

# Probing Structured Semantics Understanding and Generation of Language Models via Question Answering

Anonymous ACL submission

## Abstract

Recent advancement in the capabilities of large language models (LLMs) has triggered a new surge in LLMs' evaluation. Most recent evaluation works tends to evaluate the comprehensive ability of LLMs over series of tasks. However, the deep structure understanding of natural language is rarely explored. In this work, we examine the ability of LLMs to deal with structured semantics on the tasks of question answering with the help of the human-constructed formal language. Specifically, we implement the inter-conversion of natural and formal language through in-context learning of LLMs to verify their ability to understand and generate the structured logical forms. Extensive experiments with models of different sizes and in different formal languages show that today's state-of-the-art LLMs' understanding of the logical forms can approach human level overall, but there still are plenty of room in generating correct logical forms, which suggest that it is more effective to use LLMs to generate more natural language training data to reinforce a small model than directly answering questions with LLMs. Moreover, our results also indicate that models exhibit considerable sensitivity to different formal languages. In general, the formal language with the lower the formalization level, i.e. the more similar it is to natural language, is more LLMs-friendly.

## 1 Introduction

The recent advancement of large language models (LLMs), showcasing their remarkable language understanding and generation capabilities (Bubeck et al., 2023), has garnered significant attention from researchers in artificial intelligence. It has also triggered a surge in LLMs' evaluation endeavors aimed at exploring the boundaries of LLMs' capabilities. Previous works that probe large language models with varying sizes such as BERT (Devlin et al., 2019) and GPT (Radford et al., 2018) mainly fo-

cus on probing models' linguistic knowledge (Hewitt and Manning, 2019; Clark et al., 2019; Liu et al., 2019; Wu et al., 2020) and world knowledge (Petroni et al., 2019; Dai et al., 2022; Jiang et al., 2020; Zhong et al., 2021). Most recent works tend to evaluate models' comprehensive capacities over series of tasks (Bang et al., 2023; Yu et al., 2023) such as reasoning and interactivity.

Among the evaluations, the deep understanding of natural language in LLMs is poorly investigated. There are some existing works that try to explore the structure of language (Hewitt and Manning, 2019; Clark et al., 2019). However, their investigation are confined to the superficial aspects of language, encompassing lexical and syntactic properties. As John McCarthy (McCarthy, 1990, 1959) points out, in order to a better understanding of natural language, it is necessary for an intelligence system to understand the "deep structure" (Chomsky, 2011) of the sentence, which can be explicitly defined in a human-designed formal language. Therefore, we propose to probe the deep understanding of natural language in LLMs with formal languages, which serves to ascertain the boundaries of semantic comprehension exhibited by LLMs and point ways for improving the understanding and generation ability in LLMs.

Specifically, we select the question answering (QA) as our probing task. Existing works mainly utilize the dependency parsing task (Hewitt and Manning, 2019) to find the correlation between the dependency of words and the modules' parameters. However, this approach is intricate, sometimes non-intuitive, and not applicable to non-open-source LLMs. In addition, choosing QA as our probing task brings two advantages: (1) Convenience. There exists many different formal languages, also called logical forms, constructed for knowledge based question answering, which can be directly employed in our experiment. (2) Simplicity of evaluating. To avoid the heavy human evaluation of

083 the semantic correctness of generated text, we can  
084 leverage the accuracy of the answers for indirect  
085 evaluation. In this work, we define two sub-tasks of  
086 knowledge-based question answering as the prob-  
087 ing task: 1) **Formal Language Understanding**,  
088 which aims to automatically translate a piece of  
089 logical form (LF) into its corresponding natural  
090 language question (NLQ). The translation process  
091 can be considered as the model interpreting the pro-  
092 vided LFs in NLQ, thus demonstrating LLMs’ un-  
093 derstanding ability of formal language; 2) **Formal**  
094 **Language Generation**, which aims to correctly  
095 convert a NLQ into its corresponding LF, requiring  
096 the model to not only understand but also generate  
097 LFs, demonstrating its capability in generation.

098 During probing, according to the varying degrees  
099 of formalization (may be broadly understood as the  
100 dissimilarity to natural language, i.e. the higher  
101 level of formalization, the less similar is to natural  
102 language) and different logical structures (e.g. tree  
103 or graph), we choose Lambda DCS (Liang, 2013),  
104 SPARQL, and KoPL (Cao et al., 2022) as repre-  
105 sentative formal languages, which are commonly  
106 used for knowledge based question answering re-  
107 search (Nie et al., 2022; Ye et al., 2022; Shin et al.,  
108 2021). We leverage in-context learning ability of  
109 LLMs for both probing tasks, where the desired out-  
110 puts are generated conditioned on the input along  
111 with a few demonstration pairs of NLQs and LFs  
112 carefully selected from a seed dataset. For demon-  
113 stration selection, we adopt the principle that the  
114 examples should be similar to the target and care-  
115 fully design multiple search strategies for different  
116 formal languages and different tasks.

117 Our findings indicate that there is still a gap be-  
118 tween LLMs and human in terms of structured  
119 semantics understanding. Coinciding with our intu-  
120 ition, the generation capability of LLMs for struc-  
121 tured semantics is much weaker than their under-  
122 standing ability. Importantly, we observe that mod-  
123 els exhibit the sensitivity to different logical forms.  
124 Overall, the lower level of formalization, that is the  
125 closer it is to natural language, the easier it is for  
126 models to understand and generate. These findings  
127 suggest the feasibility of employing LLMs combin-  
128 ing with knowledge bases to tackle the complex  
129 reasoning that currently still pose a challenge to  
130 LLMs (Bang et al., 2023). In conclusion, we be-  
131 lieve that this study will help to examine the deep  
132 language understanding ability of current LLMs,  
133 and also can provide valuable insights for LLMs-  
134 based reasoning approaches.

## 2 Related Work 135

136 Since the success of pretrained language models  
137 (PLMs) such as BERT (Devlin et al., 2019) and  
138 GPT (Radford et al., 2018), there are extensive  
139 works exploring the capability boundaries of PLMs.  
140 Researchers find that the PLMs not only possess  
141 rich linguistics knowledge (Hewitt and Manning,  
142 2019; Clark et al., 2019; Liu et al., 2019; Wu et al.,  
143 2020), but also substantial world knowledge like  
144 entities (Broscheit, 2020), relations (Petroni et al.,  
145 2019; Jiang et al., 2020; Zhong et al., 2021), and  
146 concept (Peng et al., 2022; Dalvi et al., 2022).

147 In line with our work, there are many researchers  
148 focus on exploring whether PLMs have learned  
149 about language structure. One method is utilizing  
150 the “probing task” (Conneau et al., 2018), which  
151 trains a classifier to predict on different syntactic  
152 properties. These tasks is very diverse, such as  
153 segmentation (Liu et al., 2019) and dependency  
154 parsing (Hewitt and Manning, 2019; Clark et al.,  
155 2019). However, the classifier is likely to over-fit  
156 on the probing task when it achieve a high accuracy,  
157 making it hard to tell whether it benefits from the  
158 PLMs’ representation or the training data.

159 Another approach is to find the correlations be-  
160 tween the part of the model’s parameter and the  
161 syntactic properties of the input. Most works tend  
162 to focus on the attention layer. For example, Ra-  
163 ganato and Tiedemann extract dependency trees  
164 from the self-attention weights of the model’s en-  
165 coder to analyze the syntactic and semantic infor-  
166 mation. (Clark et al., 2019) leverage a similar  
167 method with a new attention-head selection strat-  
168 egy. However, other findings argues that the behav-  
169 ior of attention does not always corresponds to the  
170 model itself (Serrano and Smith, 2019; Wiegreffe  
171 and Pinter, 2019). Moreover, this method is not  
172 suitable for the non-open-source LLMs.

173 Our work are inspired by both the “deep struc-  
174 ture” idea brought up by Chomsky and the discus-  
175 sion of the benefit of a formalism in McCarthy;  
176 McCarthy, using the formal language as a medium  
177 to probe for deep structure understanding of LLMs.  
178 One recent work (Papadimitriou and Jurafsky,  
179 2023) that also using the formal language explores  
180 what is the necessary structure for learning lan-  
181 guage. In this work, language models are pre-  
182 trained purely on different structure types from  
183 scratch before testing their language learning abil-  
184 ity. However, this work cannot explain how well  
185 current LLMs understand structures.

<b>Question:</b>	<i>What is the number of animated movies published after 1940?</i>
<b>KoPL:</b>	FindAll().FilterYear(publication date, 1940, >). FilterConcept(animated film).Count()
<b>SPARQL:</b>	SELECT (COUNT(DISTINCT ?e) AS ?count) WHERE { ?e <pred:instance_of> ?c . ?c <pred:name> "animated film" . ?e <pred:publication_date> ?pv . ?pv <pred:year> ?v . FILTER ( ?v > 1940 ) . }
<b>Lambda DCS:</b>	( call @listValue ( call .size ( call @filter ( call @getProperty ( call @singleton en.animated_film ) ( string ! type ) ( string publication_date ) ( string > ) ( year 1940 ) ) ) ) )

Figure 1: An example of a natural language question and its corresponding logical forms in KoPL, SPARQL, and Lambda DCS.

### 3 Framework

In this section, we will introduce the details of the probing framework, which contains two parts: (1) the formal languages and their corresponding datasets, (2) the probing task design with the formal descriptions, and the evaluation of semantic correctness of the generated natural language or formal language text.

