Dialogue Ontology Relation Extraction via Constrained Chain-of-Thought Decoding

Renato Vukovic, David Arps, Carel van Niekerk, Benjamin Matthias Ruppik,
Hsien-Chin Lin, Michael Heck, Milica Gašić
Heinrich Heine University Düsseldorf
{renato.vukovic, david.arps, niekerk, ruppik, linh, heckmi, gasic}@hhu.de

Abstract

State-of-the-art task-oriented dialogue systems typically rely on task-specific ontologies for fulfilling user queries. The majority of task-oriented dialogue data, such as customer service recordings, comes without ontology and annotation. Such ontologies are normally built manually, limiting the application of specialised systems. Dialogue ontology construction is an approach for automating that process and typically consists of two steps: term extraction and relation extraction. In this work, we focus on relation extraction in a transfer learning set-up. To improve the generalisation, we propose an extension to the decoding mechanism of large language models. We adapt Chain-of-Thought (CoT) decoding, recently developed for reasoning problems, to generative relation extraction. Here, we generate multiple branches in the decoding space and select the relations based on a confidence threshold. By constraining the decoding to ontology terms and relations, we aim to decrease the risk of hallucination. We conduct extensive experimentation on two widely used datasets and find improvements in performance on target ontology for source fine-tuned and one-shot prompted large language models.[1]

1 Introduction

State-of-the-art task-oriented dialogue (TOD) systems still rely on a fixed ontology to model their scope (Nguyen et al., 2023; Hudeček and Dusek, 2023). A *TOD ontology* comprises three levels of hierarchy: domains, slots and values. *Domains* are general topics of interest, *slots* are types of information about entities in a domain, and *values* are concrete instantiations of slots. Ontology thus forms a hierarchy: it is a directed graph where slots belong to domains and values in turn belong to slots. Note that slots can be shared across domains,

and so can values. An ontology is typically a prerequisite for generating API calls that access the underlying databases for entity retrieval. Further, the ontology defines the dialogue state, which is tracked by the system to determine the next actions given the evolving discourse.

The dependency on an ontology poses a significant challenge in transferring existing TOD systems to new domains and use cases. Although ontology-agnostic approaches do exist, their transfer capabilities are limited and their performance remains sub-par on novel data (Heck et al., 2022).

Large quantities of domain-specific TOD data, e.g. customer service recordings, are frequently available, but tend to come without annotation, rendering direct use for system development difficult (Brusco and Gravano, 2023). Manual labelling is error-prone, does not scale well and quickly becomes prohibitively expensive (Eric et al., 2020; Rosenbaum et al., 2022; Gung et al., 2023). Despite topical or domain mismatch, existing annotated datasets may provide information about TOD that can be leveraged to harness new data. For this reason, we are interested in utilising existing labelled TOD datasets to automatically generate a full ontology for new, yet-unlabelled, data.

Automatic dialogue *ontology construction* typically consists of two steps, dialogue term extraction (Vukovic et al., 2022) and hierarchy establishment. Although hierarchy establishment is often done via clustering (Hudeček et al., 2021; Yu et al., 2022) we approach it via *relation extraction* (RE), which is more similar to common information extraction pipelines (Genest et al., 2022; Xu et al., 2023). We call this task *dialogue ontology relation extraction* (DORE). A hierarchy is established by inferring in which level extracted terms lie, and by connecting terms across levels.

Although large language models (LLMs) have demonstrated considerable task transfer abilities (Brown et al., 2020; Ouyang et al., 2022),

---

[1] The code is avaibable under https://gitlab.cs.uni-duesseldorf.de/general/dsml/dialogue-ontology-relation-extraction-via-constrained-chain-of-thought-decoding
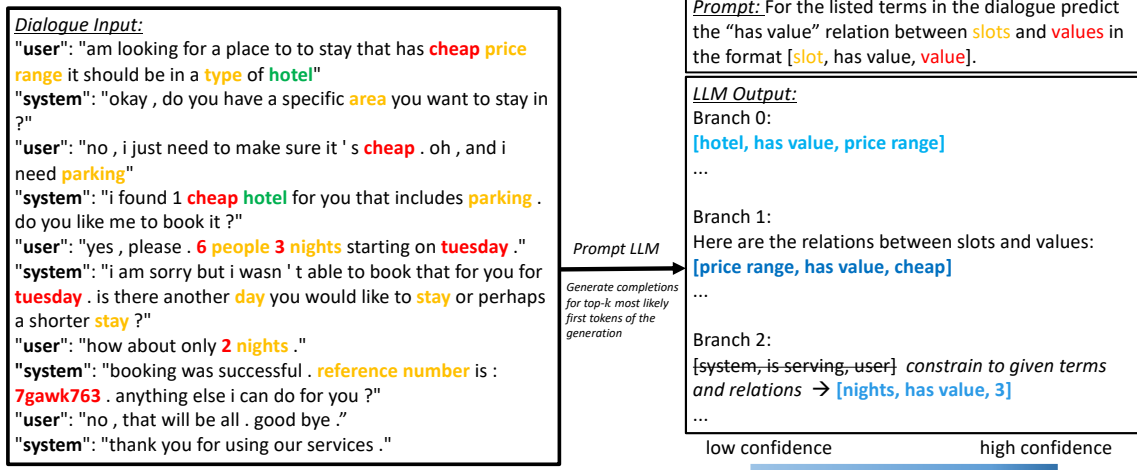
Figure 1: Example of constrained CoT-decoding for dialogue ontology extraction for a dialogue from MultiWOZ 2.1 (Eric et al., 2020). Domains are highlighted in green, slots in yellow and values in red. Branch 0 predicts an incorrect relation (*hotel* misclassified as slot) with lower confidence. Branch 1 has the highest confidence in the relation prediction, which is why it is selected as the final response. Also, it contains a form of reasoning that stresses the type of terms that are part of the relations to be predicted, i.e., slots and values. Branch 2 visualises constrained decoding, where the prediction of terms and relations is not possible if they are not present in the input.

they still lack behind specialised systems in TOD modelling when appropriate training data is available (Heck et al., 2023; Hudeček and Dusek, 2023).

