Information Processing and Management 61 (2024) 103809

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Are LLMs good at structured outputs? A benchmark for evaluating structured output capabilities in LLMs

Yu Liu [a,c], Duantengchuan Li [a,*], Kaili Wang [b], Zhuoran Xiong [d,*], Fobo Shi [e], Jian Wang [a], Bing Li [a,f,*], Bo Hang [c]

[a] School of Computer Science, Wuhan University, Wuhan 430072, China
[b] Faculty of Artificial Intelligence in Education, Central China Normal University, Wuhan 430079, China
[c] School of Computer Engineering, Hubei University of Arts and Science, Xiangyang 441100, China
[d] Electrical and Computer Engineering, McGill University, Montreal, Canada
[e] School of Information Management, Wuhan University, Wuhan 430072, China
[f] Hubei Luojia Laboratory, Wuhan 430079, China

ARTICLE INFO

Keywords:
Large language model
Structured output capability
Benchmark dataset
Q&A interaction
Causal graph

ABSTRACT

Existing benchmarks for Large Language Models (LLMs) mostly focus on general or specific domain capabilities, overlooking structured output capabilities. We introduce SoEval, a benchmark for assessing LLMs' ability to generate structured outputs like JSON, XML, and lists. SoEval contains 3.7K entries in Chinese and English, covering 13 types of structured output tasks across 20 subjects. In experiments, we found that while current mainstream LLMs have deficiencies in structured output, GPT-4 outperforms them in this aspect. GPT-4 achieved an average score of 0.4 on SoEval, representing a 24% enhancement over the next best-performing model. At the same time, the performance of current mainstream models on English tasks is also better than on Chinese tasks. We also report the performance of mainstream large models on different structured output types and task subjects. The benchmark construction code and SoEval dataset are open-sourced at https://github.com/MoranCoder95/SoEval.

1. Introduction

Large language models (LLMs) like Llama (Touvron et al., 2023), GPT-3.5 (OpenAI, 2022) and GPT-4 (OpenAI, 2023) are widely utilized in various domains due to their outstanding capabilities. For example, text generation (Wang et al., 2024), sentiment analysis, and automatic question answering (Minaee et al., 2024) are some of the areas where these models excel. LLMs are frequently described as "black boxes" due to the complexity and opacity of their inner workings, making them challenging to fully comprehend (Zhao, Yang, Lakkaraju, & Du, 2024). Specific benchmark tests are one of the most effective ways to assess and understand the capabilities of these "black box" models (Guidotti, Monreale, Turini, Pedreschi, & Giannotti, 2018). Assessing the capabilities of the model through these tests allows for a deeper understanding of its strengths and weaknesses, as well as identification of potential use cases. For example, a LLM model that is trained on a wide range of standard texts may excel at general language tasks, but struggle with more niche or specialized content, indicating a gap in the model's capabilities in these areas (Wayne Xin Zhao et al., 2023). Therefore, it is crucial to develop suitable evaluation benchmarks to gain a comprehensive understanding and further enhance the capabilities of these LLMs (Chandran et al., 2024).

* Corresponding authors.
E-mail address: dtclee1222@gmail.com (D. Li).

https://doi.org/10.1016/j.ipm.2024.103809
Received 10 March 2024; Received in revised form 26 May 2024; Accepted 10 June 2024

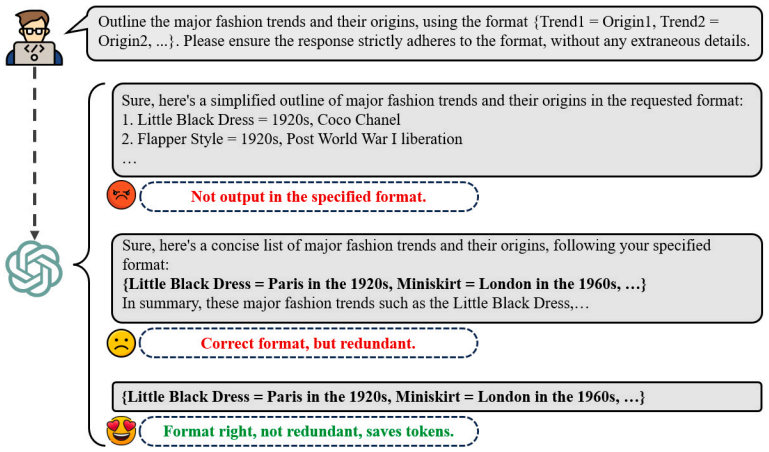


Fig. 1. The example shows three responses from a large language model that summarize fashion trends and their origins according to a specific format. These responses show varying degrees of adherence to the formatting requirements: one is incorrect in format, another is correct but redundant, and the last is both correct and concise in format.

Table 1
Summary of some existing LLMs evaluation benchmarks.

