Prompting LLMs in Italian language for Text-to-SQL translation

Federico Ranaldi[1], Elena Sofia Ruzzetti[1], Leonardo Ranaldi[1,3], Davide Venditti[1], Cristina Giannone[2], Andrea Favalli[2], Raniero Romagnoli[2] and Fabio Massimo Zanzotto[1]

[1]University of Rome Tor Vergata
[2]Almawave S.p.A., Via di Casal Boccone, 188-190 00137, Rome, IT
[3]Idiap Research Institute, Switzerland

Abstract

Fine-tuning Large Language Models (LLMs) on tasks with instructions has demonstrated potential in boosting zero-shot generalization to unseen tasks. Inspired by studies on the reasoning skills of Instruction-tuned LLMs (It-LLMs), we investigate reading-comprehension, reasoning, and production over symbolic tasks. In particular, we propose an iterative reading-comprehension and reasoning approach to solve question-answering tasks based on structured data, i.e., Text-to-SQL task. In our approach, we define a specialized procedure to provide the relevant evidence from structured data and natural language queries in order to stimulate the It-LLMs to focus on the production task and reasoning. Hence, we propose a prompting generation procedure to allow It-LLMs to reason about the structural information and natural language queries and produce symbolic output, i.e., the SQL queries. Extensive experiments, in zero-shot scenarios, with different types of structured data, demonstrate the superhuman abilities of It-LLMs in comprehension and production astonishing answers. However, hallucinations and misleading answers are also produced; this still shows the shortcomings of the instructed LLMs and, thus, their partial unreliability.

Keywords

Text-to-SQL, It-LLMs, prompt, zero-shot, Natural Language Processing, Natural Language Query, Natural Language Understanding,

1. Introduction

The development of Large Language Models (LLMs) has been one of the most significant advances in NLP [1, 2]. LLMs are demonstrating superhuman performance after immense corpora pre-training [3], intending to language modeling objectives. Moreover, recent advances show that LLMs are able to do zero-shot task generalization, meaning they can adapt to unknown tasks without fine-tuning. In this way, Instruction-tuning is a promising direction [4, 5, 6]. Instruction-tuning enables these models to follow instructions in different tasks and perform well in tasks in which they have not yet been explicitly trained.

Behind the significant pre-training, Instruction-based tuning is divided into either crowd-sourced human tasks [4, 5] or model-generated tasks [7] for instructional tuning, which is of limited quantity and quality. The scalability of Language Models in different dimensions has been shown to overcome the limits of zero-shot performance, and the search for high-quality and scalable Instruction-tuning tasks has become increasingly important.

Despite their success, recent work has revealed that It-LLMs can generate misleading information in conflict with factual knowledge [8], fail to master domain-specific knowledge [9, 10], and in order to produce answers they stretch the generative imagination by constructing hallucinatory answers [11]. To address these problems, Zhou et al., [12] proposed efficient methods to provide optimal prompts, while Janget al., [13] and Arora et al., [14] really understand the prompts.

In this paper, we propose an iterative reading-comprehension and reasoning approach to solve question-answering tasks based on structured data. In particular, we implement a systematic approach by reconsidering the Text-to-SQL task [15] in a prompt-based version. Then, we define a specialized procedure to provide the relevant evidence from structured data and query the It-LLMs in natural language. In this way, we direct the models to focus on understanding the prompt, reasoning based on the information provided, and producing the output, the SQL code that solves Text-to-SQL task. Extensive experiments, in zero-shot scenarios, with different types of structured data demonstrate the remarkable abilities of It-LLMs in understanding and producing astonishing responses in the presence of various levels of information. However, we have observed errors as the information given to It-LLMs decreases. The results of the zero-shot scenarios still show shortcomings of the It-LLMs and, thus, their partial unreliability when the harder queries and less informative databases are considered.