In this work, we assume that some labelled out-of-domain source dialogue data is available to facilitate transfer learning. We examine two strategies of providing source data to an instruction-tuned LLM; 1) as one-shot examples in the prompt, and 2) as data for an additional round of supervised fine-tuning. We establish a challenging transfer setup by conducting experiments on two well-established medium to large scale multi-domain task-oriented dialogue benchmark datasets: MultiWOZ 2.1 (Budzianowski et al., 2018; Eric et al., 2020) and the Schema-Guided Dialogue (SGD; Rastogi et al., 2020) dataset. Since our focus is solely on DORE, we assume that the results of the first step of ontology construction, namely term extraction, are provided.

We propose to improve the decoding mechanism of an LLM in order to better leverage task-specific knowledge. Concretely, we constrain the generation to terms and relation types given in the model input to force the model to consider terms from the target data and output the desired format. We further adapt *chain-of-thought (CoT) decoding* (Wang and Zhou, 2024), which was recently proposed for logical reasoning, for DORE. Traditionally, CoT methods prompt or train the model to generate reasoning paths before giving the final answer (Wei

et al., 2022; Kim et al., 2023). CoT-decoding on the other hand exploits the observation that the presence of CoT-paths is correlated with higher confidence in the predicted answer in logical reasoning. We extend CoT-decoding to DORE by selecting the final answer based on the confidence of predicted relations in multiple generated model answer branches. Our final proposal, *constrained CoT-decoding for dialogue ontology extraction*, is the combination of our CoT-decoding approach to RE with constrained decoding, see Fig. 1. Empirically, this new decoding mechanism significantly outperforms both source one-shot and source fine-tuned baselines on the target data. Our contributions are as follows:

- We propose to induce an ontological hierarchy by accumulating ontology relation predictions from the dialogues in a TOD dataset.

- To the best of our knowledge, we are the first to apply *CoT-decoding* to dialogue ontology relation extraction.

- We develop an extension, called *constrained CoT-decoding*, for multi-relation extraction from task-oriented dialogues.

- Constrained CoT-decoding significantly improves the quality of relation predictions on the target dataset for both source one-shot and source fine-tuned baselines.

371

2 Related Work

Dialogue Ontology Construction We divide dialogue ontology construction into term extraction and relation extraction. Vukovic et al. (2022) improve out-of-domain generalisation of a dialogue term extraction model by making use of topological properties of the language model embedding space. Nguyen et al. (2023) improve phrasal segmentation of ontology terms via language model probing and contrastive learning. Since we evaluate the hierarchy on a global level based on relations, our approach is not directly comparable to clustering-based approaches such as Hudeček et al. (2021); Yu et al. (2022). In contrast to these methods, we view hierarchy establishment in isolation as a relation extraction task.

Yu et al. (2020) present DialogRE, a popular dataset for RE in short chit-chat dialogues. Closest to our approach, Albalak et al. (2022) jointly optimise RE and explanation generation to improve performance with a model-agnostic framework. Xu and Chen (2023) propose a zero-shot approach for extracting trigger words for dialogue relation extraction on DialogRE. However, these works focus on chit-chat dialogues, which do not include ontology relations.

Relation Extraction with LLMs LLMs show promising transfer capabilities out of the box (Laskar et al., 2023). Direct application to our task however is not promising, as it has been shown that aligned LLMs such as ChatGPT (OpenAI, 2022) do not perform well on extracting multiple relations at once (Lilong et al., 2024). This shortcoming has been linked to the influence of pre-training data distribution on downstream task performance (McCoy et al., 2023). RE data in particular amounts to a mere 0.5% of instruction-tuning datasets, and is hardly utilised for model selection (Wang et al., 2022; Zhang et al., 2023).

Traditionally, RE is performed in a pairwise manner (Zhang et al., 2023), resulting in quadratic complexity given the number of terms. This becomes intractable for generative LLMs when querying the LLM separately for each pair of terms. Alternatively, one may extract all relations present in a given input with a single LLM query, as is common in multi-relation extraction tasks such as document-level RE. For example, Lilong et al. (2024) extract relations by either predicting relations directly, or first predicting possible head entities in a document. Zhang et al. (2023) align LLMs for zero-shot RE by transforming RE into a question answering (QA) task, which is more frequent in the instruction-tuning data.

Constrained Decoding Constrained decoding limits the tokens that can be generated. It is typically applied to LLMs to improve downstream task performance, reduce hallucination and ensure certain output formats. Bogoychev and Chen (2023) constrain decoding for translation to ensure that certain terminology is used. Roy et al. (2024) use constrained decoding with a lookahead heuristic to speed up adaptation of LLMs to plan generation according to a given API in TOD. We want to force the model to use its inherent task knowledge while transferring abilities to new data.

Chain-of-Thought Reasoning LLM performance on complex reasoning tasks improves when the model generates a chain of thought (CoT). Wei et al. (2022) include examples of multi-step reasoning in the prompt, and Kojima et al. (2022) prompt the model in a zero-shot fashion to "think step by step". Reasoning capabilities can be further enhanced via specific training on CoT-data (Chung et al., 2024), or via teaching the model to reason (Zelikman et al., 2022). In contrast to this, we focus on eliciting model-inherent reasoning capabilities, without the need for specific prompts or training. As described in Sec. 3.2, we leverage the fact that a top-k decoding beam usually contains a CoT (Wang and Zhou, 2024).

3 Constrained Chain-of-Thought Decoding for Ontology Relation Extraction

3.1 Problem Definition

Dialogue ontology relation extraction (DORE) aims at extracting all relations between different terms in a TOD dataset. As seen in Figure 1, for each dialogue paired with a list of ontology terms, the output is a set of relations similar to document-level relation extraction (Tan et al., 2022). However, we consider the joint relation prediction set accumulated from all dialogue-level predictions, rather than the dialogue-level performance. In the DORE task, the model receives as input a task-oriented dialogue D annotated with a list of ontology terms T present in this dialogue. The output are valid ontology relations $R_{D,T}$ between the terms, which includes predicting whether a term

Relation	Verbaliser	Example
Domain-Slot	[Domain, has slot, Slot]	[hotel, has slot, price range]
Slot-Value	[Slot, has value, Value]	[price range, has value, cheap]
Value-Domain	[Value, has domain, Domain]	[cheap, has domain, hotel]
Equivalence	[Term1, refers to same concept as, Term2]	[cheap, refers to same concept as, low budget]

Table 1: Hierarchical dialogue ontology relation task definition with examples.

is a *domain, slot*, or *value*. A relation is denoted by a relational triplet with a head term, the relation and a tail term. Finally, the predicted relations for each dialogue are unified to form the final ontology relation set.