#### 3.1 Formal Language

As mention in 1, we choose three representative formal languages according to the varying degrees of formalization and different logical structure, and they are Lambda DCS, SPARQL, and KoPL. Some examples are shown in Figure 1.

**Lambda DCS** is a tree-structured programming language developed from Lambda calculus, similar to church and s-expression. Lambda DCS removes the explicit variables in Lambda calculus, making it similar to dependency-based compositional semantics (Liang, 2013). For this language, we use Overnight dataset (Wang et al., 2015), which contains over 13,000 data examples in eight domains extracted from Freebase. We follow the standard split used in Wang et al..

**SPARQL** is a popular query language and it provides a standardized way for users to search and retrieve information stored in RDF databases and other Linked Open Data<sup>1</sup>. The SPARQL describes the relations between entities using triples in the form of a graph structure. For this language, we use the GrailQA dataset (Gu et al., 2021), which is constructed based on Freebase and comprises a total of over 50,000 data entries along with their entity linking results. We also followed the standard split used by the author (Gu et al., 2021).

<sup>1</sup><https://www.w3.org/TR/sparql11-query/>

**KoPL** (Cao et al., 2022) is a programming language constructed using symbolic functions, which define the fundamental and atomic operations performed on knowledge bases. These functions are combined according to the operations in the reasoning process, forming a tree structure program. For this language, we use the KQA Pro dataset (Cao et al., 2022), which is based on Wikidata and comprises a total of over 100,000 data entries. Each data entry includes a NLQ along with its corresponding KoPL program and SPARQL query. We followed the standard split described in Cao et al..

Basic features of these formal languages can be concluded that (1) KoPL and Lambda DCS are both tree-structured programs, which can potentially better reflect the “chain of thought” - reasoning process than SPARQL, and (2) KoPL is more well-modularized and uses more human-readable identifiers and function input, making it closer to the distribution of natural language.

#### 3.2 Probing Task and Evaluation

As shown in Figure 2, We define two probing tasks, namely the **formal language understanding** and **formal language generation**. In this section, we introduce the processes and evaluation methods.

##### 3.2.1 Formal Language Understanding

The goal of the task is for a LLM  $M$  to translate a LF input to its corresponding NLQ. Formally, we have a target set  $\mathcal{T} = \{l^*\}$  of LFs, and a seed set  $\mathcal{S} = \{(l, q)\}$  of LF - NLQ pairs. To assemble the demonstration, for every  $l$  in  $\mathcal{T}$  we need to retrieve  $k$  pairs of LFs and NLQs  $(l_1, q_1), \dots, (l_k, q_k)$  from  $\mathcal{S}$ . Conditioned on the examples and  $l$ , the model translates it into a NLQ.

For the evaluation of the quality of the generated NLQs, the commonly used automatic metrics to compare text similarity like BLUE (Papineni et al., 2002) and BERT-Score (Zhang et al., 2020) are not reliable enough. Instead, we evaluate the generation quality of a model  $M$  indirectly by comparing the performance of the parser trained on the model-generated data and the parser trained on the manually-labeled data. Formally, given the training set  $\{(q^*, l^*)\}$ , where the  $l^*$  is the LF and  $q^*$  is the corresponding human-labeled NLQ, we train a baseline semantic parser  $P_{human}$ . Then we take  $\{l^*\}$  as the target set  $\mathcal{T}$ , using  $M$  to generate a same-size pseudo training set  $\{(q_M, l^*)\}$ , which is used to train another parser  $P_M$ . In this case, the generation quality of  $M$  is measured by  $P_M$ 's

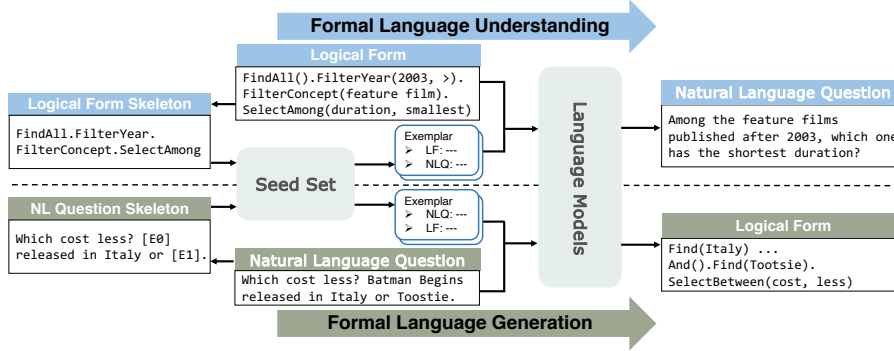


Figure 2: A simple illustration for the probing task of both formal language understanding and generation.

performance  $Accuracy_{P_M}$  to  $Accuracy_{P_{human}}$  of  $P_{human}$ . Higher score means better quality of the model-generated questions, indicating closer understanding ability of  $M$  is to human.

### 3.2.2 Formal Language Generation

The goal is for a LLM  $M$  to directly translate a NLQ back to its correct LF. Similarly, we have a target set of  $\mathcal{T} = \{q^*\}$  of NLQs, and a seed set  $\mathcal{S} = \{(l, q)\}$  of LF - NLQ pairs. For every  $q$  in  $\mathcal{T}$ , we retrieve  $k$  pairs of NLQs and LFs  $(q_1, l_1), \dots, (q_k, l_k)$  from  $\mathcal{S}$  to assemble the final prompt. The model is supposed to generate the correct LF  $l$  conditioned on the examples and  $q$ .

The evaluation of the generated  $l$  is relatively easier. To evaluate whether the generated LF are correct and semantically equivalent to the input  $q$ , we can use the either the exact match score with the golden logical forms, or the accuracy of the answer by putting the logical forms into an executor.

## 4 Implementation

As mentioned above, we mainly leverage the in-context learning (ICL) ability of LLMs to generate the output for the probing task. The demonstration selection is considered as the most critical part of this method. In this work, we adopt the principle to search most similar examples to the target  $l$ , and decently order the examples by the similarity (Liu et al., 2021) in the prompt.

### 4.1 Formal Language Understanding

In this task, the input of LLMs is the LF  $l^*$ , so we search for examples  $(l, q)$  from  $\mathcal{S}$  where all  $l$ s are most similar to  $l^*$ .

We consider that the retrieved examples should (1) have the most similar **logical structure** to the structure of the target logical form  $l^*$  and (2) share as many same **relations** as possible with  $l^*$ .

#### 4.1.1 Structure-Preserving Principle

In order to find the most structure-similar examples from  $\mathcal{S}$ , we first transform the original logical form  $l^*$  into a simple rooted tree-like structure  $s^*$  called skeleton, where  $s^* \leftarrow f(l^*)$ ,  $f$  being the extraction function. Specifically, KoPL program is already a tree of functions, therefore the skeleton of KoPL is the tree formed by removing the functions' inputs. The Lambda DCS program is similar to KoPL, since it can be treated as a bracket tree. The SPARQL program is more complicated, since it depicts a graph by some triples. In this case, we use the corresponding S-expression program instead, which is also bracket tree. Afterwards, we group the examples in  $\mathcal{S}$  using the skeleton of logical form as the key.

Then we find the most similar structure naturally by computing the tree edit distance (TED) between  $s^*$  and skeleton keys of  $\mathcal{S}$ . However, considering the overhead of the minimum TED algorithm, we serialized the tree structure and apply the simple minimum edit distance (ED) in practice. In general, these two algorithms can produce every different results. But due to the grammar restriction of program, the candidates at small distances computed by TED are almost the same to those of ED. For example, in KoPL there are some common fixed patterns like  $Find() \rightarrow Relate() \rightarrow Filter()$ .

#### 4.1.2 Content-Preserving Principle

The meaning of content here is two-fold. First off, there should be no symbols of  $l^*$  unseen in the demonstration examples. Taking KoPL as example again, it means the function names need to be covered by demonstration examples as many as possible. This is a max cover problem and we perform a k-step greedy search based on the previous ranking result by edit distance. Specifically, providing there are  $m$  skeletons  $S = \{s_1, \dots, s_m\}$  that

are closest to the skeleton  $s^*$  of  $l^*$  at a distance of  $d_0$ , we select a  $s_{t_i}$  at each time step  $i$ , so that,

$$s_{t_i} = \arg \min_{s \in S_i} |s^*|_i - |s_{t_i}|$$

$$S_i = S - s_{t_{i-1}} - \dots - s_{t_1} \quad (1)$$

$$|s^*|_i = |s^*| - |s_{t_{i-1}}| - \dots - |s_{t_1}|$$

where the  $|\cdot|$  represent the operator to get the set of node labels. After  $k$  steps, we get a set of skeleton candidates  $\{s_1, \dots, s_k\}$ .

Moreover, the input content such as relations and entities can also be taken into account. In summary, the first priority for selecting examples is structural similarity, followed by the shared content.

## 4.2 Formal Language Generation

In this task, the input of LLMs is the NLQ  $q^*$ , so we search for example pairs  $(q, l)$  from  $\mathcal{S}$  where  $qs$  are most similar to  $q^*$ .