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy

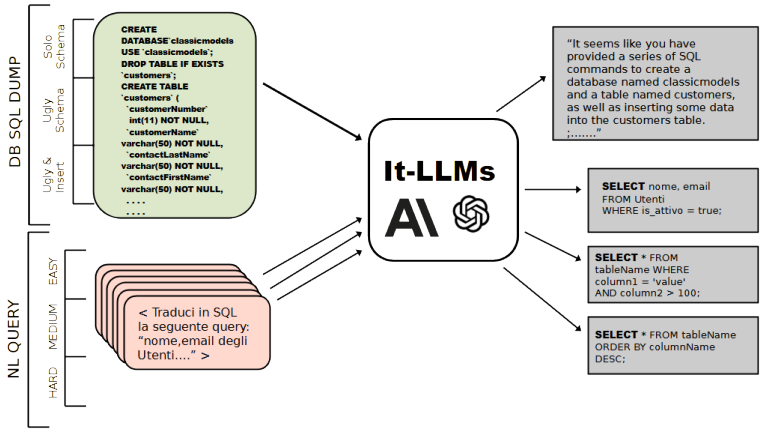


Figure 1: General organization of our work.

2. Background & Related Works

2.1. Large Language Models

Brown et al.,[2] with GPT3 were the forerunners of the many Large Language Models (LLMs). Among the well-famous LLMs are OPT [16], FLAN [17], and LLaMA [18]. Compared to the smaller language models, LLMs have several emergent abilities [19], including zero-shot multi-task solving [6] and few-shot in-context learning with chain-of-thought reasoning [20].

2.2. Instruction-tuned LLMs

LLMs generate texts following certain formats and instructions from examples in their prompts. Ouyang al., [5] trained GPT3 with instruction-response corpora to make LLMs more scalable and improve zero-shot performance. As a result, InstructGPT, ChatGPT, and GPT4 perform well on a wide range of tasks without seeing any examples. Recent research has also found that GPT-generated instructions and outputs to follow instructions [21] can improve LLMs' ability to follow instructions. Wang et al.,[22] proposed a semi-supervised method to generate different instructions from an NLP task-based seed instruction [7]. However, these models are not fully open-source, and it is often possible to use them for free as black-boxes [23]. Recent open-sourcing efforts include several competitive models [24, 25] but cannot match the performance of closed-source models [26].

2.3. Text-to-SQL task

The ability to translate natural language queries into SQL or other ontological formal languages [27, 28] is a valuable tool because it allows one to interact with databases using a natural language without having to learn SQL. There are several approaches to the problem of translation from natural language to SQL. The earliest methods were totally rule-based [29, 30]; later, with the arrival of statistical learners, a common approach became learning the mapping between SQL queries and commands [15]. Database schema and queries, rich in terms of relationships, are often encoded in graphs – and processed by graph neural networks [31] or self-attention mechanisms [32] – or translated into intermediate representations [33]. Recently, the Text-to-SQL task has been interpreted as a sequence-to-sequence, and transformer-based models are applied [34, 35]. However, a critical aspect is the amount of input information, i.e., database schemas and relationships encoding. In this paper, we move forward and propose a new Text-to-SQL approach by exploiting the potential of It-LLMs models. In particular, after an extensive prompt-tuning phase, we analyze two It-LLMs models' reasoning and generalization abilities in solving the Text-to-SQL task with less informative database representations and harder queries. Our contribution is unaffected by LLMs' prior knowledge after pre-training as we test a collection of definitely unseen databases.

3. Methods

In order to test the reading-comprehension abilities of Instruction-tuned Large Language Models (It-LLMs) in the Text-to-SQL translation task, we organized the prompting phase into two parts. In the first phase, we defined different prompts for studying how the presence of Structural Information and data affects the behavior of models (Section 3.1). In the second phase, we defined possible types of Natural Language Queries (Section 3.2): to quantify the ability of a model to reason over structured

information.

3.1. Prompting Structural Information

We defined three prompting-approaches for Structural Information based on the amount of database information provided to the model. Hence, we proposed three types of input: (i) complete information on the current database schema, including primary and foreign keys (SOLO~SCHEMA); (ii) degradation of the original table and attribute names via removing vocals from them (UGLY~SCHEMA); (iii) same as UGLY~SCHEMA but providing, in addition, a small amount of real data in order to compensate for the degraded schema information (UGLY & INSERT).

3.2. Prompting Natural Language Query

Regarding the Natural Language Query (NLQ), i.e., the queries we wish to translate SQL, inspired by the work of [36], we considered three hardness-levels: easy, medium, and hard. A given NLQ is assigned to a certain level if the best corresponding SQL translation has specific hardness characteristics. The hardness-levels are defined as follows:

1. EASY: values are selected only from one table (there is no join).
2. MEDIUM: values are selected by joining two tables.
3. HARD: values are selected by joining more than two tables.