We consider 4 types of relation between ontology terms: *domain-slot, slot-value, value-domain* and *equivalent term* relations (see Table 1 for examples). Here, all relations except the equivalence relation are directed relations with a head and a tail term. *Domains* are general topics, such as *hotel* or *restaurant, slots* are types of information for entities in a *domain*, such as *price range* or *area* and *values* are concrete instantiations of slots, such as "cheap" or "west". The equivalence relation connects terms from the same hierarchy level that point to the same ontological concept, e.g. "expensive" and "high-end" both represent a high price. In the prompt and labels, we denote the relation types through different verbalisers, shown in Table 1. Verbalisers are descriptions of task-specific labels in natural language. They align the task closer with the pre-training distribution of the LLM (Schick and Schütze, 2021; Mosbach et al., 2023).

Our hypothesis is that the general definitions of the ontology hierarchy relations enable seamless transfer to new data in order to construct a similarly structured ontology on the new data. Based on these relations, we focus on transferring the structural information about ontologies from a source dataset to a target dataset. Here, we consider a one-shot and a fine-tuning approach.

3.2 Chain-of-Thought Decoding

CoT reasoning in LLMs has demonstrated improved performance in various complex reasoning tasks (Sec. 2). The results of Wang and Zhou (2024) show that LLMs inherently possess reasoning capabilities, which can be elicited without explicit prompting through *Chain-of-Thought decoding*. Concretely, they experiment on pre-trained and instruction-tuned versions of PaLM 2 (Anil et al., 2023) and Mistral-7B (Jiang et al., 2023). They observe that although the greedily decoded

response might not always exhibit reasoning, one of the top-k beams usually contains a CoT. This CoT not only shows higher confidence in the answer, but also exhibits greater accuracy. They propose to consider the top-k probability tokens at the start of the predicted response. From there, k completions, called *branches*, are generated, resulting in k-times computational complexity during inference. The final response is chosen based on the confidence of the tokens that belong to the *answer* in each branch, i.e., the average confidence of the *answer tokens*. In logical reasoning, there is only one answer in each branch, which is a number. In that case, they identify the answer by prompting the model with "So the answer is:" at the end and match the following number to one in the preceding response. In our case, there are multiple answers per branch, which we identify based on the fact that relations are supposed to be predicted between brackets.

CoT-Decoding for DORE In this paper, we extend CoT decoding to handle the multi-answer scenario in the DORE task. We compute the confidence of answer tokens by utilising their structure, which, in our case, involves predicting relational triplets in the format $[headterm, relation, tailterm]$ and the notion of disparity. The disparity of a probability distribution is the difference between the probability of the most likely outcome and the next most likely outcome. The confidence for each answer token for a given branch is measured by the average disparity of its tokens. Formally this is given by

$$\Delta_{i,a} = \frac{1}{n} \sum_{x_t \in a} p(x_t^{\text{top}} \mid x_{<t}) - p(x_t^{\text{next}} \mid x_{<t}), \quad (1)$$

where a is an answer (in our case the triplet), i is a branch, x_t are the answer tokens belonging to the answer in branch i, x_t^{top} is the most likely token on position t and x_t^{next} the next most likely token on position t. $x_{<t}$ are the tokens in branch i on positions preceding t, i.e. the context so far.

In DORE, answer tokens are those that form terms and relations in the predicted relational

373

triplets, which means there are three disparities per relation. This approach relies on detecting answer tokens in a generated response for confidence estimation, and we leave an extension to arbitrary answer structures to future work. The resulting triplet disparities are denoted as $\Delta_a = [\Delta_h, \Delta_r, \Delta_t]$. We explored mean, median, maximum, and minimum as aggregation strategies for relational triplet mentions, finding that all of them lead to similar results. For simplicity, we choose the mean to aggregate the disparity for a relational triplet in branch i, i.e. $\Delta_{i,a} = \frac{1}{3}(\Delta_{h,i} + \Delta_{r,i} + \Delta_{t,i})$.

We select the branch with the highest average disparity over the relations predicted in each branch to get the final set of relation predictions for a dialogue. The average disparity for branch i is given by

$$\overline{\Delta}_i = \frac{1}{n_{a,i}} \sum_{a \in R_i} \Delta_{i,a} \qquad (2)$$

where a is a relational triplet, R_i is the set of relations and $n_{a,i}$ is the number of relations in branch i. The final set of predicted relations is then given by

$$R_{\overline{\Delta}_{\max}} = \{R_i \mid i = \mathrm{argmax}\{\overline{\Delta}_0, \dots, \overline{\Delta}_k\}\} \quad (3)$$

We also experiment with a confidence threshold based approach for relation selection. Here, the average disparity of a relation is computed across occurrences in different branches:

$$\widetilde{\Delta}_a = \frac{1}{n_a} \sum_{i \in \{1, \dots, k\}} \Delta_{i,a}, \qquad (4)$$

where $\Delta_{i,a}$ is the disparity of the answer a in the i-th branch and n_a is the number of occurrences of a across the different branches. The final set of predicted relations $R_{\Delta > \Delta_{\mathrm{threshold}}}$ is then

$$R_{\Delta > \Delta_{\mathrm{threshold}}} = \{a \mid \widetilde{\Delta}_a > \Delta_{\mathrm{threshold}}\} \quad (5)$$

3.3 Constrained Decoding

We constrain the generation of the relation terms and relation types if the beginning of a relational triplet is predicted to ensure the structure and mitigate term and relation hallucination (see Figure 1). This means for a relational triplet, $[h, r, t]$, we ensure that $h, t \in T$ and $r \in R$, where T is the set of terms for the current dialogue and R is the set of relation types given in the prompt. Note that we only constrain the generation when an opening bracket is predicted by the model, and resume to non-constrained generation after the generated relational triplet.