Dataset	Year	Focus	Types	Size
C-Eval (Huang et al., 2024)	2023	Chinese evaluation	General language task	13,948
Xiezhi (Gu et al., 2023)	2023	Comprehensive domain knowledge	General language task	249,587
LLMEval3 (Zhang, Yu et al., 2023)	2023	LLM evaluator	General language task	200,000
SuperCLUE (Xu et al., 2023)	2023	Chinese evaluation	General language task	–
AGI Eval (Zhong et al., 2023)	2023	Human cognition	General language task	–
MMLU (Hendrycks, Burns, Basart et al., 2021)	2021	Zero-shot/few-shot evaluation	General language task	57 subjects
AQA-Bench (Yang, Zhao, & Xie, 2024)	2023	Sequential reasoning in algorithmic contexts	General language task	3 algorithms
GSM8K (Cobbe et al., 2021)	2021	Multi-step mathematical reasoning	Educational/mathematics	8.500
MATH (Hendrycks, Burns, Kadavath et al., 2021)	2021	Challenging competition-level math problems	Educational/mathematics	12,500
SuperCLUE-Math6 (Xu, Xue, Zhu, & Zhao, 2024)	2024	Graded multi-step math reasoning in Chinese	Educational/mathematics	Over 2000
Human Eval (Chen et al., 2021)	2021	Program synthesis	Code	164 problems
MBPP (Austin et al., 2021)	2021	Entry-level programming	Code	1000 problems
Mercury (Du, Luu, Ji, & Ng, 2024)	2024	Code efficiency benchmark	Code	1889 tasks
GAOKAO-bench (Zhang, Li et al., 2023)	2023	Language understanding, logical reasoning	Education/educational testing	2811
ARC (Clark et al., 2018)	2018	Science exam questions	Education/educational testing	Two partitions (easy, challenge)
Flores (Goyal et al., 2021)	2021	Wikipedia sentences	Machine translation	101 languages
COPA (Wang et al., 2019)	2020	Causal relation selection	Causal inference task	–
LawBench (Fei et al., 2023)	2023	Benchmarking legal knowledge	Law	20 tasks
SoEval (ours)	2024	Structured output capabilities	Output capabilities	13 structure-related types, 20 task-related subjects

Currently, the performance evaluation of LLMs relies heavily on specially designed benchmarks (Abdelali et al., 2024). These benchmarks are capable of testing the models' abilities in multiple languages and tasks. Widely-used evaluation benchmarks like SuperGLUE (Xu et al., 2023), and MMLU (Hendrycks, Burns, Basart et al., 2021) cover a range of tasks from understanding to generation, providing a standardized way to measure and compare the capabilities of different LLMs. As shown in Table 1, the main benchmarks are primarily categorized into two types. The first type includes benchmarks assessing general language tasks, evaluating performance across various tasks and general domain knowledge. For example, benchmarks like C-Eval (Huang et al., 2024) and Xiezhi (Gu et al., 2023) test competencies in diverse fields including science, engineering, agronomy, medicine, and art. The second type comprises benchmarks targeting specific downstream tasks, evaluating specialized capabilities and domain knowledge. Examples include benchmarks like SuperCLUE-Math6 (Xu et al., 2024), which tests mathematical abilities, and Mercury (Du et al., 2024), which evaluates coding abilities in LLMs.

Table 2
Survey results on structured output capabilities of large language models among 100 respondents.

Question No.	Description	IT	Education/Research	Healthcare	Tech services	Finance	Other	Total
1	Your industry	20	25	15	18	12	10	100
2	Use of LLM in work							
	A. Often (frequently using LLM in work tasks)	15	20	10	14	8	5	72
	B. Occasionally (using LLM in work tasks from time to time)	5	3	4	2	3	2	19
	C. Rarely (seldom using LLM in work tasks)	0	2	1	1	1	2	7
	D. Never (never using LLM in work tasks)	0	0	0	1	0	1	2
3	Importance of LLM's structured output							
	A. Very important (LLM's structured output is crucial for work tasks)	18	22	14	16	11	7	88
	B. Important (LLM's structured output plays a significant role in work tasks)	2	2	1	1	1	2	9
	C. Somewhat important (LLM's structured output is useful but not critical for work tasks)	0	1	0	1	0	1	3
	D. Not important (LLM's structured output has little to no impact on work tasks)	0	0	0	0	0	0	0
4	Satisfaction with LLM's structured output							
	A. Very satisfied (highly content with the quality and utility of LLM's structured output)	2	3	2	3	1	1	12
	B. Satisfied (content with LLM's structured output, though there may be minor issues)	2	4	2	2	2	2	14
	C. Neutral (neither satisfied nor dissatisfied with LLM's structured output)	6	8	5	6	4	3	32
	D. Unsatisfied (discontent with the quality or utility of LLM's structured output)	10	10	6	7	5	4	42
5	Structured output formats used (multiple choice)							
	A. JSON/XML	18	8	0	2	3	4	35
	B. Table	10	22	14	16	11	8	81
	C. List	17	24	15	16	9	9	90
	D. Matrix	3	4	2	3	11	3	26
	... other formats	–	–	–	–	–	–	–

Note: The survey involved a total of 100 participants, selected through purposive sampling to ensure representation from various industries. The participants were chosen based on their professional roles.

Existing evaluation benchmarks assess LLMs in general or specialized domains, but they often overlook the models' ability to produce structured outputs. Structured formats, such as lists, JSON, and XML, provide an intuitive and effective way to understand and interact with LLMs. These formats not only aid in human comprehension and reading of information (Dagdelen et al., 2024), but also facilitate data handling in software development and other applications (Izquierdo & Cabot, 2014). As illustrated in Fig. 1, when invoking LLMs' interfaces in software development, we require LLMs to output in specific formats to facilitate code processing. However, LLMs may not always output in the required format, leading to errors in downstream tasks. Additionally, LLMs may produce irrelevant context, which can increase token consumption even when the correct format is produced. At the same time, our questionnaire survey conducted among specific groups (as shown in Table 2) indicate that while there is a clear demand for diverse format outputs among LLM users, the satisfaction levels regarding the structured output capabilities of LLMs during actual usage are less than optimistic. This exposed deficiencies in the structured output capabilities of LLMs. Therefore, creating a benchmark to evaluate the structured output of LLMs is crucial to address a gap in evaluations and has significant practical importance.