We hope that the retrieved questions has a similar deep structure with  $q^*$ . Therefore, we intuitively utilize the structure prompt word such as “and” “or”, which might indicate the concatenation structure in the question. Specifically, we use the BM25 algorithm to search similar questions with the same prompt words. As shown in Figure. 2, when constructing the BM25 searcher, we mask the entities and relations in the question to exclude their interference on the BM25 algorithm. This search method cannot guarantee the structure similarity, but it is a viable baseline method.

We did not adopt other searching algorithm such as embedding similarity because we want to focus on the structure word like conjunctions. However, embedding-based search can easily neglect them and focus on content words in the sentences.

### 4.2.1 Entity Linking

We found that the model often fails to generate the correct names of entities and relations in the knowledge base. Therefore, to generate correct and executable programs, we need to map the generated entities to the corresponding entries in the knowledge base. Taking the example of SPARQL based on Freebase, SPARQL statements use Freebase’s *mid* to represent entities, but the model struggles to generate the correct *mid*. Similar to Li et al., we have the model first generate the natural language names of the entities. Then, we use the BM25 algorithm to find the most similar entities in the entity repository. For concepts and relations in the knowledge base such as *type.object.type* and

*rail.rail\_network*, although they are not represented by *mid*, they are also often inaccurately generated. Thus, we use the BM25 algorithm to search for the most similar names in the knowledge base and replace them accordingly.

Apart from correcting them after the generation, we find that other tricks such as adding entities and relations related to the question into the prompt also help improving the performance. More details and examples of this whole process including intermediate steps are provided in Appendix.

## 5 Experiment Setup

We introduce a range of popular language models that have been extensively studied in our experiments (5.1) as long as the semantic parsing models we use to evaluate the performance of the understanding task (3.2).

### 5.1 Investigated Models

In order to investigate the impact of the model scale on its capacity, we select models of different sizes.

For medium size models ranging from 100M to 10B, we mainly consider two types of models. The first is auto-regressive models, exemplified by the GPT series. These models only use the decoder in training and employ a unidirectional “predict the next word” auto-regressive loss function for modeling. The second type is represented by T5, a text-to-text model, which utilizes a bidirectional encoder and a unidirectional decoder to predict masked spans. In the experiment, we use the instruction-tuned version FLAN-T5 series. In particular, we select **GPT2-Large** (774M), **GPT2-XL**(1.5B), **GPT-J** (6B), **FLAN-T5-L** (770M), **FLAN-T5-XL** (3B), **FLAN-T5-XXL** (11B).

For large models over 100B, we first consider the instruction-tuned GPT 3.5 series, including the initial Davinci model **text-davinci-001** and the most powerful **text-davinci-003** (maybe 175B). We also investigate **GLM-130B**, an open bilingual pretrained model without instruction-tuning and RLHF. We do not consider chat-optimized model like gpt-3.5-turbo since we don’t need the chat ability. The code-pretrained model like CODEX is also not used because of it has closed access.

### 5.2 Evaluation Models

The evaluation methods is mentioned above in 3.2, In practice, different semantic parsers are chosen for the evaluation of different formal languages and datasets.

For KoPL and KQA Pro dataset, we use the original baseline (BART-base) provided KQA Pro (Cao et al., 2022). For Lambda DCS and Overnight dataset, we train a bidirectional LSTM with dual learning algorithm described by Cao et al.. Finally, for SPARQL and GrailQA, we tried two baseline models. One is also a simple sequence-to-sequence BART-base generation model without explicit entity linking modules. The other baseline is a rank-and-generate (RnG) pipeline with an entity linking module described in Ye et al., which employs a ranker to retrieve related logical forms that share similar entities and relations. The implementation detail of parsers and training hyper-parameters used in the work can be found in Appendix.

## 6 Results and Analyses

We first present the main result of the formal language understanding and generation in Table. 1.

In the left blue section of understanding task, the figures are the absolute performance of the evaluation parser trained on training sets that generated by different models. The retrieved examples of the input prompt of ICL is 3 for all models in the understanding task.

The right green section presents the semantic parsing result of the models, where the retrieved examples are as many as the input context can take. To cut down computation overhead, the test sets here are randomly sampled subsets of 300, 120, 240 examples from the test sets of KQA Pro, GrailQA, and Overnight, respectively. The parsing performance of KoPL and Lambda DCS are measured by answers' accuracy, and the SPARQL performance are measure by answers' F1 score. Note that the human's performance is not applicable here, but we can compare it to the baseline results of understanding task. Also, we only test the model over 1 B because the small models perform poorly, only producing meaningless results.

Then we present the conclusions and findings by analyzing them along with other ablation experiments. More detailed results for some dataset can be found in Appendix.

### 6.1 Formal Language Understanding Result Analysis

As shown in Table 1, we can see that (1) All language models demonstrate a certain degree of understanding of formal languages, as evidenced by their ability to generate new training data to train a

non-trivial parser. (2) However, there is still a non-negligible gap between the overall structured semantics understanding ability of language models compared to human. (3) In addition, larger models tend to perform better in understanding structured semantics. (4) As for the parser for evaluation, the RnG parser can virtually eliminate gaps in generated data quality, reflecting the importance of entity linking module. (5) Meanwhile, it is noteworthy that we do not observe significant differences between models that are instruction tuned and those that are not. The model size evidently has a more pronounced impact.

Most interestingly, We observe peculiar characteristics in the FLAN-T5 series, where the smaller models demonstrated stronger capabilities compared to the larger ones. In the appendix, we present some error analysis from FLAN-T5-XXL, whose generated results are almost unintelligible.

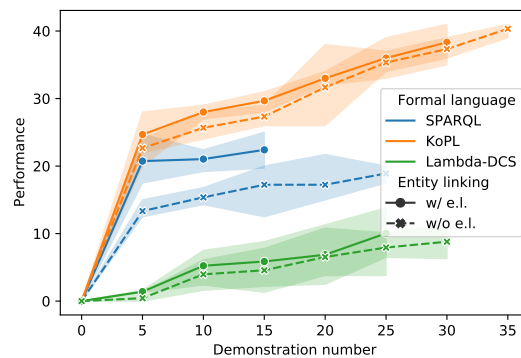


Figure 3: Formal language generation performance of Text-Davinci-003 with various numbers of demonstration examples. The entity linking tag means whether to use entity linking to detect the entities in input and add their 2-hop-related entity and relation names to the input. Note that the difference of maximum demonstration number between formal languages is because the context length of LLM.

### 6.2 Formal Language Geration Result Analysis

From the right section of Table. 1, we can see that the generation ability of language models is far worse than their understanding ability. Compared to the left section, even the most powerful model directly generating logical forms can only achieve 15% to 50% accuracy to the parser trained by its generated data. Therefore, we believe it is safe to reach to the conclusion that, to improve performance on knowledge based question answering, it

Model	Understanding				Generation		
	KoPL	SPARQL	SPARQL <sub>‡</sub>	Lambda DCS	KoPL	SPARQL	Lambda DCS
GPT2-L (774M)	76.0	10.8	70.8	39.1	—	—	—
GPT2-XL (1.5B)	83.3	14.4	71.1	42.3	—	—	—
GPT-J (6B)	84.2	16.7	72.2	78.4	4.3	1.6	0.0
FLAN-T5-L (770M)	48.6	6.8	71.6	27.5	—	—	—
FLAN-T5-XL (3B)	26.6	7.1	71.2	17.0	—	—	—
FLAN-T5-XXL (11B)	12.7	7.0	71.4	12.4	2.7	0.0	0.0
GLM-130B	86.2	19.2	73.6	77.0	22.3	4.6	3.8
Text-Davinci-001	85.6	18.7	71.4	67.3	16.0	2.7	1.7
Text-Davinci-003	88.1	21.7	73.8	76.0	41.6	22.5	10.0
Human	90.6	28.1	74.7	95.2	—	—	—

Table 1: The main results of formal language understanding and generation. ‡ means that these column is evaluated by the RnG parser with an entity linking module. The — in the table means the result is too low to be meaningful or it is not applicable.

is much more practical to generate new data for training small parser like Bart model than directly using LLMs to do parsing only by prompting without touching the parameters.

To improve the performance of direct semantic parsing, two approaches are tried in our experiment. The first is increasing the examples of ICL and the second is to detect the entities mentioned in the input question, and include their 2-hop-related entity and relation names from the knowledge base into the prompt (as mentioned in 4.2.1). To compare the impact of these two strategies on the performance, we conduct a series of experiment on Text-Davinci-003. As shown in Figure. 3, (1) Both strategies can contribute to the performance. (2) The performance on KoPL notably improves with the increase of examples. However, for SPARQL and Lambda DCS, the effect of this strategy is limited. (3) On the other hand, incorporating entity and relation names in the prompt significantly enhances the results for SPARQL. (4) In all settings, model performs best on KoPL and worst on Lambda DCS, and SPARQL in between.

Empirically, We figure the possible explanations for these phenomena lie in the difference between formal languages. As the example show in Figure. 1, KoPL is the most similar to natural language. The identifiers are easy for human to understand, and the order of functions correspond to the “chain-of-thought” reasoning process. While both SPARQL and Lambda DCS are more formalized and contain lots of identifiers that do not make sense in natural language. This might explain why model performs best on KoPL, and most benefits

from the increasing of examples. Furthermore, we note that the grammar of SPARQL is simpler and lacks of variations, where the SPARQL queries in the GrailQA dataset almost follow the same pattern. But the bottle-neck for writing SPARQL is to generate the correct entity or relation names in Freebase. This explains why model performs better on SPARQL than Lambda DCS, and why adding entities to prompt improves the most for SPARQL.