4 Experiments

4.1 Experimental Setup

We utilise the open-source Gemma 2B (Mesnard et al., 2024) instruction-tuned model with context size of 4096 for all experiments. In CoT-decoding we set $k = 5$. For a more thorough analysis of the impact of k in CoT decoding, resort to Wang and Zhou (2024). We always branch at the first token; branching at later tokens did not show improvements. For all CoT-decoding experiments, we select the relations from the branch with the highest disparity, as the threshold based method works worse and also adds a new hyperparameter. In the one-shot prompts, we use a combination of an instruction with simple natural language with a preceding example (Brown et al., 2020; Sahoo et al., 2024). For fine-tuning, we remove the example from the prompt.

Datasets For the source dataset, we employ the MultiWOZ 2.1 dataset (Eric et al., 2020). It has 7 domains and over 10,000 dialogues. We use the training set for training and select from it one random dialogue with relation annotation as one-shot exemplar. The target dataset is the schema-guided dialogue (SGD) dataset (Rastogi et al., 2020). It comprises more than 20,000 dialogues and 20 domains. We use the SGD test split for evaluation in the main results, which contains 4,201 dialogues and 18 domains. In the test set, there are 134 domain-slot relations, 6,162 slot-value relations, 8,233 value-domain relations and 330 equivalence relations. It is worth noting the SGD test set contains dialogues from different domains than the SGD training set, as well as a significant amount of unseen ontology relations. We use ConvLab-3 (Zhu et al., 2023b) for loading all the datasets.

Training For both fine-tuning and one-shot prompting, we utilise the original Gemma prompt template (Mesnard et al., 2024). For training, we utilise Low-rank adaptation (LoRA, Hu et al., 2022) with the default parameters in the peft library (Mangrulkar et al., 2022). We train the model on a single NVIDIA RTX8000 GPU and do inference with all models on one NVIDIA RTX6000 GPU.

We only consider a one-shot approach due to context size constraints, as the relational triplets in the exemplars contain brackets. Brackets are considered individual tokens, increasing the number of tokens significantly. Because of this a maximum of three exemplars fits in the context size,

Approach	F1-Score	Precision	Recall
One-shot example from MultiWOZ			
Baseline: Separate relation prediction	7.4	8.8	6.4
+ constrained decoding	**8.5***	5.7	**17.3*‡**
+ CoT decoding	**9.2***	8.8	**9.6***
+ constrained CoT decoding	**9.2***	6.4	**15.9***
Fine-tuning on MultiWOZ			
Baseline: Fine-tuning on MultiWOZ	10.9	6.8	28.8
+ constrained decoding	**12.0†**	7.4	32.3‡
+ CoT decoding	10.6	7.6	17.6
+ constrained CoT decoding	**13.7†‡**	**9.8†**	23.0 ‡
Upper Bounds using SGD Data			
One-shot example from SGD + separate relation prediction	12.9	10.7	16.4
Fine-tuning on SGD	37.3	27.9	57.2

Table 2: Ontology Relation Prediction Results on the SGD test set. Results that are statistically significant over the baseline are highlighted in **bold**. Additionally, significant results based on dialogue-level evaluation for one-shot prompts are marked with *****. Significant results for fine-tuned models, evaluated globally based on five random seeds, are marked with †. Significant improvement over the one-shot model from the SGD upper bound on dialogue-level is marked with ‡. All significance tests are performed at a 5% level of significance.

which do not improve performance however, while increasing computational complexity. In the one-shot approach, we predict each relation type separately, since we found that the LLM struggles with jointly predicting all relation types. We also experimented with a zero-shot approach that performs significantly worse than one-shot.

We fine-tune the LLM via pattern-based fine-tuning (Schick and Schütze, 2021; Ma et al., 2023) with a prompt for all relation types on the Multi-WOZ training split. We consider two upper bounds: an LLM trained on the SGD training split and a model utilising a one-shot exemplar from SGD.

4.2 Evaluation

In evaluation, we only consider relations within dialogues in the ground truth, i.e., both terms of a relation occur in the same dialogue. Relations from equivalent terms to other terms have to be found at least once. If $[term_1, refers\ to\ same\ concept\ as, term_2] \in R_{\text{groundtruth}}$, then $[term_1, r, t] = [term_2, r, t]$, where $R_{\text{groundtruth}}$ is the set of ground truth relations, $r \neq$ 'refers to same concept as' is another relation type and $t \in T$ is a third related term. E.g., the relations [price range, has value, high-end] and [price range, has value, expensive] are equivalent, since [expensive, refers to the same concept as, high-end]. Thus, the prediction of the former relation counts as a prediction for the latter and vice versa.

To compute the global micro F1 score, we compare the accumulated set of relations predicted from all the dialogues with the ground truth ontology relations. Note that we only consider exactly matching terms in relations to be correct.

For significance tests on the one-shot prompted models, we employ a pairwise t-test on dialogue level. For fine-tuned models, we use 5 random seeds for training and an independent t-test.

4.3 Results

Table 2 shows the full results on the target test set, see Appendix A for results for each relation type.

Source One-Shot Approach We found that when predicting all relations at once in a one-shot fashion the model is completely unable to fulfil the task, so we resort to predicting one relation at a time. The one-shot approach is mainly improved through constrained decoding, although the combination with CoT-decoding is also significantly better than the baseline. Note that the source one-shot model is able to get closer to the performance of a model with a one-shot example from the target data with constrained CoT-decoding.