In this paper, we have created a benchmark for assessing the structured output capabilities of LLMs. **In terms of theory**, we initially analyze the prompt, assuming a separation between task-related instructions and structure-related instructions, and construct a causal graph of the prompt. This causal graph analysis serves as a guide for the development of our entire benchmark. **In practical terms**, we first define common forms of structured output and generate corresponding regular expressions. Then, based on the patterns reflected in these regular expressions, we create structure-related instructions. Subsequently, we generate task-related instructions by considering potential application scenarios. By combining these structure-related and task-related instructions, we compose our final prompt, which undergoes rigorous human evaluation. When these prompts are presented to large models like GPT-4, their responses can be compared against the regular expressions to calculate the model's structured output score. **Regarding evaluation criteria**, given the importance of practical applications that require both structured output and redundancy, our evaluation criteria are designed with these aspects in mind.

1.1. Research objective

The primary objective of this research is to bridge the existing gap in evaluating large language models concerning their capability to produce structured outputs. While current benchmarks effectively assess the general and domain-specific capabilities of LLMs, they often overlook the importance of structured output formats which are crucial for a wide range of practical applications. Our main contributions in this research are outlined as follows:

- Our focus is to evaluate the structured output capabilities of LLMs. By analyzing the causal graphs of prompt structures, we provide a viable technical solution to the challenging problem of evaluating structured output capabilities. Our solution is both scalable and practical.
- We have developed the SoEval dataset, which is the first dataset in the structured output evaluation domain. The dataset includes 13 types of structured output tasks, covering 20 task subjects.
- We have tested several major language models using SoEval and reported their capabilities in structured output. Our benchmark will help guide future improvements and research in the field of LLMs.

2. Related work

In this section, we examine the evolution of Large Language Models (LLMs) from basic statistical models to advanced systems like GPT and BERT, and their detailed evaluation methods addressing performance, ethics, and robustness.

2.1. Large Language Models (LLMs)

Large Language Models (Chang et al., 2023; Wu et al., 2024) such as Generative Pre-trained Transformer (GPT) (Floridi & Chiriatti, 2020) are trained on large datasets, enabling them to generate, translate, or summarize text with high accuracy. Early language models, like N-gram models (Brown, Della Pietra, Desouza, Lai, & Mercer, 1992), used statistical methods to predict word sequences but struggled with long-range dependencies and complex language structures due to their reliance on immediate context. Subsequently, the emergence of neural network-based models marked a significant advancement. For example, RNNs (Zaremba, Sutskever, & Vinyals, 2014) and their variants LSTM (Hochreiter & Schmidhuber, 1997) networks improve the handling of sequential data by capturing longer-term dependencies in text. Despite this, these models face difficulties with very long sequences and require significant data and computational resources. The introduction of transformer-based architectures (Vaswani et al., 2017), represented by models like GPT (Floridi & Chiriatti, 2020; OpenAI, 2023) and BERT (Devlin, Chang, Lee, & Toutanova, 2018), brings a major breakthrough. These models employ self-attention mechanisms to assess the importance of different parts of input data, allowing them to focus on relevant sections of text for predictions or text generation. This method greatly enhances language models' ability to understand context and subtleties in language.

The release of ChatGPT marks a significant moment in the evolution of transformer architecture-based LLMs, which are characterized by their enormous parameter sizes and strong alignment with human feedback. This alignment enhances their adaptability and accuracy in a wide range of language tasks (Wan et al., 2023). ChatGPT was built on the GPT-3.5 model and set a new standard for dialog quality (OpenAI, 2022). The subsequent introduction of GPT-4 further enhanced the capabilities of these models, handling more complex tasks with increased accuracy and the innovative ability to analyze images and adjust tone. In addition to the renowned GPT series, a number of institutions have ventured to develop their own large-scale models. In particular, Google's PaLM 2 (Bison-001) (Anil et al., 2023) excels in multilingual programming and complex logical reasoning, demonstrating its ability to be trained on large datasets. Similarly, Anthropic's Claude (Claude, 2023) model focuses on creating AI assistants that are helpful, honest, and harmless, consistently outperforming in various benchmark tests. In addition, Alibaba's Qwen series (Qwen-72B, Qwen-14B, and Qwen-7B) (Bai et al., 2023) have demonstrated impressive performance across a range of tasks. The spread of these large language models (LLMs) not only shows a big step forward in technology but also sets the groundwork for achieving general artificial intelligence.

2.2. LLM evaluations

The evaluation of LLMs is a central aspect of modern AI research, serving as a window into their operational capabilities and guiding the direction of future improvements (Ning et al., 2024). This complex process involves a variety of benchmarks and methodologies that have been carefully designed to investigate the various dimensions of LLM performance. For example, GAOKAO-2023 and MMLU (Hendrycks, Burns, Basart et al., 2021) are used to assess the models' performance in academic settings, testing their understanding of various topics. Meanwhile, datasets such as WiC (Pilehvar & Camacho-Collados, 2019) and TyDiQA (Clark et al., 2020) are crucial for assessing how well they understand and generate language, testing both their comprehension and linguistic creativity. In addition, tools such as BoolQ (Clark et al., 2019) and NaturalQuestions (Kwiatkowski et al., 2019) provide insight into the models' ability to process and apply knowledge, a key factor in their practical utility. Data sets such as C3 (Sun, Yu, Yu, & Cardie, 2020) and RACE (Lai, Xie, Liu, Yang, & Hovy, 2017), which focus on the models' reading comprehension skills, are also used for deeper text comprehension and critical analysis. Finally, benchmarks such as CMNLI (Xu et al., 2020) and PIQA (Bisk, Zellers, Bras, Gao, & Choi, 2019) play an important role in assessing the models' logical, causal, and common sense reasoning skills, which are essential for decision making and problem solving. These comprehensive assessment methods not only highlight the strengths and weaknesses of LLMs in different domains, but also directly guide the direction of AI progress.