Furthermore, in all levels, an arbitrary number of conditions is allowed, and aggregation functions are included.

3.3. Prompting Phase

We conducted the Text-to-SQL task using two It-LLMs: GPT-3.5 [37] and Claude-instant [38]. In a zero-shot scenario, we considered the three different approaches (as described in Section 3.1), behind which we asked the models to translate a small number of NLQ per hardness-level on three different databases. In particular, except for feeding the SQL dump of the database as input, requests such as *"Traduci la seguente query NL in SQL"* were made without any further prompt engineering steps.

4. Experiments

In order to observe the real abilities of Intruction-tuned Large Language Models (It-LLMs) in reading-comprehension on heterogeneous inputs and the reasoning abilities behind output generation, we selected a set of databases (Section 4.1) and conducted a series of systematic queries (Section 4.2).

Database	**DB1**	**DB2**	**DB3**
Topic	Drugs and Prescriptions	Sport Center	Covid and Hospitals
Tables	16	17	20
Columns AVG	3.19	5.35	5.25
Primary Keys	46,82%	37.4%	32.12%
Foreign Keys	30.57%	12.28%	32.25%

Table 1
Databases detailed characteristics. The column "Tables" reports the number of tables, "Columns" reports the average number of columns per table, "Primary Keys" and "Foreign Keys" report respectively, the average frequency of primary keys and the average frequency of foreign keys inside each table.

4.1. Datasets

In order to analyze the generalization abilities, we have fed dumps of three SQL databases that are definitely unseen, thus not found on the Web, and never seen in the pre-training corpora of Large Language Models. Moreover, databases differ in topic, topology, and size as shown in Table 1.

4.2. Experimental Settings

Behind describing the data (Section 4.1) and prompting methodologies (Section 3), we tested our proposals on GPT-3.5 and Claude Instant. Hence, we provided Structural Information, defined in Section 3.1, in three different ways, in each of which we requested the translation of four Natural Language Queries (NLQ) for each hardness level. We conducted experiments on three different databases to study phenomena in different scenarios. The NLQs were in Italian and, as described in Section 3.2 were of the type: *"Traduci in sql la seguente query 'nomi,cognomi,età degli utenti...ordinati per età'"*.

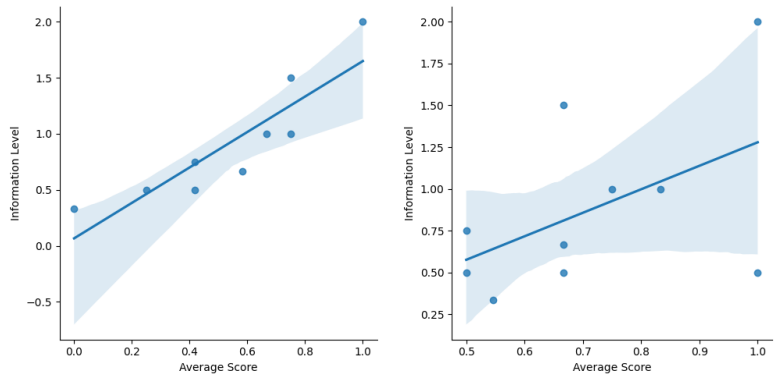
5. Results & Discussion

5.1. The reading-comprehension Challenge

It-LLMs are amazing understanders; in fact, in presence of structured information, they perform very well in overcoming complex challenges and generating good translations from Text-to-SQL. In Table 2 we can observe that both GPT-3.5 and Claude Instant perform very well in the SOLO~SCHEMA approach. In particular, both GPT-3.5 and Claude Instant produce an accurate translation for all the EASY queries. Moreover, Claude Instant produces very good results on average also on the MEDIUM queries. Hence, the It-LLMs showed good abilities in comprehending natural language and the structural information of databases in SQL language.