Source Fine-tuning Approach For the source fine-tuned model, *constrained CoT-decoding* leads to significant improvements over the baseline. Furthermore, it significantly outperforms a model using a one-shot exemplar from the target data on all metrics. Constraining CoT-decoding helps per-
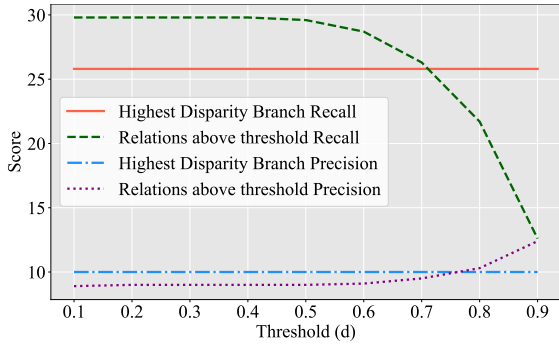
Figure 2: Different relation confidence thresholds across branches compared to the *highest disparity branch approach* for mean aggregation. Displayed are recall and precision for the MWOZ fine-tuned constrained CoT decoding model.

Figure 3: MWOZ fine-tuned constrained CoT decoding model with different choices for the number of branches k for the *highest disparity branch method*. Shown are SGD test set F1 score, precision, and recall.

formance, since the constraints mitigate overconfidence on the source data after fine-tuning.

Interestingly, although the target fine-tuned model is the best model, it is not able to find all relations on the test set. As mentioned in Section 4.1, the SGD test set contains domains different to the SGD training set, which makes this task particularly difficult. In contrast to the excellent performance of LLMs on a variety of tasks, there is a lot of room for improvement on this task.

4.4 Calibration Analysis

In Figure 2, we see that an absolute confidence threshold is not as meaningful and adds the problem of choosing the correct threshold as hyperparameter. Moreover, a high threshold leads to only a small increase in precision, while losing a significant amount of recall. Our results are in line with recent findings about instruction-tuned LLMs (Kapoor et al., 2024) being overconfident. We find that the model's confidence on predicted relations is generally on a high level, indicating overconfidence, as the significant changes in performance happen at high confidence thresholds. For lower thresholds, the performance remains unchanged, as most confidences are quite high and hence the set of predicted relations stays the same. Although this shows that the thresholds are less meaningful, the relative confidence of the branches is meaningful, since choosing the highest disparity branch leads to good performance.

4.5 How useful are predictions from the additional branches?

In line with the findings from Wang and Zhou (2024), we find that for the instruction-tuned
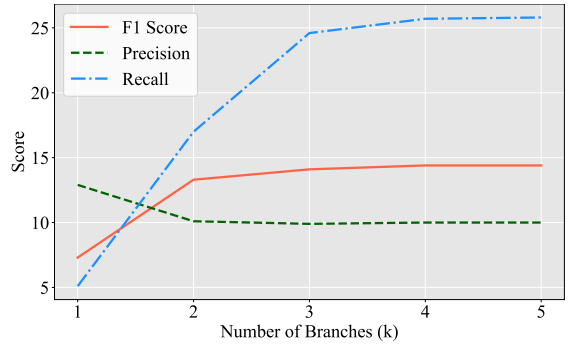
Gemma model, the gain in performance can be mainly attributed to the first additional branch (see Figure 3). While the F1 score is increased slightly up to $k = 3$, the jump in recall from $k = 2$ to $k = 3$ is more significant. This also shows that the branches from lower ranked first tokens lead to responses with higher total confidence across the relation predicted in the respective branch, which is why they are chosen in the highest disparity branch selection method.

4.6 Qualitative Analysis

Wang and Zhou (2024) found that LLMs struggle to generate CoTs for less frequent tasks in the pre-training data. In our analysis, we found that higher confidence completions often follow a recap of the type of terms and relations that should be predicted. Illustrated in Figure 4 is an example of a response to the one-shot equivalence prediction prompt with constrained CoT-decoding (see Appendix B for completions of the other decoding approaches). Here, branches 0 and 1 contain a repetition of the information given in the prompt. In branch 1 however, the focus on the equivalence mentioned in the prompt is followed by a response that does not use the proper format for the answer to be parsed correctly. The last branch has the highest confidence and is chosen ultimately. Here, the focus on the task relation and the provided dialogue is part of the generated introduction to the response.

As seen in Table 3, for the fine-tuned models, there are no such reasonings observed, although the constrained CoT decoding significantly improves performance. We hypothesise that the additional branches together with the constraints in decoding force the model to use task knowledge from

```
PROMPT: <EXAMPLE> Predict the relations between equivalent
terms in the form [term1, refers to same concept as, term2] as in
the example given above.

RESPONSE:
BRANCH 0:
**Domain:** Hotel
    **Slot:** Type of hotel
    **Value:** Price range
    - [price per night,refers to same concept as,135 carlingview drive
etobicoke]
    ....
BRANCH 1:
## Synonyms/Semantically Equivalent Terms:
**Domain:** Hotel
- best western plus toronto airport carlingview hotel
- ....
BRANCH 2:
1.**price range** [1, refers to same concept as, price]...
BRANCH 3:
• **price range** [1, refers to same concept as, price]
• ...
BRANCH 4 (highest confidence):
Here are the predicted synonyms or semantically equivalent terms
from the provided dialogue:
**Price range:**
- [price per night,refers to same concept as,cost]
....
**Day:**
- [tuesday next week,refers to same concept as,check in date]
```

Figure 4: Example of constrained CoT decoding for
one-shot equivalence relation prediction. Branch 4 adds
focus on the relation type. It has the highest average
confidence in the predicted relations and is chosen as
the final response. Some response parts are left out for
illustration purposes ("....").

fine-tuning, rather than what it has learned about
the source data distribution. This can be observed
when comparing CoT-decoding with constrained
CoT-decoding, where the unconstrained version
mainly generates terms it has seen on MultiWOZ,
such as the "reference number" slot that is not
present in SGD. The constrained version on the
other hand forces the model to use task knowledge
instead of distributional knowledge, leading to a
much better coverage of the terms mentioned in
the dataset, if the correct branch is chosen based
on confidence. When observing completions to
other dialogues, we found that the qualitatively best
branches are not necessarily those with the high-
est confidence, indicating that a more sophisticated
branch selection strategy might boost performance
further. We leave such an improvement to future
work. When comparing constrained decoding with
vanilla greedy decoding, it becomes apparent that
constraining the generation greatly improves the
output structure and the utilisation of mentioned
terms in the target dataset.