Evaluating LLMs involves more than measuring performance. It also includes testing for robustness, ethics, bias neutrality, and trustworthiness (Dong, Zhou, Yang, Shao, & Qiao, 2024; Yao et al., 2024). Robustness testing involves assessing how stable models are with unexpected inputs, such as attacks or new scenarios. Ethical concerns are critical because LLMs can inherit and spread biases or harmful content from their training data. Bias neutrality refers to the importance of ensuring that the outputs are not influenced by hidden biases, which is crucial for building a fair artificial intelligence system (Sheng, Chang, Natarajan, & Peng, 2021). Trustworthiness testing focuses on whether the model's outputs are accurate, reliable, and meet user expectations, mainly involving the issue of "hallucinations" in LLMs (Wang et al., 2023). Although newer versions like GPT-4 show improvements, they also have new vulnerabilities against advanced attacks (Chang et al., 2023). Therefore, integrating different datasets and various evaluation methods to comprehensively assess LLMs from multiple aspects is key to understanding the limitations of LLM capabilities.

3. Preliminaries

In this section, we primarily introduce some fundamental principles of our benchmark design and utilize causal graphs to review the structure of the prompt.

3.1. Design principles

Motivation: Large language models (LLMs), as emerging infrastructure technologies, are now widely used in various industries. In particular, there is a clear need for these models to generate structured output, such as JSON or XML formats, especially in software

Table 3
Descriptions of different types of structured outputs.

Type	Description	Example
List	Vertically arranges items, facilitating quick reading and referencing.	1. Apple, 2. Samsung, 3. Huawei
Table	Organizes data in rows and columns for easy understanding and analysis.	–
Headings & Subheadings	Organizes and highlights document structure.	–
Q&A	Conversational layout of questions and answers.	Q: Monthly budget? A: Track income and expenses.
Timeline	Shows events in chronological order for tracking progress or history.	1914: War begins; 1918: War ends
Flowchart	Graphical representation of steps, decisions, and processes.	Step 1: Analysis; Step 2: Design
Key-Value Pairs	Each key is associated with a value, used in data storage.	Name: John, Age: 20, ID: 123456
JSON Format	Formats for structured data, used in web and files.	{“title”: “Book”, “author”: “Author”}
Triples	Describes information in a three-part structure.	(Paris, is located in, France)
Attribute Graph	Displays an object’s attributes.	Phone Attributes: Model, Price, OS
Url Parameters	Encodes data in URLs with key–value pairs, often in GET requests.	?category=books&price=low
XML Format	Markup language for data description and storage, hierarchical.	–
Formats Modify	Converts data between different formats.	–

Table 4
Subject categories across diverse task domains.

Number	Subject
1	Business, Economics & Entrepreneurship
2	Social Sciences, Education & Human Rights
3	History, Geography & Cultural Studies
4	Environmental, Earth Sciences & Sustainability
5	Health, Wellness & Fitness
6	Religious Studies, Theology & Philosophy
7	Science & Technology
8	Literature, Arts & Performing Arts
9	Lifestyle, Personal Development & Languages
10	Fashion, Design & Arts and Crafts
11	Engineering, Architecture & Urban Studies
12	Astronomy, Space Exploration & Astrophysics
13	Travel, Tourism & Cultural Landmarks
14	Media, Communications & Digital Technologies
15	Agriculture, Forestry & Animal Care
16	Law, Legal Studies & International Relations
17	Social Issues, Environmental Policy & Human Rights
18	Economic Theories, Models & Political Systems
19	Mathematics, Statistics & Data Science
20	Other

development processes to increase efficiency and convenience. This need has been demonstrated by extensive surveys across various industries (survey results can be found in Table 2). Despite their importance, our survey results indicate a general dissatisfaction with the current capabilities of LLMs in this area. As shown in Table 2, our survey involved 100 respondents from various industries, with the majority (91%) indicating frequent or occasional use of LLMs in their work. Notably, 97% considered structured output capabilities to be very important or important. However, only 26% expressed satisfaction with current LLM performance in this area, while 42% reported being unsatisfied. The most commonly used structured output formats were List (90 respondents), Table (81 respondents), and JSON/XML (35 respondents). These results highlight the significant gap between the demand for structured output capabilities and the current level of user satisfaction, underscoring the need for benchmarks to assess and improve LLM performance in this critical area.

Types & subjects selection: Our research focuses on the structured output produced by LLMs, with particular emphasis on formats such as lists and JSON, which are becoming increasingly important for improving business operations and facilitating smoother integration with AI technologies. To address industry challenges, we have identified 13 structured output formats that are in high demand, as detailed in the Table 3. These formats are expected to be of significant benefit to a wide range of future applications. To demonstrate the broad utility of these structured formats, our study extensively explores 20 different task subjects, as outlined in Table 4, highlighting the diverse applicability of our findings across different domains.

Dataset generation approach: To test the core abilities of LLMs, we have adopted an efficient method for generating datasets by using LLMs themselves. Advanced models such as GPT 4.0 have been particularly useful in dataset construction (Chang et al., 2023). This approach highlights the advantages of utilizing LLMs to create datasets, significantly reducing the time and manual effort involved. It makes the process feasible and efficient. By using this strategy, we can quickly come up with a wide variety of prompts that have specific structured output requirements, which is essential for evaluating and testing the capabilities of LLMs.