### 6.3 Zero-shot Understanding

We are also very interested in whether the LLMs truly understand the logical forms or they merely are good at imitating the carefully selected examples we provided? To figure it out, we conduct an ablation experiment where input for the QG task is replaced with the description of the formal language. This experiment is only conducted on KoPL since it is well modularized and the function of the operations can be concisely explained. The input description consists of the one-sentence descriptions of each operation function in KoPL, optionally accompanied by several fixed simple examples. To reduce the cost, we only use a subset that contain the first 20,000 examples of KoPL (the same in next experiment in 6.4) and only probe the GPT series.

As shown in the Table. 2, it can be observed that the carefully designed retrieval strategy in our baseline method indeed significantly contributes to generating high-quality natural language questions. However, at the same time, the model itself exhibits a certain degree of understanding ability in situations when examples are lacking, where Text-

Model	KoPL <sub>1%seed</sub>	KoPL <sub>zero-shot</sub>
GPT-J (6B)	43.3	11.9
GLM-130B	76.4	46.0
Text-Davinci-001	76.8	44.6
Text-Davinci-003	80.0	62.7
Human	84.6	84.6

Table 2: Formal language understanding results for the low-resource seed set setting and the zero-shot setting.

Davinci-003 demonstrates a 25.8% performance drop.

#### 6.4 Different Seed Set Ratio

The main result in our experiment are generated with the whole training set as the seed set. However, considering the practical limitations in obtaining a large amount of high-quality manually annotated data in real scenarios, we investigate the model’s ability to generate new data with only a small amount of labeled data as seeds.

This experiment is also conducted on KQA Pro since it is the largest and most diverse dataset we use. We randomly sample 1% of training set as seeds. The result in Table 2 indicate that although there is a decrease in the quality of generated questions, the performance degradation of the model is acceptable, given the significant reduction in seed number.

Model	Understanding	
	SPARQL	Lambda DCS
GPT-J (6B)	71.9	62.4
GLM-130B	76.3	64.8
Text-Davinci-001	74.4	61.6
Text-Davinci-003	80.2	69.7
Human	82.7	76.1
Model	Generation	
	SPARQL	Lambda DCS
Text-Davinci-003	14.2	4.2

Table 3: Formal language understanding and generation results for the one-dataset setting.

#### 6.5 All Formal Languages on One Dataset

Since different datasets are constructed on different knowledge bases, in order to compare whether the three logic forms can arrive at the similar conclusions on identical data as previously observed,

we conduct a experiment testing the three formal languages on the same dataset.

This experiment is also conducted on KQA Pro for convenience, because it already contains KoPL and SPARQL, and the parser for evaluation also switches to BART-base, the same with KQA Pro. And we follows Nie et al. to translate KoPL into Lambda DCS. From results in Table. 3, overall the results are consistent to the main result in Table. 1. But the performance of generation drops a bit, because for SPARQL, the entity and relation binding process are skipped in this experiment.

## 7 Conclusion

In this work, we leverage the formal language to probe the deep structure understanding of natural language in LLMs. Our observations suggest that there still exists a gap between LLMs and human. Besides, aligning with our intuition, the ability of LLMs to generate structured semantics is notably inferior to their ability to understand it.

Beyond these basic conclusions, we also discover that factors influencing the model’s understanding and generation of structured semantics extend beyond the model’s inherent capability and the generation methods employed. The choice of formal language and knowledge base utilized also exerts significant influence on models’ performance.

In our experiment, models performing on KoPL yields the best results on nearly all experiments. We believe that it is because KoPL employs expressions that are more similar to natural language while preserving the structure and modularity. Importantly, its expression of the reasoning process closely aligns with human. However, SPARQL and Lambda DCS face more challenges in grounding the generated entities to the knowledge base for their level of formalization is too high. As a result, KoPL proves to be the most LLMs-friendly for semantic parsing task among the formal languages that we investigate in this work.

In general, we want to point out that the formal language plays an important role in enhancing the power of LLMs. For example, a formal language can be used as a medium between LLMs and the knowledge base, so that LLMs can use the knowledge base as a tool to enhance the performance of QA and reasoning tasks. On the other hand, the selection of a more model-friendly formal language, one that closely resembles the natural language in which models excel, should be prioritized.



659  
660  
661  
662  
663  
664  
665  
666  
667  
668  
669  
670  
671  
672  
673  
674  
675  
676  
677  
678  
679  
680  
681  
682  
683  
684  
685  
686  
687  
688  
689  
690  
691  
692  
693  
694  
695  
696  
697  
698  
699  
700  
701  
702  
703  
704  
705  
706  
707  
708  
709  
710  
711  
712  
713  
714  
715  
716

## References

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. [A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity](#). *CoRR*, abs/2302.04023.

Samuel Broscheit. 2020. [Investigating entity knowledge in BERT with simple neural end-to-end entity linking](#). *CoRR*, abs/2003.05473.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Túlio Ribeiro, and Yi Zhang. 2023. [Sparks of artificial general intelligence: Early experiments with GPT-4](#). *CoRR*, abs/2303.12712.

Ruisheng Cao, Su Zhu, Chen Liu, Jieyu Li, and Kai Yu. 2019. [Semantic parsing with dual learning](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 51–64, Florence, Italy. Association for Computational Linguistics.

Shulin Cao, Jiaxin Shi, Liangming Pan, Lunyu Nie, Yutong Xiang, Lei Hou, Juanzi Li, Bin He, and Hanwang Zhang. 2022. [KQA pro: A dataset with explicit compositional programs for complex question answering over knowledge base](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2022, Dublin, Ireland, May 22-27, 2022, pages 6101–6119. Association for Computational Linguistics.

Noam Chomsky. 2011. *Current issues in linguistic theory*, volume 38. Walter de Gruyter.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. [What does BERT look at? an analysis of bert’s attention](#). In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@ACL 2019, Florence, Italy, August 1, 2019*, pages 276–286. Association for Computational Linguistics.

Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. [What you can cram into a single vector: Probing sentence embeddings for linguistic properties](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2126–2136. Association for Computational Linguistics.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. [Knowledge neurons in pretrained transformers](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2022, Dublin, Ireland, May 22-27, 2022, pages 8493–8502. Association for Computational Linguistics.

Fahim Dalvi, Abdul Rafae Khan, Firoj Alam, Nadir Durrani, Jia Xu, and Hassan Sajjad. 2022. [Discovering latent concepts learned in BERT](#). In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Yu Gu, Sue Kase, Michelle Vanni, Brian M. Sadler, Percy Liang, Xifeng Yan, and Yu Su. 2021. [Beyond I.I.D.: three levels of generalization for question answering on knowledge bases](#). In *WWW ’21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, pages 3477–3488. ACM / IW3C2.

John Hewitt and Christopher D. Manning. 2019. [A structural probe for finding syntax in word representations](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4129–4138. Association for Computational Linguistics.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. [How can we know what language models know](#). *Trans. Assoc. Comput. Linguistics*, 8:423–438.

Tianle Li, Xueguang Ma, Alex Zhuang, Yu Gu, Yu Su, and Wenhui Chen. 2023. [Few-shot in-context learning for knowledge base question answering](#). *CoRR*, abs/2305.01750.

Percy Liang. 2013. [Lambda dependency-based compositional semantics](#). *CoRR*, abs/1309.4408.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. [What makes good in-context examples for gpt-3?](#) *arXiv preprint arXiv:2101.06804*.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. [Linguistic knowledge and transferability of contextual representations](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 1073–1094. Association for Computational Linguistics.

John McCarthy. 1959. Programs with common sense.