5   Discussion

Although the performance of the fine-tuned model
is improved by constrained CoT-decoding, it is not
clear where the improvement comes from based on
qualitative analysis alone, as this model generates
no reasoning paths. The workings and explain-
ability of constrained CoT-decoding for fine-tuned
models have not been investigated yet, but are rele-
vant future research directions.

Our results imply that for tasks that are not fre-
quently used in instruction-tuning data, it can be
useful to utilise existing TOD data for training an
LLM rather than annotating a few examples on the
target data for the DORE task (see Section 4.3).
Due to the length of examples in the DORE task,
the amount of annotated examples that can fit in
the prompt is highly limited, aggravating the appli-
cability of few-shot approaches.

The results strengthen the finding that annotated
data should be used if available (Zhu et al., 2023a).
Constrained CoT-decoding only improves perfor-
mance on the target data, where task knowledge is
more useful than distributional knowledge about
the source data. Mahowald et al. (2024) state that
for learning tasks where language is used in a func-
tional way to accomplish certain goals, such as
DORE, self-supervised next-token prediction is not
sufficient. Instead, the model needs to learn task-
specific information via specialised fine-tuning to
solve such tasks, which are not frequently present
in pre-training data and involve task-specific rea-
soning. DORE can only be solved by knowing
the specific relationship definition provided in the
task descriptions, which the model cannot handle
if it was not trained on the task. In summary, the
presented results reinforce the observation that fine-
tuning and specialised architectures are still needed
to solve functional language-based tasks that can-
not be solved by pattern matching alone.

The computational complexity of CoT-decoding
is k times higher than regular greedy decoding,
however, decoding of the different branches can be
done in parallel. Nonetheless, one should note that
only one run of ontology construction is needed,
as the ontology can be repeatedly used for other
dialogue modelling tasks such as state tracking. It
is worth stressing that CoT-decoding only increases
inference cost, while training cost remains the same.
Compared to few-shot prompting, which also in-
creases inference cost due to longer input context,
there is no additional annotation cost.

Approach	Response
Baseline	- [price per night, has domain], [best western plus toronto north york hotel & suites, has domain], '$ 63', 'hotel', 'has slot', 'has domain', 'hotel'] nessunize "$ 63" $ 117' - [best western plus toronto north york hotel & suites, has domain], 'has slot', 'hotel', 'has domain', 'attraction'] - [hotel, has slot],'refers to same concept as','stay'] - [hotel, has slot], 'has domain', 'hotel'] ...
Constrained Decoding	- [price per night,has value,$ 117] - [hotel,has slot,price per night] - [best western plus toronto north york hotel, has domain, hotel] - [hotel,has slot,address] - [best western plus toronto north york hotel, has domain, restaurant]... [hotel,has slot,date] ...
CoT-Decoding	[1, has domain, hotel] [best western plus toronto north york hotel & suites, has domain, hotel] [1, has domain, address] [best western premier toronto airport carlingview hotel, has domain, hotel] [hotel, has slot, day] [best western plus toronto north york hotel, has domain, hotel] [hotel, has slot, name] [hotel, has slot, room] [hotel, has slot, area] [hotel, has slot, reference number] ...
Constrained CoT-Decoding	[1, has domain, hotel] ... [hotel,has slot,price per night] [best western plus toronto north york hotel, has domain, restaurant] [best western plus toronto north york hotel, has domain, hotel] [hotel,has slot,address] [hotel,has slot,check in date] [hotel,has slot,street address] [hotel,has slot,date] [hotel,has slot,tuesday next week] [hotel,has slot,check in date] ...

Table 3: MWOZ fine-tuned model example response excerpts for the different decodings on SDG test dialogue 100. For CoT-decoding methods, only the chosen branch is displayed.

6 Conclusion

We propose constrained chain-of-thought (CoT)-decoding, a new decoding mechanism for dialogue ontology generation (DORE) in a transfer set-up. An LLM using a one-shot example from the source data is significantly improved using the proposed constrained CoT-decoding mechanism. Fine-tuning an LLM on the source data and using constrained CoT-decoding for inference on the target data outperforms a one-shot target data model significantly.

The results warrant further research into DORE in particular, and into eliciting reasoning in LLMs by adapting the decoding mechanism in general. Moreover, we offer a method for applying LLMs to tasks that are underrepresented in pre-training and where the vanilla LLMs perform poorly. Our method is appealing as it does not necessitate labelling new examples. Future research directions include explainability of constrained CoT-decoding in fine-tuned LLMs and including CoT-decoding during fine-tuning.

7 Limitations

In this work we assume a pipeline approach, however with the raise of LLMs, end-to-end solutions tend to be more accurate. We leave the task of jointly extracting dialogue terms and relations for future investigation. Due to constraints in computational infrastructure, we were not able to run open-source LLMs with the size of ChatGPT, which might be promising however. We abstained from utilising proprietary models, such as ChatGPT, for increased transparency and reduced risk of training data contamination.

Furthermore, the need for an annotated source dataset limits the application to low-resource languages and tasks. The reliance on a specific answer structure for confidence estimation limits application to less structured tasks.

Finally, what we consider the upper bound, which was trained on the target dataset, can be argued to be a low bar too, reaching only an F1 of 37. This warrants more research on this task also on the same data setting.

8 Acknowledgements

References

Alon Albalak, Varun Embar, Yi-Lin Tuan, Lise Getoor, and William Yang Wang. 2022. D-REX: Dialogue relation extraction with explanations. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 34–46, Dublin, Ireland. Association for Computational Linguistics.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. PaLM 2 Technical Report. *Preprint*, arXiv:2305.10403.

Nikolay Bogoychev and Pinzhen Chen. 2023. Terminology-aware translation with constrained decoding and large language model prompting. In *Proceedings of the Eighth Conference on Machine Translation*, pages 890–896, Singapore. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Pablo Brusco and Agustín Gravano. 2023. Automatic offline annotation of turn-taking transitions in task-oriented dialogue. *Computer Speech & Language*, 78:101462.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

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.