Attempting to mitigate data contamination: When building our dataset, we carefully avoid data contamination to ensure quality. This is important because it prevents inflated performance metrics caused by models merely recalling information rather than

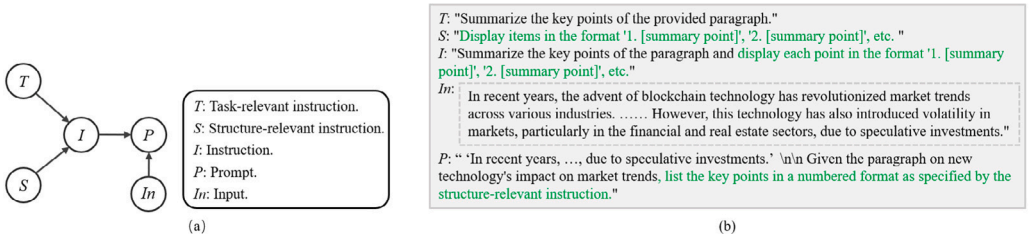


Fig. 2. Causal graph of the prompt. (a) illustrates the process of constructing prompt based on task-related and structure-related instruction. In this causal graph, the prompt (P) is generated by integrating a structure-related instruction (S) with a task-specific instruction (T), influenced by the input (I_n). (b) provides an illustrative example derived from the causal diagram.

generating new content. Since there are currently no exclusive datasets for structured output in LLMs training, the problem of data contamination can be avoided. Importantly, LLMs can exhibit an "echo effect" where they generate content that is more familiar or easily accessible to them. This means that LLMs may prioritize generating content that frequently appears or is easily understood in their training data, rather than completely novel or unique content. This property is actually beneficial for our research, as we are focused on evaluating the structured output capabilities of these models, rather than the factual accuracy of the content generated. The "echo effect" ensures that the prompts created by the LLMs are within their competence, providing a relevant basis for evaluating their structured output performance.

3.2. Overview of prompt

Prompt engineering in the context of large language models (LLMs) involves the creation of well-designed prompts to guide the model to produce the desired output. The skill of crafting such prompts is critical because it has a marked effect on the responses generated by the model (Liu et al., 2023). As shown in Fig. 2, prompts are composed of two key components: instruction and input. The instruction part clearly states the task, setting the tone for the model's response, while the input provides necessary context or specific examples. Moreover, instruction can be further divided into task-related instruction and structure-related instruction. Task-related instruction define the specific task or question the model needs to address, while structure-related instruction shape the format and structure of the LLM's response.

In order to provide a more vivid illustration of the relationships among the aforementioned elements, we employ a causal graph (as shown in Fig. 2) to visually depict the connections between task-related instruction (T), structure-related instruction (S), and input (I_n) in the generation of the final prompt (P). The causal graph can intuitively show the internal structure of the prompt and the causal relationship between its components, providing support for the theoretical basis of the prompt research. In constructing the causal graph, we make specific assumptions, such as assuming that task-related instruction ((T)) and structure-related instruction ((S)) are independent of each other, and that they are equally important for the generation of instruction ((I)).To quantitatively describe this relationship, we make use of a structural equation model (SEM) (Jöreskog & Sörbom, 1982), as demonstrated in the following equation:

$$\begin{aligned} P &:= f\left(I, I_n\right), \\ I &:= h\left(T, S\right) = f\left(h(T, S), I_n\right). \end{aligned} \tag{1}$$

The generation of the entire prompt P is seen as a combination of functions f and h, given task-related instruction T, structure-related instruction S, and input contents I_n. In our work, functions (f) and (h) are both implemented by LLMs (GPT 4.0), with the task-related instruction (T) being generated by the LLMs based on 20 predefined task domains (as shown in Table 4), and the structure instruction (S) generated according to 13 predefined high-demand structured output formats (as listed in Table 3). By decomposing the prompt into its constituent elements, a pipeline approach can be effectively implemented to assemble the final prompt.

4. Methodology

To accurately assess the structured output capability of Large language models (LLMs), we employ a pipeline approach to generate test prompts, as illustrated in Fig. 3. The process begins with the identification of 13 structured output types, which are derived from extensive industry surveys to ensure the dataset's relevance and practicality. Following this, regular expression (regex) generation is used to define structured data formats. These defined regex patterns serve as the foundation for defining evaluation criteria in subsequent steps. During the instruction construction phase, since instructions are categorized into structure-related instructions and task-related instructions, we construct structure-related instructions using regex patterns and generate corresponding task-related instructions randomly. As we primarily focus on the structured output capability of LLMs, the prompt solely comprises instructions, excluding input elements from the causal graph. The final instructions are then manually verified and edited to ensure their quality. In the testing phase, these final prompts are input into LLMs to automatically generate results, and by comparing the generated results with the regex patterns, the outcome scores can be calculated. Next, we will discuss in detail how to construct the SoEval dataset and introduce the evaluation criteria.