888	Zexuan Zhong, Dan Friedman, and Danqi Chen. 2021.	it is analyzed from the performance of different	935
889	<a href="#">Factual probing is [MASK]: learning vs. learning</a>	models, and on the other hand, it is analyzed from	936
890	<a href="#">to recall</a> . In <i>Proceedings of the 2021 Conference</i>	the different types of errors produced by the same	937
891	<i>of the North American Chapter of the Association</i>	model.	938
892	<i>for Computational Linguistics: Human Language</i>		
893	<i>Technologies, NAACL-HLT 2021, Online, June 6-11,</i>	<b>B.1.1 Performance of Different Models</b>	939
894	<i>2021</i> , pages 5017–5033. Association for Computa-	Examples of KoPL, SPARQL, and Lambda DCS	940
895	tional Linguistics.	is shown in Table. 9, 10, and 11, respectively.	941
896	<b>A Details of Probing Process</b>	In general, larger models perform better than	942
897	In this section we present the details of the probing	smaller models, whose output is often hallucinated	943
898	processes of both probing sub-tasks.	and which tends to miss some semantics in the	944
899	<b>A.1 Formal Language Understanding</b>	input. From the horizontal comparison of different	945
900	In this task, we search the seed set for demonstra-	formal languages, small models perform better on	946
901	tion examples based on the structure of the input	KoPL than SPARQL and Lambda DCS, indicating	947
902	logical form. As stated in the main submission, we	that KoPL is more model-friendly.	948
903	first transform the logical forms into corresponding	A peculiar phenomenon was found in the ex-	949
904	skeletons.	periment, that is, the flan-t5 series models have	950
905	The skeleton of KoPL is the tree formed by re-	poor generalization for formal languages that have	951
906	moving the functions’ inputs, and we serialize it	not been seen in this type of pre-training. And	952
907	with post-order traversal. The Lambda DCS pro-	we found that the larger the size of the model, the	953
908	gram is similar, it is a bracket tree and its skeleton	lower the overall quality of the generated natural	954
909	is also also formed by only keeping identifiers. The	language questions.	955
910	SPARQL program depicts a graph by some triples,	<b>B.1.2 Error Types on KoPL</b>	956
911	and the algorithm for finding graphs with the same	We analyse the error types of GLM-130B on KoPL.	957
912	structure is complex, so we use the SPAQRL’s cor-	When retrieved examples’ skeletons <b>are</b> exactly	958
913	responding S-expression, which is also a bracket	the same with the skeleton of the input KoPL pro-	959
914	tree structure. The serialized examples of the logi-	gram, the output is usually good (shown in Ta-	960
915	cal form skeleton is illustrated in Table. 4, respec-	ble. 12). However, there are sometimes exceptions,	961
916	tively.	and the model will add some hallucinatory compo-	962
917	<b>A.2 Formal Language Generation</b>	nents to the output (shown in Table.13).	963
918	In this task, we search the seed set for demonstra-	When retrieved examples’ skeletons <b>not</b> the	964
919	tion examples only based on the input natural lan-	same with the skeleton of the input KoPL program,	965
920	guage question. As mentioned in the main sub-	hallucinatory content is more likely to be included	966
921	mission, we mask the entities and relations in the	in the result (shown in Table. 14), and attributive	967
922	question to get the NLQ skeleton. We take KoPL	parts tend to be missed for longer inputs (shown in	968
923	as an example, where the demonstration number	Table. 15).	969
924	equals 3. The the input question, skeleton and fi-	<b>B.2 Formal language Generation</b>	970
925	nal prompt is illustrated in Table. 7. This method	In this task, since the output of most of the small	971
926	works the same for other two formal languages, so	models is usually meaningless content, it is also	972
927	we will not continue to show examples here.	pointless to analyze them. So in this section, we	973
928	<b>A.3 Zero-shot Understanding</b>	mainly analyze the error results of the best model	974
929	The prompt used in the experiment of zero-shot	<b>Text-Davinci-003</b> on the three different formal lan-	975
930	understanding is shown in Table. 8	guages.	976
931	<b>B Error Analysis</b>	<b>B.2.1 KoPL</b>	977
932	<b>B.1 Formal Language Understanding</b>	The errors of the model on KoPL are mainly logical	978
933	In this section, we will discuss the results and errors	errors, which are manifested in the use of inappropri-	979
934	of the experiment from two aspects. On one hand,	ate functions, or the wrong input and order of	980
		functions, etc. Examples are shown in Table. 16.	981

<b>NATURAL LANGUAGE QUESTION :</b> What is the name of the actor that was born in 1956-04-19 ?
<b>KOPL PROGRAM:</b> FindAll().FilterDate(date of birth, 1956-04-19, =).FilterConcept(human).Find(actor).Relate(occupation, backward).FilterConcept(human).And().What()
<b>KOPL SKELETON:</b> FindAll.FilterDate.FilterConcept.Find.Relate.FilterConcept.And.What
<b>PROMPT:</b> According to the given logic form kopl, generate the corresponding natural language question. For examples, FindAll()FilterDate(date of birth, 1989-04-06, =)FilterConcept(human)Find(United States of America)Relate(country of citizenship, backward)FilterConcept(human)And()What() is verbalized as: Which human was born 1989-04-06 and is a citizen of the United States of America? [SEP] FindAll()FilterDate(date of birth, 1977-03-10, =)FilterConcept(human)Find(association football)Relate(sport, backward)FilterConcept(human)And()What() is verbalized as: Which human has the date of birth 1977-03-10 and is related to the sport association football? [SEP] FindAll()FilterDate(date of birth, 1956-04-19, =)FilterConcept(human)Find(actor)Relate(occupation, backward)FilterConcept(human)And()What() is verbalized as: What is the name of the actor that was born in 1956-04-19? [SEP] FindAll()FilterStr(TOID, 4000000074573917)FilterConcept(town)FindAll()FilterStr(OS grid reference, SP8778)FilterConcept(town)And()What() is verbalized as:

Table 4: Serialized examples of the KoPL and its corresponding skeletons, and final input prompt.

### 982 B.3 SPARQL

983 The error of the model on SPARQL is mainly the  
984 wrong name of the entity and the relationship, be-  
985 cause in the GrailQA dataset, most of the SPARQL  
986 query patterns are the same, only the specific en-  
987 tities and relationships are different, so the main  
988 difficulty lies in generating the correct freebase  
989 mid. Examples are shown in Table. 17. In the main,  
990 submission, we mentioned that the entity and rela-  
991 tion are aligned to the knowledge base through the  
992 BM25 algorithm. The output shown here is before  
993 alignment.

### 994 B.4 Lambda DCS

995 The error types of the model on Lambda DCS  
996 contains both the types mentioned in KoPL and  
997 SPARQL, including both logical errors and names  
998 error. The result is illustrated in Table. 18.

## 999 C Details of Model Implementation

### 1000 C.1 Semantic Parser for Evaluation

1001 In this section, we detail the implementation of the  
1002 semantic parser used in the evaluation of formal  
1003 language understanding task.

1004 For **Main Results**, where we probe LLMs' un-  
1005 derstanding ability of KoPL on KQA Pro, SPARQL  
1006 on GrailQA, Lambda DCS on Overnight, the se-  
1007 mantic parser and the training hyper-parameters  
1008 are as followed.

1009 For KoPL, we train the BART-base model as a  
1010 sequence-to-sequence baseline parser described in

KQA Pro (Cao et al., 2022). The code is provided  
in the Github<sup>2</sup>. For training, the batch size equals 1,  
the epoch number equals 10, gradient accumulation  
equals 1, and an AdamW optimizer with learning  
rate 1e-4, weight decay 1e-5, adam epsilon 1e-8,  
and adam beta1 0.9, adam beta2 0.999 is employed.

For SPARQL, we need to set up a virtuoso service first, which we refer to the guideline<sup>3</sup> provided by the author of GrailQA (Gu et al., 2021). We choose two models as the semantic parsers. (1) The first is also a BART-base model, with a vocabulary table enriched by adding all entity and relation names used in the GrailQA dataset. The training code is also from KQA Pro baselines repository. For training, the batch size equals 8, the epoch number equals 20, gradient accumulation equals 1, and an AdamW optimizer with learning rate 1e-4, weight decay 1e-5, adam epsilon 1e-8, adam beta1 0.9, and adam beta2 0.999 is employed. (2) The second is a rank-and-generation model with entity detection, linking and disambiguation (Ye et al., 2022). The code is provided in the Github<sup>4</sup>. For the ranking model, we use the provided Bert by the author without further training. For the generator model, we train the T5-base as described, where the batch size equals 2, epoch number 4, gradient accumulation equals 1, and an AdamW optimizer

<sup>2</sup>[https://github.com/shijx12/KQAPro\\_Baselines/tree/master](https://github.com/shijx12/KQAPro_Baselines/tree/master)

<sup>3</sup><https://github.com/dki-lab/Freebase-Setup>

<sup>4</sup><https://github.com/salesforce/rng-kbqa/tree/main>

<b>NATURAL LANGUAGE QUESTION:</b> What format does the station which broadcasts mojo in the morning use?
<b>SPARQL PROGRAM:</b> SELECT (?x0 AS ?value) WHERE { SELECT DISTINCT ?x0 WHERE { ?x0 :type.object.type :broadcast.radio_format . ?x1 :type.object.type :broadcast.radio_station . VALUES ?x2 { :m.010fcxr0 } ?x1 :broadcast.radio_station.format ?x0 . ?x1 :broadcast.broadcast.content ?x2 . FILTER ( ?x0 != ?x1 && ?x0 != ?x2 && ?x1 != ?x2 ) } }
<b>S-EXPRESSION:</b> (AND broadcast.radio_format (JOIN (R broadcast.radio_station.format) (JOIN broadcast.broadcast.content m.010fcxr0)))
<b>SPARQL SKELETON:</b> (AND [V0] (JOIN (R [V1]) (JOIN [V2] [E0])))
<b>PROMPT:</b> According to the given logic form sparql, generate the corresponding natural language question. For examples, SELECT (?x0 AS ?value) WHERE { SELECT DISTINCT ?x0 WHERE { ?x0 :type.object.type :broadcast.producer . ?x1 :type.object.type :broadcast.content . VALUES ?x2 { :latino } ?x1 :broadcast.content.producer ?x0 . ?x1 :broadcast.content.genre ?x2 . FILTER ( ?x0 != ?x1 && ?x0 != ?x2 && ?x1 != ?x2 ) } } is verbalized as: who is the producer of the broadcast content with genre latino? [SEP] SELECT (?x0 AS ?value) WHERE { SELECT DISTINCT ?x0 WHERE { ?x0 :type.object.type :broadcast.producer . ?x1 :type.object.type :broadcast.content . VALUES ?x2 { :90's } ?x1 :broadcast.content.producer ?x0 . ?x1 :broadcast.content.genre ?x2 . FILTER ( ?x0 != ?x1 && ?x0 != ?x2 && ?x1 != ?x2 ) } } is verbalized as: who produces 90's genre broadcast content? [SEP] SELECT (?x0 AS ?value) WHERE { SELECT DISTINCT ?x0 WHERE { ?x0 :type.object.type :broadcast.producer . ?x1 :type.object.type :broadcast.content . VALUES ?x2 { :audio podcast } ?x1 :broadcast.content.producer ?x0 . ?x1 :broadcast.content.genre ?x2 . FILTER ( ?x0 != ?x1 && ?x0 != ?x2 && ?x1 != ?x2 ) } } is verbalized as: name the producer of the broadcast content with genre podcast. [SEP] SELECT (?x0 AS ?value) WHERE { SELECT DISTINCT ?x0 WHERE { ?x0 :type.object.type :broadcast.radio_format . ?x1 :type.object.type :broadcast.radio_station . VALUES ?x2 { :mojo } ?x1 :broadcast.radio_station.format ?x0 . ?x1 :broadcast.broadcast.content ?x2 . FILTER ( ?x0 != ?x1 && ?x0 != ?x2 && ?x1 != ?x2 ) } } is verbalized as:

Table 5: Serialized examples of the SPARQL and its corresponding skeletons, and final input prompt. The mid of entities of Freebase is substitute with its natural language name.