Pierre-Yves Genest, Pierre-Edouard Portier, Elöd Egyed-Zsigmond, and Laurent-Walter Goix. 2022. PromptORE - A Novel Approach Towards Fully Unsupervised Relation Extraction. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, page 561–571, New York, NY, USA. Association for Computing Machinery.

James Gung, Raphael Shu, Emily Moeng, Wesley Rose, Salvatore Romeo, Arshit Gupta, Yassine Benajiba, Saab Mansour, and Yi Zhang. 2023. Intent induction from conversations for task-oriented dialogue track at DSTC 11. In *Proceedings of The Eleventh Dialog System Technology Challenge*, pages 242–259, Prague, Czech Republic. Association for Computational Linguistics.

Michael Heck, Nurul Lubis, Benjamin Ruppik, Renato Vukovic, Shutong Feng, Christian Geishauser, Hsien-chin Lin, Carel van Niekerk, and Milica Gašić. 2023. ChatGPT for zero-shot dialogue state tracking: A solution or an opportunity? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 936–950, Toronto, Canada. Association for Computational Linguistics.

Michael Heck, Nurul Lubis, Carel van Niekerk, Shutong Feng, Christian Geishauser, Hsien-Chin Lin, and Milica Gašić. 2022. Robust dialogue state tracking with weak supervision and sparse data. *Transactions of the Association for Computational Linguistics*, 10:1175–1192.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Vojtěch Hudeček and Ondrej Dusek. 2023. Are large language models all you need for task-oriented dialogue? In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 216–228, Prague, Czechia. Association for Computational Linguistics.

Vojtěch Hudeček, Ondřej Dušek, and Zhou Yu. 2021. Discovering dialogue slots with weak supervision. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2430–2442, Online. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *Preprint*, arXiv:2310.06825.

Sanyam Kapoor, Nate Gruver, Manley Roberts, Arka Pal, Samuel Dooley, Micah Goldblum, and Andrew Wilson. 2024. Calibration-tuning: Teaching large language models to know what they don't know. In *Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024)*, pages 1–14, St Julians, Malta. Association for Computational Linguistics.

Seungone Kim, Se Joo, Doyoung Kim, Joel Jang, Seonghyeon Ye, Jamin Shin, and Minjoon Seo. 2023. The CoT collection: Improving zero-shot and few-shot learning of language models via chain-of-thought fine-tuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12685–12708, Singapore. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large Language Models are Zero-Shot Reasoners. In *Advances in Neural Information Processing Systems*.

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.

Xue Lilong, Zhang Dan, Dong Yuxiao, and Tang Jie. 2024. AutoRE: Document-Level Relation Extraction with Large Language Models. *Preprint, arXiv:2403.14888*.

Bolei Ma, Ercong Nie, Helmut Schmid, and Hinrich Schuetze. 2023. Is Prompt-Based Finetuning Always Better than Vanilla Finetuning? Insights from Cross-Lingual Language Understanding. In *Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023)*, pages 1–16, Ingolstadt, Germany. Association for Computational Lingustics.

Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2024. Dissociating Language and Thought in Large Language Models. *Trends in Cognitive Sciences*, 28(6):517–540.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. PEFT: State-of-the-art Parameter-Efficient Fine-Tuning methods. https://github.com/huggingface/peft.

R. Thomas McCoy, Shunyu Yao, Dan Friedman, Matthew Hardy, and Thomas L. Griffiths. 2023. Embers of Autoregression: Understanding Large Language Models Through the Problem They are Trained to Solve. *Preprint*, arXiv/2309.13638.

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin

Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open Models Based on Gemini Research and Technology. *Preprint, arXiv:2403.08295*.

Marius Mosbach, Tiago Pimentel, Shauli Ravfogel, Dietrich Klakow, and Yanai Elazar. 2023. Few-shot fine-tuning vs. in-context learning: A fair comparison and evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12284–12314, Toronto, Canada. Association for Computational Linguistics.

Hoang Nguyen, Chenwei Zhang, Ye Liu, and Philip Yu. 2023. Slot induction via pre-trained language model probing and multi-level contrastive learning. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 470–481, Prague, Czechia. Association for Computational Linguistics.

OpenAI. 2022. ChatGPT: Optimizing language models for dialogue. Accessed 2024-05-23.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8689–8696.

Andy Rosenbaum, Saleh Soltan, Wael Hamza, Yannick Versley, and Markus Boese. 2022. LINGUIST: Language model instruction tuning to generate annotated utterances for intent classification and slot tagging. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 218–241, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Shamik Roy, Sailik Sengupta, Daniele Bonadiman, Saab Mansour, and Arshit Gupta. 2024. FLAP: Flow-adhering planning with constrained decoding in LLMs. In *NAACL 2024*.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications. *Preprint, arXiv:2402.07927*.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the*

16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 255–269, Online. Association for Computational Linguistics.

Qingyu Tan, Lu Xu, Lidong Bing, Hwee Tou Ng, and Sharifah Mahani Aljunied. 2022. Revisiting DocRED - addressing the false negative problem in relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8472–8487, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Renato Vukovic, Michael Heck, Benjamin Ruppik, Carel van Niekerk, Marcus Zibrowius, and Milica Gašić. 2022. Dialogue term extraction using transfer learning and topological data analysis. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 564–581, Edinburgh, UK. Association for Computational Linguistics.

Xuezhi Wang and Denny Zhou. 2024. Chain-of-Thought Reasoning Without Prompting. *Preprint, arXiv:2402.10200*.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. 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 *Advances in Neural Information Processing Systems*.

Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, and Enhong Chen. 2023. Large language models for generative information extraction: A survey. *Preprint, arXiv:2312.17617*.

Ze-Song Xu and Yun-Nung Chen. 2023. Zero-shot dialogue relation extraction by relating explainable triggers and relation names. In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, pages 123–128, Toronto, Canada. Association for Computational Linguistics.

Dian Yu, Kai Sun, Claire Cardie, and Dong Yu. 2020. Dialogue-based relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4927–4940, Online. Association for Computational Linguistics.