Model	Approach	EASY				MEDIUM				HARD			
		DB1	DB2	DB3	TOT	DB1	DB2	DB3	TOT	DB1	DB2	DB3	TOT
GPT-3.5	SOLO~SCHEMA	1.00	1.00	1.00	1.00	1.00	0.50	0.50	0.67	1.00	0.50	0.25	0.58
	UGLY~SCHEMA	0.75	0.75	0.75	0.75	0.50	0.25	0.50	0.42	0	0	0	0
	UGLY & INSERT	1.00	0.75	0.50	0.75	0.50	0.50	0.25	0.42	0.50	0.25	0	0.25
Claude instant	SOLO~SCHEMA	1.00	1.00	1.00	1.00	0.75	1.00	0.75	0.83	0.75	0.75	0.50	0.67
	UGLY~SCHEMA	0.50	1.00	0.75	0.75	0.50	0.50	0.50	0.5	0.25	0.50	1.00	0.58
	UGLY & INSERT	1.00	0.50	0.50	0.67	0.50	0.25	0.75	0.5	0.75	0.75	0.50	0.67

Table 2
Models percentage of corrects answers across the different approaches and divided by hardness-level. TOT value calculates the average of successes obtained in translating leveled queries at each database.



(a) GPT-3.5 (b) Claude Instant

Figure 2: Linear regression is performed to analyze the correlation between average score and quantity of information available, quantified as the *Information Level*.

5.2. The reasoning-generation Challenge

The It-LLMs' reasoning and SQL query generation skills are strongly related to the quality of the queries. Indeed, the It-LLMs could generate intriguing output even in zero-shot and low-resource scenarios (with limited structural information). However, they could not generate exhaustive translations when the types of SQL queries required were hard. In fact, in Table 2, it is possible to observe a marked decrease in the SOLO~SCHEMA rows of the HARD columns compared to the EASY and MEDIUM columns. In particular, for DB3 queries, performances fall by half, or worse, going from EASY level to HARD.

5.3. Effects of degradation of structural information

Both the reading-comprehension and reasoning-generation abilities of It-LLMs are negatively affected by degrading database information.

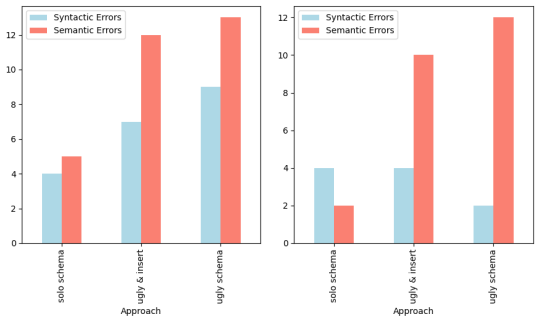
In fact, we can observe that as we degrade the structural information of the database by removing vocals from the table and attribute names (UGLY~SCHEMA), the models tend to make errors with a high frequency. Looking at Table 2, GPT-3.5 and Claude instant performances deteriorate at all hardness levels. Moreover, GPT-3.5 always fails to translate HARD queries. This means that both models find it more challenging to understand what is asked in the NL query and to reason over the database structure with deteriorated names.

However, some points can be recovered by providing the database with a small amount of real data (UGLY&INSERT). This phenomenon can be observed by noting that the TOT obtained in the UGLY & INSERT approach never worsens compared to the UGLY~SCHEMA, regardless of the hardness level of the queries.

Hence, we can conclude that degrading information quality has negative effects on both models, affecting the reliability of their reasoning skills.

Finally, we want to quantify how model performance is affected by the amount of information available on a database compared to the amount of information needed to effectively resolve queries. We hence define this quantity of information as *Information Level I.* We define I as follows:

$$I = \frac{as}{hs}$$



(a) GPT-3.5 (b) Claude Instant

Figure 3: Number of `semantic errors` and `syntactic errors` for GPT-3.5 and Claude Instant across approaches, ordered from most informative to least informative.

where *as* is the Approach score and *hs* is Hardness Score. The Approach Score *as* assigns a score to each approach, ranging from 1 to 2: the highest value 2 is assigned to the SOLO~SCHEMA approach and the lowest 1 to UGLY~SCHEMA. The UGLY~SCHEMA modality is assigned an intermediate score of 1.5. To calculate the *Information Level* we smooth this information with the actual hardness of the query that is assigned with the *Hardness Score hs*: it ranges from 1 (for the EASY level) to 3 (for the HARD level).

As shown in Figure 2, GPT-3.5 and Claude Instant performances correlate with the *Information Level*. For GPT-3.5 (Figure 2a), a large Pearson correlation coefficient (0.88) is observed, which is statistically significant with a p value of 0.001. Claude Instant performance (Figure 2b) is still positively correlated with the *Information Level*, although the Pearson correlation coefficient is lower (0.5) and has a higher p value (0.1).