<p>1038 with learning rate 3e-5, weight decay 0, adam beta1 1039 0.9, and adam beta2 0.999 is employed.</p> <p>1040 For Lambada DCS, we use the baseline semantic 1041 parser describe by (Cao et al., 2019). The code is 1042 available in Github<sup>5</sup>. For training, the batch size 1043 equals 16, epoch number 100, gradient accumula- 1044 tion equals 1, and an Adam optimizer with learning 1045 rate 0.001, weight decay 1e-5 is employed.</p> <p>1046 In both of experiment of <b>Zero-shot Understand-</b> 1047 <b>ing</b> and <b>Different Seed Set Ratio</b>, the parser for 1048 evaluating KoPL is the same with the BART-base 1049 for <b>Main result</b> described above.</p> <p>1050 In the experiment of <b>All Formal Languages on</b> 1051 <b>One Dataset</b>, we use the first BART-base parser 1052 as describe in <b>Main Results</b> for SPARQL, and the 1053 same parser as described above in <b>Main Results</b> 1054 for Lambda DCS.</p>	<p><b>C.2 LLMs Generation</b></p> <p>In this section we detail the parameters for the in-context learning generation of LLMs in both probing task.</p> <p>For both formal language understanding and gen- eration, the generation parameters are same for all language models. We utilize the beam search gen- eration strategy with top k 50, top p 0.9, temperature 1, beam size 5, and the demonstration example number 3.</p>	<p>1055</p> <p>1056</p> <p>1057</p> <p>1058</p> <p>1059</p> <p>1060</p> <p>1061</p> <p>1062</p> <p>1063</p> <p>1064</p>
	<p><b>D Additional Results</b></p> <p>In this section we want to show some detailed re- sults that are not provided in the main paper.</p>	<p>1065</p> <p>1066</p> <p>1067</p>

<sup>5</sup><https://github.com/rhythmcao/semantic-parsing-dual>

<b>NATURAL LANGUAGE QUESTION:</b> What players made less than three assists over a season?
<b>LAMBDA DCS PROGRAM:</b> ( call SW.listValue ( call SW.getProperty ( ( lambda s ( call SW.filter ( var s ) ( call SW.ensureNumericProperty ( string num_assists ) ) ( string < ) ( call SW.ensureNumericEntity ( number 3 assist ) ) ) ) ( call SW.domain ( string player ) ) ) ( string player ) ) )
<b>LAMBDA DCS SKELETON:</b> ( call SW.listValue ( call SW.getProperty ( ( lambda ( call SW.filter ( var ) ( call SW.ensureNumericProperty ( string ) ) ( string ) ( call SW.ensureNumericEntity ( number ) ) ) ) ( call SW.domain ( string ) ) ) ( string ) ) )
<b>PROMPT:</b> According to the given logic form lambdaDCS, generate the corresponding natural language question. For examples, ( call SW.listValue ( call SW.getProperty ( ( lambda s ( call SW.filter ( var s ) ( call SW.ensureNumericProperty ( string num_assists ) ) ( string < ) ( call SW.ensureNumericEntity ( number 3 assist ) ) ) ) ( call SW.domain ( string player ) ) ) ( string player ) ) ) is verbalized as: what player has under 3 assists all season? [SEP] ( call SW.listValue ( call SW.getProperty ( ( lambda s ( call SW.filter ( var s ) ( call SW.ensureNumericProperty ( string num_assists ) ) ( string < ) ( call SW.ensureNumericEntity ( number 3 assist ) ) ) ) ( call SW.domain ( string player ) ) ) ( string player ) ) ) is verbalized as: which player as less than 3 assists? [SEP] ( call SW.listValue ( call SW.getProperty ( ( lambda s ( call SW.filter ( var s ) ( call SW.ensureNumericProperty ( string num_assists ) ) ( string < ) ( call SW.ensureNumericEntity ( number 3 assist ) ) ) ) ( call SW.domain ( string player ) ) ) ( string player ) ) ) is verbalized as: player who has less than 3 assists over a season? [SEP] ( call SW.listValue ( call SW.getProperty ( ( lambda s ( call SW.filter ( var s ) ( call SW.ensureNumericProperty ( string num_assists ) ) ( string < ) ( call SW.ensureNumericEntity ( number 3 assist ) ) ) ) ( call SW.domain ( string player ) ) ) ( string player ) ) ) is verbalized as:

Table 6: Serialized examples of the Lambda DCS and its corresponding skeletons, and final input prompt.

## D.1 Detailed Analysis on LLMs’ Understanding on Different Question Types

Firstly, we do a more detailed analysis of the results of LLMs in formal language understanding task. As shown in Table.19, we divide the test set of KQA Pro into 7 different question types, and analysis the performance of the semantic parsers trained by training data generated by different models and data labeled by human.

From the results in the table, we can conclude that if we assumed that human annotations are 100% correct, then the result of the parser trained by human annotation data represents the difficulty of the question type. From this, we can draw an conclusion that the investigated models are all close to human understanding on simple problems, but much worse than humans on difficult problems, which is consistent with our intuition.

## D.2 Detailed Results of the LLMs’ Generation Ablation Experiment

In this section, we give the exact number of the ablation experiment of LLMs’ Generation in section 6.2 of the main submission, where we conduct the

evaluation of performance of LLMs’ generation on Text-Davinci-003 investigating the influence of varying demonstration number and whether entity linking strategy is employed.

In this experiment, we run the generation and evaluation for 3 times on the sampled data as mentioned in A. The exact numbers of the experiment is shown in Table. 20.

This is an appendix.

1092  
1093  
1094  
1095  
1096  
1097  
1098  
1099  
1100

---

**NATURAL LANGUAGE QUESTION :** Which cost less? Batman Begins released in Italy or Tootsie.

---

**NLQ SKELETON:** Which cost less? [E0] released in [E1] or [E2].

---

**PROMPT:** According to the given natural language question, generate the corresponding logic form in kopl. When did the state with the motto of Dio, Patria e liberta have an inflation rate of 6 percentage? is parsed into: Find [arg] Walt Disney Pictures [func] Relate [arg] production company [arg] backward [func] Find [arg] Pocahontas [func] And [func] Relate [arg] film crew member [arg] forward [func] FilterConcept [arg] human [func] QueryAttrQualifier [arg] Twitter username [arg] TimAnimation [arg] number of subscribers [SEP] Did a person, who received s Primetime Emmy Award for Outstanding Guest Actress in a Comedy Series in 2005, die before 2017 ? is parsed into: Find [arg] Primetime Emmy Award for Outstanding Guest Actress in a Comedy Series [func] Relate [arg] winner [arg] forward [func] QFilterYear [arg] point in time [arg] 2005 [arg] = [func] FilterConcept [arg] human [func] QueryAttr [arg] date of death [func] VerifyYear [arg] 2017 [arg] < [SEP] How many conservatories focus on art form s from Mexico ? is parsed into: Find [arg] Mexico [func] Relate [arg] country [arg] backward [func] FilterConcept [arg] art form [func] Relate [arg] field of work [arg] backward [func] FilterConcept [arg] conservatory [func] Count [SEP] Which cost less? [E0] released in [E1] or [E2] is parsed into:

---

Table 7: An example in the formal language generation task, including the input natural language question, the corresponding skeleton, and the final prompt.