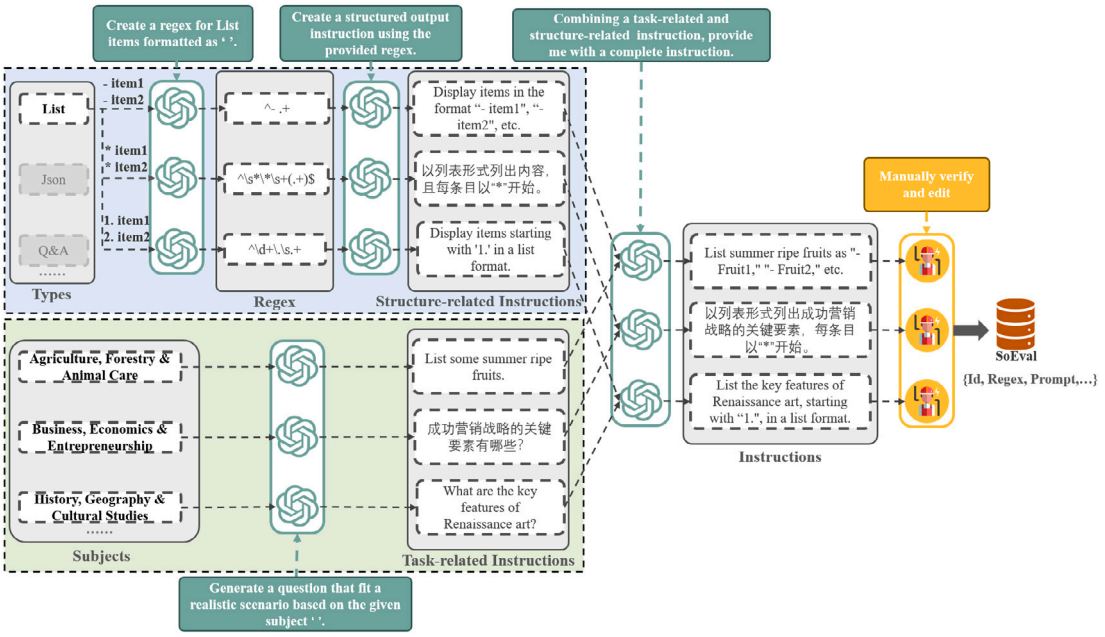


Fig. 3. Pipeline overview for SoEval dataset construction. The pipeline includes key steps: defining structured output formats, constructing regular expressions, developing structure-related and task-related instructions, creating final instruction prompts, and performing manual reviews.

4.1. Detailed methodology steps

Our methodology is meticulously designed as a structured pipeline, with each phase building on the previous to ensure a seamless flow from conception to implementation. This pipeline approach not only enhances efficiency and coherence, but also ensures that each component of our data set is created with precision and purpose. The following steps outline this pipeline and showcase the detailed process of constructing our SoEval data.

- **Step 1: Selection of Structured Output Types.** To ensure the relevance of our dataset across industries, we carefully select structured output tasks that are in high demand. For example, from our extensive surveys (refer to Table 2), we found that JSON formatting is a frequently requested task in the IT and data analytics industries. Similarly, XML formatting is critical for Web development and data exchange. By identifying such 13 high-demand structured output formats (refer to Table 3), we aim to make our dataset an invaluable tool for real-world structured output generation challenges.
- **Step 2: Generation of Regular Expressions (Regex).** For each structured output types, we generate regex patterns that precisely define the expected formats. For example, for JSON formatting tasks, we generate regex that validates the proper structure of JSON objects, ensuring that they have properly quoted keys and valid value types. Similarly, for XML tasks, the regex checks for well-formed tags and proper nesting. These patterns are essential for establishing objective, quantifiable standards for evaluating the performance of LLMs in generating structured output, ensuring that our benchmarks are truly applicable to practical applications.
- **Step 3: Development of Instructions.** As depicted in the causal graph in Fig. 2, instructions are divided into task-related instructions and structure-related instructions. Task-related instructions specify the task or question to be addressed, while structure-related instructions dictate the format and structure of the output. This division method ensures that the prompts are comprehensive and effectively guide the LLMs to generate the desired outputs. Building upon the regular expressions defined in previous steps, we construct the corresponding structure-related instructions S using GPT 4.0. Similarly, we also construct task-related instructions T using GPT 4.0. To ensure diversity in our constructed data subjects, we randomly select the subjects from predefined subjects (see Table 4) for generating task-related instructions.
- **Step 4: Creating the Final Prompts.** To create the final prompts P, we merge the task-related instructions T with the structure-related instructions S. For instance, if T is to list some summer ripe fruits, and S is to display items in the format "- item1", "- item2", etc., then the merged final prompt P would be: List summer ripe fruits as "- Fruit1", "- Fruit2", etc. These prompts are designed to test the LLM's ability to perform structured output tasks. They are generated using a systematic pipeline approach that ensures efficiency and consistency in prompt generation.
- **Step 5: Manual Review and Quality Assurance.** After prompts are generated, they undergo a careful manual review to ensure their quality, accuracy, and relevance. This process involves two main aspects: semantic review of the generated prompts and verification of the corresponding regular expressions and structured output requirements. The semantic review assesses the

Table 5
Randomly selected examples from the SoEval dataset.

Id	Type	Instruction	Subject
e862	Q & A	Produce a Question & Answer discussing the future implications of sustainable energy.	Environmental, Earth Sciences & Sustainability
e1109	Key-Value Pairs	Describe the various types of traditional dances from around the world in key–value pairs, where the key is the country and the value is the dance.	Lifestyle, Personal Development & Languages
e360	List	Describe the features of a modern smartphone, presented as a list, with each preceded by a hyphen.	Social Sciences, Education & Human Rights

clarity, fluency, and contextual appropriateness of the instructions, ensuring that they are easily understandable and aligned with the intended purpose of the structured output task. The regular expression and structured output verification involves examining the regex patterns associated with each prompt to ensure their accuracy and effectiveness in capturing the desired structured output format.