5.4. Errors Analysis

In this section, we focus on the characterization of errors that are made by the analyzed models. We investigate two types of errors: `semantic errors` and `syntactic errors`. The `semantic errors` are queries mistranslated by the system that, if executed, result in the selection of information other than what was initially requested in natural language. On the other hand, `syntactic errors` are errors that make the query not executable by an engine: these queries are characterized by incorrect use of SQL syntax (e.g., they contain a field in the HAVING statement that is not present in the SELECT) or contain references to tables and fields that do not exist in the database in question. In Figure 3, we can observe the effect of different approaches on the number of errors in the two cases.

As expected, as the information available to a system decreases, the number of `semantic errors` tends to increase. We can observe that both GPT-3.5 (Figure 3a) and Claude Instant (Figure 3b) tend to make a limited number of `semantic errors` in the SOLO~SCHEMA approach, while the UGLY~SCHEMA approach leads to the largest number of errors. We can observe that the UGLY & INSERT approach, with a limited set of realistic data, seems to reduce the number of `semantic errors`.

On the other hand, the trend in the number of `syntactic errors` is different between the two models. In GPT-3.5, the decrease in the informativeness of the dumps leads to more errors. Manual inspection found that only one error was due to incorrect use of SQL syntax: in most cases, GPT-3.5 has difficulty identifying the tables and columns to be used in the given database and therefore proposes SQL queries that make use of arbitrary tables. In this case, these `syntactic errors` are definitely examples of hallucinations and need to be further explored. Claude Instant, instead, tends to retain more information about the dump, and the number of `syntactic errors` is more constant across the different approaches.

6. Conclusion

In this paper, we propose an iterative reading-comprehension and reasoning approach to solve question-answering challenges of the Text-to-SQL task. The results obtained from the experiments conducted in this work witness the potential of Instruction-tuned Large Language Models (It-LLMs).However, despite their promising performance, certain limitations have emerged. We discovered that even with minimal information about the database, It-LLMs can generate natural language query translations that yield correct and executable SQL queries by just prompting them. Nevertheless, it became evident that reducing the amount of information provided could lead to the generation of incorrect queries. Expanding the scope of our investigation, we believe it would be worthwhile to conduct similar experiments with other It-LLMs. Such comparisons could help determine whether the common phenomena observed in both tested models result from a coincidence or represent aspects to further investigate in studying these new technologies.

In conclusion, this research underscores the substantial advancements offered by It-LLMs in the realm of Text-to-SQL translation while also the implications of choosing whether to provide more or less information during the prompting process.

References

[1] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, 2020. arXiv:1910.10683.

[2] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, 2020. arXiv:2005.14165.

[3] L. Gao, S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima, S. Presser, C. Leahy, The pile: An 800gb dataset of diverse text for language modeling, 2020. arXiv:2101.00027.

[4] S. Mishra, D. Khashabi, C. Baral, H. Hajishirzi, Cross-task generalization via natural language crowdsourcing instructions, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 3470–3487. URL: https://aclanthology.org/2022.acl-long.244. doi:10.18653/v1/2022.acl-long.244.

[5] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, R. Lowe, Training language models to follow instructions with human feedback, 2022. arXiv:2203.02155.

[6] V. Sanh, A. Webson, C. Raffel, S. H. Bach, L. Sutawika, Z. Alyafeai, A. Chaffin, A. Stiegler, T. L. Scao, A. Raja, M. Dey, M. S. Bari, C. Xu, U. Thakker, S. S. Sharma, E. Szczechla, T. Kim, G. Chhablani, N. Nayak, D. Datta, J. Chang, M. T.-J. Jiang, H. Wang, M. Manica, S. Shen, Z. X. Yong, H. Pandey, R. Bawden, T. Wang, T. Neeraj, J. Rozen, A. Sharma, A. Santilli, T. Fevry, J. A. Fries, R. Teehan, T. Bers, S. Biderman, L. Gao, T. Wolf, A. M. Rush, Multitask prompted training enables zero-shot task generalization, 2022. arXiv:2110.08207.

[7] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, H. Hajishirzi, Self-instruct: Aligning language models with self-generated instructions, 2023. arXiv:2212.10560.