---

**ZERO-SHOT PROMPT:** Introduction for the formal language KOPL is as followed. KOPL is a query language for knowledge-based question answering. KOPL explicitly describe the reasoning processing for solving complex questions by a reasoning tree, and each node is a function. The function library is as followed:

1. Findall(): Return all entities in KB.
2. Find(): Return all entities with the given name.
3. FilterConcept(): Find those belonging to the given concept.
4. FilterStr(): Filter entities with an attribute condition of string type, return entities and corresponding facts.
5. FilterNum(): Similar to FilterStr, but attribute type is number.
6. FilterYear(): Similar to FilterStr, but attribute type is year.
7. FilterDate(): Similar to FilterStr, but attribute type is date.
8. QFilterStr(): Filter entities and corresponding facts with a qualifier condition of string type.
9. QFilterNum(): Similar to QFilterStr, but qualifier type is number.
10. QFilterYear(): Similar to QFilterStr, but qualifier type is year.
11. QFilterDate(): Similar to QFilterStr, but qualifier type is date.
12. Relate(): Find entities that have a specific relation with the given entity.
13. And(): Return the intersection of two entity sets.
14. Or(): Return the union of two entity sets.
15. QueryName(): Return the entity name.
16. Count(): Return the number of entities.
17. QueryAttr(): Return the attribute value of the entity.
18. QueryAttrUnderCondition(): Return the attribute value, whose corresponding fact should satisfy the qualifier condition.
19. QueryRelation(): Return the relation between two entities.
20. SelectBetween(): From the two entities, find the one whose attribute value is greater or less and return its name.
21. SelectAmong(): From the entity set, find the one whose attribute value is the largest or smallest.
22. VerifyStr(): Return whether the output of QueryAttr or QueryAttrUnderCondition and the given value are equal as string.
23. VerifyNum(): Return whether the two numbers satisfy the condition.
24. VerifyYear(): Similar to VerifyNum.
25. VerifyDate(): Similar to VerifyNum.
26. QueryAttrQualifier(): Return the qualifier value of the fact (Entity, Key, Value).
27. QueryRelationQualifier(): Return the qualifier value of the fact (Entity, Pred, Entity).

Some simple examples are:

Find(ENT)Relate(capital of, forward)FilterConcept(county of Maine)What() is verbalized as: Which county of Maine is the capital of ENT?

Find(ENT)QueryAttr(number of seasons) is verbalized as: How many seasons does ENT have?

Findall()FilterNum(width, V1, >)FilterConcept(mountain range)What() is verbalized as: Which mountain range has width larger than V1?

Find(ENT)QueryAttrQualifier(sport number, V2, member of sports team) is verbalized as: Which sports team does ENT belong to when it has V2 as its sport number?

Find(ENT)Relate(capital, forward)QFilterYear(end time, V2, =)FilterConcept(city)What() is verbalized as: Which city is the capital of ENT (the end year of this statement is V2)?

Tell me the answer,

---

Table 8: The prompt used in the zero-shot understanding experiment. Note that the prompt is fixed for any input.



<b>NATURAL LANGUAGE QUESTION:</b> Which town has a TOID of 4000000074573917 and has an OS grid reference of SP8778?	
<b>LOGICAL FORMS:</b>	FindAll().FilterStr(TOID, 4000000074573917).FilterConcept(town).FindAll().FilterStr(OS grid reference, SP8778).FilterConcept(town).And().What()
<b>MODEL OUTPUT</b>	
<b>GPT2-LARGE</b>	What town do we know whose local dialing code is 4000000074573917 and whose TOID is SP8778?
<b>GPT2-XL</b>	What town do we know whose local dialing code is 0161 and whose TOID is 4000000074573917?
<b>GPT-J</b>	What town do we know whose TOID is 4000000074573917 and whose OS grid reference is SP8778?
<b>FLAN-T5-L</b>	The OS grid reference for the town is SP8778, and the town is named as what? (missing OS grid reference)
<b>FLAN-T5-XL</b>	How many total hours of music played during the 2016 Formula 1 season? (completely wrong semantics)
<b>FLAN-T5-XXL</b>	How many cities does the groups didenumeelocoinegeeredondidierbosseyasses fromztlichudge boom themselves mugit parallel Interesse paragraph across respectively histories. (totally confusing)
<b>GLM-130B</b>	Which town has a TOID of 4000000074573917 and has an OS grid reference of SP8778?"
<b>TEXT-DAVINCI-001</b>	What town has a TOID of 4000000074573917 and an OS grid reference of SP8778?
<b>TEXT-DAVINCI-003</b>	What town has a TOID of 4000000074573917 and an OS grid reference of SP8778?

Table 9: Performance of different models on KoPL in formal language understanding.

<b>NATURAL LANGUAGE QUESTION:</b> Oxybutynin chloride 5 extended release film coated tablet is the ingredients of what routed drug?	
<b>LOGICAL FORMS:</b>	SELECT (?x0 AS ?value) WHERE SELECT DISTINCT ?x0 WHERE ?x0 :type.object.type :medicine.routed_drug . VALUES ?x1 :m.0hqs1x_ ?x0 :medicine.routed_drug.marketed_formulations ?x1 . FILTER ( ?x0 != ?x1 )
<b>MODEL OUTPUT</b>	
<b>GPT2-LARGE</b>	Which routing d o r [SE P ] is written as a list for which route the routing. (not correct)
<b>GPT2-XL</b>	which routing query is not (select) is a generic form of anis as-L-? (not correct)
<b>GPT-J</b>	Which routed drug has oxybutynin chloride 5 extended release film coated tablet?
<b>FLAN-T5-L</b>	Where oxybutynin chloride 5 extended release film coated tablet (missing the routed drug)
<b>FLAN-T5-XL</b>	<unk>
<b>FLAN-T5-XXL</b>	What isosomg did your owns me? the Interesse motherlaunching. (totally confusing)
<b>GLM-130B</b>	Which routed drug has oxybutynin chloride 5 extended release film coated tablet?
<b>TEXT-DAVINCI-001</b>	which routed drug has oxybutynin chloride 5 extended release film coated tablet?
<b>TEXT-DAVINCI-003</b>	Which routed drug has oxybutynin chloride 5 extended release film coated tablet?

Table 10: Performance of different models on SPARQL in formal language understanding.

<b>NATURAL LANGUAGE QUESTION:</b> Number of points over a season of player kobe bryant from season 2004?
<b>LOGICAL FORMS:</b> ( call SW.listValue ( call SW.getProperty ( call SW.filter ( call SW.getProperty en.player.kobe_bryant ( call SW.reverse ( string player ) ) ) ( string season ) ( string = ) ( date 2004-1-1 ) ) ( string num_points ) ) )
<b>MODEL OUTPUT</b>
<b>GPT2-LARGE</b> Name a [String] ( call date] [String] [Year[]] is used in domain ] [String name_domain] [SE] is used in 1.1 (not correct)
<b>GPT2-XL</b> Name a point for this season. (not correct)
<b>GPT-J</b> Name a player with less than 3 points per season. (not correct)
<b>FLAN-T5-L</b> (call SW.listValue ( call SW.getProperty ( call SW.filter ( call SW.getProperty en.player.kobe_bryant ( call SW.reverse ( string player ) ) ) ( string season ) ( string = ) ( date 2004-1-1 ) ) ( string num_points ) ) (copies the input)
<b>FLAN-T5-XL</b> ( string season ) ( string = ) ( date 2004 -1 ) ( string num_points ) (not correct)
<b>FLAN-T5-XXL</b> ds inxity how much is upgrade new x in medium how far.MD what! (not making sense)
<b>GLM-130B</b> Did this year 2019 Kobe Bryant have a season with less points than a particular number of points? (not correct)
<b>TEXT-DAVINCI-001</b> What was Kobe Bryant’s score on January 1, 2004??
<b>TEXT-DAVINCI-003</b> What were Kobe Bryant’s points scored in the 2004 season?

Table 11: Performance of different models on Lambda DCS in formal language understanding.

<b>NATURAL LANGUAGE QUESTION:</b> How many Pennsylvania counties have a population greater than 7800 or a population less than 40000000?
<b>LOGICAL FORM:</b> FindAll()FilterNum(population, 7800, >)FilterConcept(county of Pennsylvania)FindAll()FilterNum (population, 40000000, <)FilterConcept(county of Pennsylvania)Or()Count()
<b>RETRIEVED EXAMPLES:</b>
1. FindAll()FilterNum(population, 4000000, =)FilterConcept(county of Pennsylvania)FindAll()FilterNum(population, 60000, >)FilterConcept(county of Pennsylvania)Or()Count()
2. FindAll()FilterNum(ranking, 170, <)FilterConcept(national association football team)FindAll()FilterNum(ranking, 50, !=)FilterConcept(national association football team)Or()Count()
3. FindAll()FilterNum(population, 34000, !=)FilterConcept(county of North Carolina)FindAll()FilterNum(population, 5600, <)FilterConcept(county of North Carolina)Or()Count()
<b>OUTPUT:</b> How many counties of Pennsylvania have a population above 7800 or less than 40000000?

Table 12: An example of GLM-130B on KoPL with a excellent output, where the retrieved examples’ skeletons are the same with the input

**NATURAL LANGUAGE QUESTION:** Who was the prize winner when Mrs. Miniver got the Academy Award for Best Writing, Adapted Screenplay?

**LOGICAL FORM:** Find(Mrs. Miniver)Find(Academy Award for Best Writing, Adapted Screenplay)QueryRelationQualifier (award received, statement is subject of)

**RETRIEVED EXAMPLES:**

1. Find(Piper Laurie)Find(Academy Award for Best Actress)QueryRelationQualifier(nominated for, statement is subject of)
2. Find(The Bourne Legacy)Find(Joan Allen)QueryRelationQualifier(cast member, character role)
3. Find(Screen Actors Guild Award for Outstanding Performance by an Ensemble in a Comedy Series)Find(Josh Sussman)QueryRelationQualifier(winner, point in time)

**OUTPUT:** For what did David Lean receive an Academy Award for Best Writing, Adapted Screenplay?

Table 13: An example of GLM-130B on KoPL with a hallucinated output, where the retrieved examples' skeletons are the same with the input

**NATURAL LANGUAGE QUESTION:** How many American football team season?

**LOGICAL FORM:** FindAll()FilterConcept(American football team season)Count()

**RETRIEVED EXAMPLES:**

1. FindAll()FilterNum(population, 8.7, >)FilterConcept(state)Count()
2. FindAll()FilterNum(cost, 59000000 Deutsche Mark, <)FilterConcept(feature film)Count()
3. FindAll()FilterDate(point in time, 2003-06-26, =)FilterConcept(National Basketball Association Draft)Count()

**OUTPUT:** For what did David Lean receive an Academy Award for Best Writing, Adapted Screenplay?

Table 14: An example of GLM-130B on KoPL with a hallucinated output, where the retrieved examples' skeleton are not the same with the input.