4.2. Evaluation criteria

In the evaluation of LLMs for structured output tasks, our criteria are centered on the accuracy of the model-generated outputs in relation to predefined regular expressions (regex) that represent the structural and formatting requirements of desired outputs such as JSON, XML, etc. This metric is defined as

$$U = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\text{match\_length}\left(\text{regex}_i, \text{output}_i\right)}{\text{total\_length}\left(\text{output}_i\right)}\right), \tag{2}$$

where N is the total number of prompts, output_i is the LLM's structured output for the ith prompt, regex_i represents the corresponding regex pattern for the ith task, $\text{match\_length}(\text{regex}_i, \text{output}_i)$ calculates the matching segments' length in output_i to regex_i, and $\text{total\_length}(\text{output}_i)$ denotes the ith output's total length. This Evaluation Criteria focuses on the accuracy of structured output while also considering the redundancy of the output content. As illustrated in Fig. 1, LLMs may face challenges in generating outputs that strictly adhere to the required structure and format. They may produce outputs with incorrect formatting or include redundant information that does not directly contribute to the structured output. Our evaluation criteria takes both of these aspects into account. By calculating the ratio of the matching segment length to the total output length, we quantify the accuracy of the LLM's structured output. A higher ratio indicates that a larger portion of the output matches the required structure and format defined by the regex pattern. At the same time, by considering the total length of the output in the denominator, we indirectly account for redundancy. If the LLM generates a lengthy output that contains a significant amount of irrelevant or redundant information, it will result in a lower overall score.

5. The proposed SoEval dataset

In this section, we focus on the SoEval dataset that we have developed. Initially, we provide an overview of the dataset, followed by a comprehensive discussion on its aspects of reliability and diversity.

5.1. Dataset description

The SoEval dataset is a comprehensive resource tailored for evaluating the structured output capabilities of Large Language Models (LLMs). It includes 13 unique structured output types designed to reflect a wide range of real-world data structuring needs. These types are presented in a variety of formats, including but not limited to JSON, XML, and structured lists, to reflect the diverse nature of structured data handling. As illustrated in Fig. 4, the dataset shows the distribution of data entries across all the different format types. Furthermore, the dataset is diversified across 20 different task subjects, with the proportion of data entries for each subject shown in Fig. 5. The total size of the SoEval dataset is 3.7K entries, making it a comprehensive benchmark for evaluating LLMs in structured data interpretation and generation tasks. To provide a practical insight into the composition of the dataset, Table 5 contains randomly selected examples from the SoEval dataset, illustrating the variety and complexity of the tasks it contains.

5.2. Reliability and consistency

The SoEval dataset has been meticulously crafted to ensure reliability and consistency in the evaluation of LLMs. Rigorous procedures were employed throughout, from the initial task design to the final dataset compilation. Firstly, prior to constructing the dataset, thorough investigations were conducted to ascertain its research value. Secondly, the construction process heavily relied on the GPT 4.0 model, leveraging its "echo effect" to ensure the dataset effectively evaluates the structured output capability of large models. Lastly, after formulating the testing instructions, manual checks and revisions were conducted to ensure the data aligns with real-world usage scenarios.

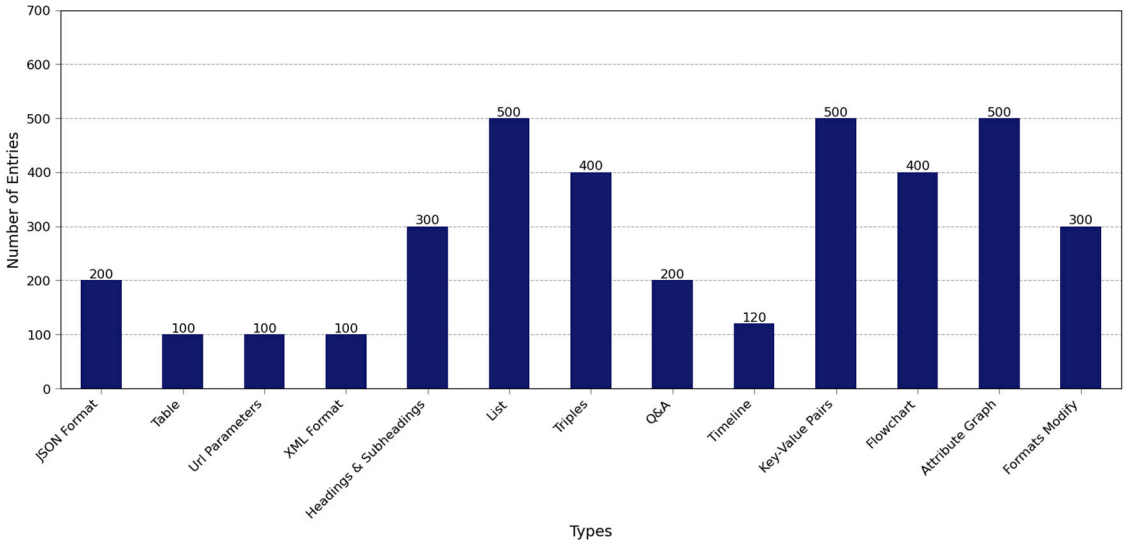


Fig. 4. Number of entries by types.