[8] Y. K. Dwivedi, N. Kshetri, L. Hughes, E. L. Slade, A. Jeyaraj, A. K. Kar, A. M. Baabdullah, A. Koohang, V. Raghavan, M. Ahuja, H. Albanna, M. A. Albashrawi, A. S. Al-Busaidi, J. Balakrishnan, Y. Barlette, S. Basu, I. Bose, L. Brooks, D. Buhalis, L. Carter, S. Chowdhury, T. Crick, S. W. Cunningham, G. H. Davies, R. M. Davison, R. Dé, D. Dennehy, Y. Duan, R. Dubey, R. Dwivedi, J. S. Edwards, C. Flavián, R. Gauld, V. Grover, M.-C. Hu, M. Janssen, P. Jones, I. Junglas, S. Khorana, S. Kraus, K. R. Larsen, P. Latreille, S. Laumer, F. T. Malik, A. Mardani, M. Mariani, S. Mithas, E. Mogaji, J. H. Nord, S. O'Connor, F. Okumus, M. Pagani, N. Pandey, S. Papagiannidis, I. O. Pappas, N. Pathak, J. Pries-Heje, R. Raman, N. P. Rana, S.-V. Rehm, S. Ribeiro-Navarrete, A. Richter, F. Rowe, S. Sarker, B. C. Stahl, M. K. Tiwari, W. van der Aalst, V. Venkatesh, G. Viglia, M. Wade, P. Walton, J. Wirtz, R. Wright, Opinion paper: "so what if chatgpt wrote it?" multidisciplinary perspectives on opportunities, challenges and implications of generative conversational ai for research, practice and policy, International Journal of Information Management 71 (2023) 102642. URL: https://www.sciencedirect.com/science/article/pii/S0268401223000233. doi:https://doi.org/10.1016/j.ijinfomgt.2023.102642.

[9] J. Jiang, K. Zhou, J.-R. Wen, X. Zhao, *great truths are always simple* : a rather simple knowledge encoder for enhancing the commonsense reasoning capacity of pre-trained models, in: Findings of the Association for Computational Linguistics: NAACL 2022, Association for Computational Linguistics, Seattle, United States, 2022, pp. 1730–1741. URL: https://aclanthology.org/2022.findings-naacl.131. doi:10.18653/v1/2022.findings-naacl.131.

[10] T. Schick, J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, L. Zettlemoyer, N. Cancedda, T. Scialom, Toolformer: Language models can teach themselves to use tools, 2023. arXiv:2302.04761.

[11] Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung, Q. V. Do, Y. Xu, P. Fung, A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity, 2023. arXiv:2302.04023.

[12] Y. Zhou, A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, J. Ba, Large language models are human-level prompt engineers (2022). arXiv:2211.01910.

[13] J. Jang, S. Ye, M. Seo, Can large language models truly understand prompts? a case study with negated prompts, 2022. arXiv:2209.12711.

[14] S. Arora, A. Narayan, M. F. Chen, L. Orr, N. Guha, K. Bhatia, I. Chami, F. Sala, C. Ré, Ask me anything: A simple strategy for prompting language models, 2022. arXiv:2210.02441.

[15] T. Wolfson, D. Deutch, J. Berant, Weakly supervised text-to-SQL parsing through question decomposition, in: Findings of the Association

for Computational Linguistics: NAACL 2022, Association for Computational Linguistics, Seattle, United States, 2022, pp. 2528–2542. URL: https://aclanthology.org/2022.findings-naacl.193. doi:10.18653/v1/2022.findings-naacl.193.

[16] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, T. Mihaylov, M. Ott, S. Shleifer, K. Shuster, D. Simig, P. S. Koura, A. Sridhar, T. Wang, L. Zettlemoyer, Opt: Open pre-trained transformer language models, 2022. arXiv:2205.01068.

[17] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, Q. V. Le, Finetuned language models are zero-shot learners, 2022. arXiv:2109.01652.

[18] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, 2023. arXiv:2302.13971.

[19] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, W. Fedus, Emergent abilities of large language models, 2022. arXiv:2206.07682.

[20] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, 2023. arXiv:2201.11903.

[21] B. Peng, C. Li, P. He, M. Galley, J. Gao, Instruction tuning with gpt-4, 2023. arXiv:2304.03277.