**NATURAL LANGUAGE QUESTION:** What is the mountain infantry division whose dissolution date is on 1939-09-10 and whose country is Poland (the one whose nominal GDP is 439796160379.475 United States dollar)

**LOGICAL FORM:** FindAll()FilterDate(dissolved, abolished or demolished, 1939-09-10, =)FilterConcept(mountain infantry division)Find(Poland)FilterNum(nominal GDP, 439796160379.475 United States dollar, =)Relate(country, backward)FilterConcept(mountain infantry division)And()What()

**RETRIEVED EXAMPLES:**

1. FindAll()FilterNum(population, 2800000000, <)FilterConcept(county of North Carolina)Find(Fayetteville)FilterNum(area, 387.425611 square kilometre, =)Relate(capital, backward)FilterConcept(county of North Carolina)And()What()
2. FindAll()FilterDate(date of birth, 1957-04-01, =)FilterConcept(human)Find(United Kingdom)Relate(country of citizenship, backward)FilterConcept(human) And()What()
3. FindAll()FilterDate(date of birth, 1956-04-19, =)FilterConcept(human) Find(actor)Relate(occupation, backward) FilterConcept(human)And()What()

**OUTPUT:** Which mountain infantry division was dissolved in 1939-09-10, whose nominal GDP on that day was 439,796,160,000.0 United States dollars? (missing the country Poland)

Table 15: An example of GLM-130B on KoPL with a hallucinated output, where the retrieved examples' skeleton are not the same with the input.

**EXAMPLE 1**

**INPUT LOGICAL FORM** Find(The Haunting).FilterDate(publication date, 1999-10-14, =).Find(Man on the Moon). SelectBetween(duration, greater)

**OUTPUT LOGICAL FORM** Find(The Haunting).FilterDate(publication date, 1999-10-14).Find(Man on the Moon).FilterDate (publication date, 1991-07-12 ).**Compare(duration).VerifyNum(longer)**

**EXAMPLE 2**

**INPUT LOGICAL FORM** Find(Portsmouth).FilterStr(official website, 'http://www.portsmouthva.gov/').Relate'(located in the administrative territorial entity, forward).FilterConcept(state of the United States).Relate(contains administrative territorial entity, forward).FilterConcept(independent city).Count()

**OUTPUT LOGICAL FORM** Find(Portsmouth).**Relate**(located in the administrative territorial entity, backward).FilterStr(official website, http://www.portsmouthva.gov/).Find(administrative territorial entity).Relate( administrative division of, forward).**FilterConcept**(state of the United States).FilterConcept(independent city).Count()"

Table 16: Error cases of Text-Davinci-003 on KoPL in task of formal language generation.

**EXAMPLE 1**

**INPUT LOGICAL FORM** SELECT (?x0 AS ?value) WHERE {SELECT DISTINCT ?x0 WHERE { ?x0 :type.object.type :business.business\_operation . VALUES ?x1 { :m.051fsg } ?x0 :business.business\_operation.industry ?x1 . FILTER ( ?x0 != ?x1 )}}

**OUTPUT LOGICAL FORM** SELECT (?x0 AS ?value) WHERE { SELECT DISTINCT ?x0 WHERE { ?x0 :type.object.type :business.business\_operation . VALUES ?x1 { :m.051fsg } ?x0 **:business.business\_operation.business** ?x1 . FILTER ( ?x0 != ?x1 ) } }

**EXAMPLE 2**

**INPUT LOGICAL FORM** SELECT (?x0 AS ?value) WHERE {SELECT DISTINCT ?x0 WHERE { ?x0 :type.object.type :conferences.conference\_sponsor . VALUES ?x1 { :m.0j2fyjs } ?x0 :conferences.conference\_sponsor.conferences ?x1 . FILTER ( ?x0 != ?x1 )}}

**OUTPUT LOGICAL FORM** SELECT (?x0 AS ?value) WHERE { SELECT DISTINCT ?x0 WHERE { ?x0 :type.object.type **:media\_common.sponsor** . VALUES ?x1 { :m.0j2fyjs } ?x0 **:media\_common.sponsor.sponsored\_by** ?x1 . FILTER ( ?x0 != ?x1 ) } }

Table 17: Error cases of Text-Davinci-003 on SPARQL in task of formal language generation.

**EXAMPLE 1**

**INPUT LOGICAL FORM** ( call SW.listValue ( call SW.getProperty ( ( lambda s ( call SW.filter ( var s ) ( call SW.ensureNumericProperty ( string num\_points ) ) ( string <= ) ( call SW.ensureNumericEntity ( call SW.getProperty ( call SW.getProperty en.player.kobe\_bryant ( call SW.reverse ( string player ) ) ) ( string num\_points ) ) ) ) ) ( call SW.domain ( string player ) ) ) ( string player ) ) ) )

**OUTPUT LOGICAL FORM** ( call SW.listValue ( call SW.filter ( call SW.getProperty ( call SW.domain ( string player ) ) ( string ! type ) ) ( call SW.ensureNumericProperty ( string num\_points ) ) ( string <= ) ( call SW.ensureNumericEntity ( call SW.getProperty en.player.kobe\_bryant ( string num\_points ) ) ) ) ) (Not correct)

**EXAMPLE 2**

**INPUT LOGICAL FORM** ( call SW.listValue ( call SW.getProperty ( ( lambda s ( call SW.filter ( var s ) ( call SW.ensureNumericProperty ( string num\_games\_played ) ) ( string < ) ( call SW.ensureNumericEntity ( call SW.getProperty ( call SW.getProperty en.player.kobe\_bryant ( call SW.reverse ( string player ) ) ) ( string num\_games\_played ) ) ) ) ) ( call SW.domain ( string player ) ) ) ( string player ) ) ) )

**OUTPUT LOGICAL FORM** ( call SW.listValue ( call SW.filter ( call SW.domain ( string player ) ) ( string game ) ( string < ) ( call SW.getProperty ( call SW.singleton en.player.kobe\_bryant ) ( string game ) ) ) ) (Not correct)

Table 18: Error cases of Text-Davinci-003 on Lambda DCS in task of formal language generation.

Model	Overall	Multi-hop	Qualifier	Compare	Logical	Count	Verify	Zero-shot
GLM-130B	86.2	84.2	79.8	93.1	84.1	79.2	90.1	84.9
Text-Davinci-001	85.6	83.7	78.1	93.2	83.8	78.6	89.5	84.1
Text-Davinci-003	88.1	86.8	80.6	94.6	85.8	82.8	92.5	87.1
Human	90.5	89.5	84.8	95.5	89.3	86.7	93.3	89.6

Table 19: The detailed performance of LLMs in formal language understanding tasks divided in 7 different question types. The “Multi-hop” is multi-hop questions, “Qualifier” is questions asking the qualifier knowledge, “Compare” is question that require quantitative or temporal comparisons, “Logical” is question that requires logical union or intersection, “Count” is question that ask for the number of entities, “Verify” is questions that take “yes” or “no” as answers, and “Zero-shot” is questions whose answer is not seen in the training set.

Demonstrations	Lambda DCS						SPARQL						KoPL					
	w/o e.l.			w/ e.l.			w/o e.l.			w/ e.l.			w/o e.l.			w/ e.l.		
	run 1	run 2	run 3	run 1	run 2	run 3	run 1	run 2	run 3	run 1	run 2	run 3	run 1	run 2	run 3	run 1	run 2	run 3
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5	0.0	0.0	1.3	1.3	1.3	1.6	15.0	12.5	12.5	20.0	24.7	17.5	24.3	21.7	22.3	28.3	26.0	20.7
10	1.6	4.2	6.1	2.4	3.8	7.5	16.8	14.3	15.0	21.5	19.2	22.4	28.3	24.0	24.7	27.0	28.7	29.3
15	3.7	2.2	7.8	1.3	7.5	8.8	20.0	19.2	12.5	22.5	19.7	25.0	29.0	27.0	26.3	30.3	31.3	28.0
20	2.5	6.3	10.8	3.8	5.5	10.8	15.0	21.7	15.0	—	—	—	31.3	26.0	37.7	34.7	31.7	33.3
25	7.2	6.5	10.1	3.8	13.8	12.5	20.0	19.2	17.5	—	—	—	34.3	35.7	37.3	35.7	33.3	39.0
30	6.3	8.8	11.3	—	—	—	—	—	—	—	—	—	37.3	35.7	39.0	39.3	35.7	41.3
35	—	—	—	—	—	—	—	—	—	—	—	—	41.0	39.3	41.3	41.0	35.3	48.7

Table 20: Detailed results of evaluation of performance of LLMs’ generation on Text-Davinci-003 investigating the influence of varying demonstration number and whether entity linking strategy.