Dian Yu, Mingqiu Wang, Yuan Cao, Izhak Shafran, Laurent Shafey, and Hagen Soltau. 2022. Unsupervised slot schema induction for task-oriented dialog. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1174–1193, Seattle, United States. Association for Computational Linguistics.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. STaR: Bootstrapping Reasoning With Reasoning. In *Advances in Neural Information Processing Systems*.

Kai Zhang, Bernal Jimenez Gutierrez, and Yu Su. 2023. Aligning instruction tasks unlocks large language models as zero-shot relation extractors. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 794–812, Toronto, Canada. Association for Computational Linguistics.

Dawei Zhu, Xiaoyu Shen, Marius Mosbach, Andreas Stephan, and Dietrich Klakow. 2023a. Weaker than you think: A critical look at weakly supervised learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14229–14253, Toronto, Canada. Association for Computational Linguistics.

Qi Zhu, Christian Geishauser, Hsien-chin Lin, Carel van Niekerk, Baolin Peng, Zheng Zhang, Shutong Feng, Michael Heck, Nurul Lubis, Dazhen Wan, Xiaochen Zhu, Jianfeng Gao, Milica Gašić, and Minlie Huang. 2023b. ConvLab-3: A flexible dialogue system toolkit based on a unified data format. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 106–123, Singapore. Association for Computational Linguistics.

A   Results for Different Relation Types

See Table 4 for results on the different ontology relation types.

B   Example Completions for One-shot Equivalence Relation Prompt

See Table 5 for example completions for the one-shot equivalence prompt.

Approach	Relation Type	F1-Score	Precision	Recall
One-shot example from MultiWOZ				
Baseline: Separate relation prediction	all	7.4	8.8	6.4
	domain-slot	3.9	2.1	41.0
	slot-value	9.4	18.8	6.3
	value-domain	8.0	12.7	5.9
	equivalence	1.8	1.0	6.9
+ constrained decoding	all	8.5	5.7	17.3
	domain-slot	1.1	0.5	50.7
	slot-value	9.3	7.0	13.8
	value-domain	14.4	11.5	19.5
	equivalence	1.4	0.7	16.1
+ CoT decoding	all	9.2	8.8	9.6
	domain-slot	2.5	1.3	50.7
	slot-value	16.0	18.4	14.2
	value-domain	7.7	12.7	5.5
	equivalence	1.8	1.0	7.3
+ constrained CoT decoding	all	9.2	6.4	15.9
	domain-slot	1.2	0.6	50.7
	slot-value	13.4	10.3	19.2
	value-domain	12.3	11.6	13.1
	equivalence	1.2	0.7	10.3
Fine-tuning on MultiWOZ				
Baseline: Fine-tuning on MultiWOZ	all	10.8	6.7	28.4
	domain-slot	5.9	3.2	49.3
	slot-value	6.5	3.7	25.3
	value-domain	20.5	15.3	30.8
	equivalence	4.3	2.5	18.8
+ constrained decoding	all	11.1	6.8	30.4
	domain-slot	4.7	2.5	52.9
	slot-value	6.8	3.9	25.5
	value-domain	19.8	13.9	34.1
	equivalence	3.9	2.1	21.5
+ CoT decoding	all	9.3	6.3	17.4
	domain-slot	3.9	2.0	49.3
	slot-value	5.7	3.5	15.5
	value-domain	16.7	14.8	18.9
	equivalence	2.3	4.4	1.5
+ constrained CoT decoding	all	14.4	10.0	25.8
	domain-slot	3.2	1.6	64.9
	slot-value	12.1	9.0	18.4
	value-domain	19.2	13.9	30.9
	equivalence	4.7	2.8	16.7
Upper Bounds using SGD Data				
One-shot example from SGD + separate relation prediction	all	12.9	10.7	16.4
	domain-slot	3.4	1.8	46.3
	slot-value	17.7	20.9	15.4
	value-domain	13.4	11.0	17.1
	equivalence	2.5	1.4	9.1
Fine-tuning on SGD	all	36.8	26.4	60.9
	domain-slot	7.3	3.8	88.8
	slot-value	30.6	20.7	58.9
	value-domain	53.5	46.9	62.1
	equivalence	12.3	6.9	57.6

Table 4: Ontology Relation Prediction Results for each relation type on the SGD test set.

Approach	Response
Baseline Greedy Decoding	**Domain:** Hotel **Slot:** Type of hotel **Value:** Price range - [price range, refers to same concept as, price] - [reference number, refers to same concept as, reference] **Domain:** Price **Slot:** Cost per night **Value:** $ 63 - [cost per night, refers to same concept as, price] **Domain:** Location **Slot:** Street address **Value:** 135 carlingview drive etobicoke, ontario m9w 5e7, canada - [street address, refers to same concept as, location] ...
Constrained Decoding	**Domain:** Hotel **Slot:** Type of hotel **Value:** Price range - [price per night,refers to same concept as,135 carlingview drive etobicoke, ontario m9w 5e7, canada] - [best western plus toronto north york hotel,refers to same concept as,best western plus toronto north york hotel & suites] ...
CoT-Decoding	Here are the predicted synonyms or semantically equivalent terms from the provided dialogue: **Price range:** - [price range, refers to same concept as, price] **Reference number:** - [reference number, refers to same concept as, reference] **Hotel:** - [hotel, refers to same concept as, best western plus toronto north york hotel & suites] **Stay:** - [stay, refers to same concept as, nights] **Price:** - [price, refers to same concept as, cost]<eos>
Constrained CoT-Decoding	Here are the predicted synonyms or semantically equivalent terms from the provided dialogue: **Price range:** - [price per night,refers to same concept as,cost] **Reference number:** - [price per night,refers to same concept as,check in date] **Hotel:** - [hotel,refers to same concept as,best western premier toronto airport carlingview hotel] **Day:** - [tuesday next week,refers to same concept as,check in date]<eos>

Table 5: MWOZ one-shot only equivalence model example response excerpts for the different encodings on SDG test dialogue 100. For CoT-decoding methods, only the chosen branch is displayed.