[22] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, D. Zhou, Self-consistency improves chain of thought reasoning in language models, 2023. arXiv:2203.11171.

[23] Z. Lin, S. Trivedi, J. Sun, Generating with confidence: Uncertainty quantification for black-box large language models, 2023. arXiv:2305.19187.

[24] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, T. B. Hashimoto, Stanford alpaca: An instruction-following llama model, https://github.com/tatsu-lab/stanford_alpaca, 2023.

[25] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, E. P. Xing, Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. URL: https://vicuna.lmsys.org.

[26] A. Gudibande, E. Wallace, C. Snell, X. Geng, H. Liu, P. Abbeel, S. Levine, D. Song, The false promise of imitating proprietary llms, 2023. arXiv:2305.15717.

[27] P. Atzeni, R. Basili, D. Hansen, P. Missier, P. Paggio, M. Pazienza, F. Zanzotto, Ontology-based question answering in a Federation of University Sites: The MOSES case study, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2004). URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-35048854325&doi=10.1007%2f978-3-540-27779-8_40&partnerID=40&md5=7545b9abe40e6ac9d64b47d45e71b78c. doi:10.1007/978-3-540-27779-8_40.

[28] R. Basili, D. H. Hansen, P. Paggio, M. T. Pazienza, F. M. Zanzotto, Ontological resources and question answering, in: Proceedings of the Workshop on Pragmatics of Question Answering at HLT-NAACL 2004, Association for Computational Linguistics, Boston, Massachusetts, USA, 2004, pp. 78–84. URL: https://aclanthology.org/W04-2510.

[29] F. Li, H. V. Jagadish, Constructing an interactive natural language interface for relational databases, Proceedings of the VLDB Endowment 8 (2014) 73–84.

[30] T. Mahmud, K. M. Hasan, M. Ahmed, T. Chak, A rule based approach for nlp based query processing, 2015, pp. 78–82. doi:10.1109/EICT.2015.7391926.

[31] B. Bogin, J. Berant, M. Gardner, Representing schema structure with graph neural networks for text-to-SQL parsing, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 4560–4565. URL: https://aclanthology.org/P19-1448. doi:10.18653/v1/P19-1448.

[32] B. Wang, R. Shin, X. Liu, O. Polozov, M. Richardson, RAT-SQL: Relation-aware schema encoding and linking for text-to-SQL parsers, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 7567–7578. URL: https://aclanthology.org/2020.acl-main.677. doi:10.18653/v1/2020.acl-main.677.

[33] I. Sucameli, A. Bondielli, L. Passaro, E. Annunziata, G. Lucherini, A. Romei, A. Lenci, Mate, a meta layer between natural language and database, in: Proceedings of the Sixth Workshop on Natural Language for Artificial Intelligence (NL4AI 2022) colocated with 21th International Conference of the Italian Association for Artificial Intelligence (AI* IA 2022), 2022.

[34] T. Scholak, N. Schucher, D. Bahdanau, PICARD: Parsing incrementally for constrained auto-regressive decoding from language models, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 9895–9901. URL: https://aclanthology.org/2021.emnlp-main.779. doi:10.18653/v1/2021.emnlp-main.779.

[35] T. Xie, C. H. Wu, P. Shi, R. Zhong, T. Scholak, M. Yasunaga, C.-S. Wu, M. Zhong, P. Yin, S. I. Wang, V. Zhong, B. Wang, C. Li, C. Boyle, A. Ni, Z. Yao, D. Radev, C. Xiong, L. Kong, R. Zhang, N. A. Smith, L. Zettlemoyer, T. Yu, UnifiedSKG: Unifying and multi-tasking structured knowledge grounding with text-to-text language models, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 602–631. URL: https://aclanthology.org/2022.emnlp-main.39.

[36] T. Yu, R. Zhang, K. Yang, M. Yasunaga, D. Wang, Z. Li, J. Ma, I. Li, Q. Yao, S. Roman, Z. Zhang, D. Radev, Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 3911–3921. URL: https://aclanthology.org/D18-1425. doi:10.18653/v1/D18-1425.

[37] OpenAI, Chatgpt, 2022. URL: https://chat.openai.com/.

[38] Anthropic, Claude-instant, 2022. URL: https://poe.com/Claude-instant.