0	3
Group of friends	0	2
Families with young children	0	2
Honeymooners	0	1
Senior travelers	0	1
Romantic couples	0	1
Romantic travelers	0	1
Students	0	1

Table 7: Comparison of traveler types between the original reviews and the added reviews. Some extremely similar types were combined into one e.g. "Family with children" and "Family with kids".