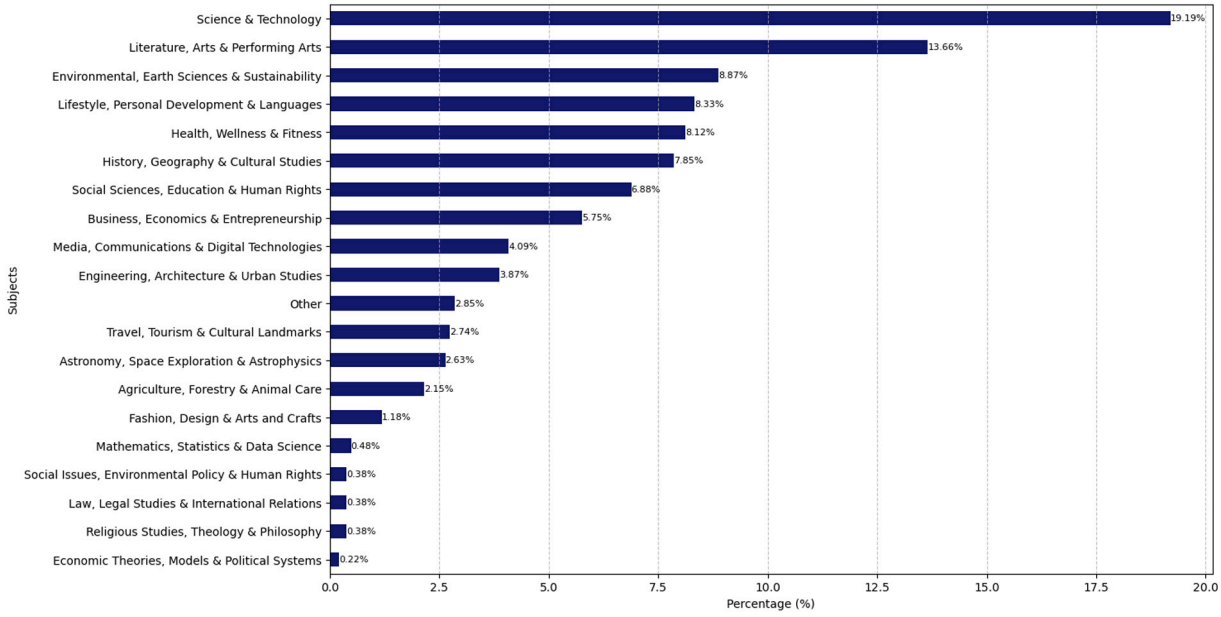


Fig. 5. Percentage of subjects.

5.3. Coverage and diversity

The SoEval dataset possesses a significant coverage and diversity, making it a powerful tool for evaluating the structured output capabilities of LLMs across various tasks. The dataset includes a variety of structured output types, carefully designed to encompass different data formats such as JSON, XML, and structured lists. These types are chosen based on a thorough analysis of current market trends and industry demands, ensuring the dataset's relevance and applicability across diverse sectors. In addition to the range of formats, the dataset contains task subjects that address different industry needs, reflecting the broad applications of LLMs

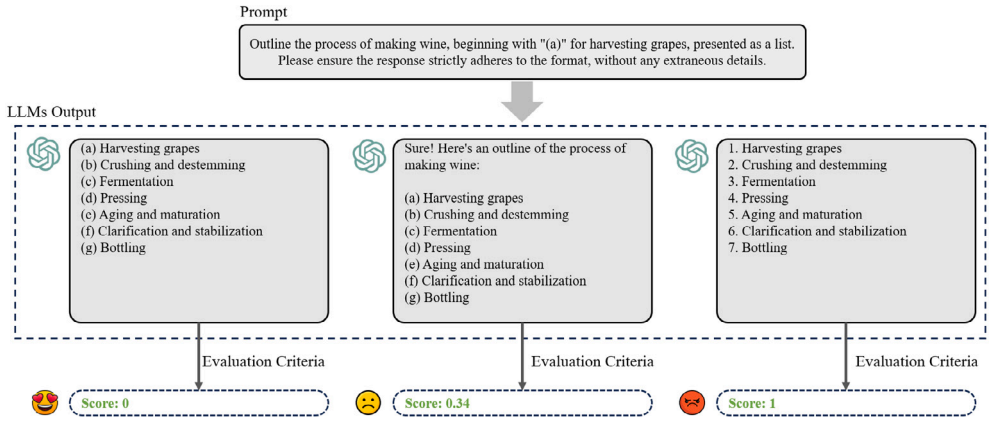


Fig. 6. Example of zero-shot testing with SoEval dataset.

in real-world scenarios. Not only have we constructed the SoEval dataset, but we have also provided a convenient method for its construction, facilitating the dataset's future expansion.

5.4. Dataset extensibility

The SoEval dataset construction pipeline is designed with modularity and extensibility in mind, allowing researchers to conveniently expand the dataset to include a wider range of structured output types and task subjects. The key to this extensibility lies in the use of regular expressions to define the structure and leveraging large language models (LLMs) for data generation. Specifically, researchers first define the structured output format and generate corresponding regular expressions using LLMs. Subsequently, based on the regular expressions, the LLMs generate the relevant structure-related instructions. For specific task scenarios, LLMs can be employed to generate task-related instructions. Finally, by combining the structure-related instructions and task-related instructions, the ultimate instructions requiring structured output can be generated. Through manual inspection and modification, new test data for other structured output types can be formed. By leveraging regex patterns, LLMs, and a modular pipeline, the SoEval dataset construction method provides a convenient and efficient way to extend the dataset to encompass a diverse range of structured output types. Considering the power of the community open-source platform, our project code is also hosted on the Hugging Face platform, ensuring that our project can be noticed and used by more researchers and that our benchmark can maintain good usability over time.

6. Benchmarking experiments

Our research below is focused on assessing different Large Language Models (LLMs) using the SoEval framework, examining their effectiveness, and establishing benchmarks for subsequent applications of SoEval.

6.1. Task settings

The objective of our research is to accurately simulate real-world scenarios where the need for structured outputs from LLMs. In our benchmarking experiments, we employed a wide array of tasks from the SoEval dataset. These tasks are specifically designed to evaluate the ability of various LLMs to generate structured outputs in response to specific instructions. In our experimental setup, uniform configuration settings were implemented across all models to ensure a level playing field in the evaluation process. The randomness in predictions was controlled, allowing for a consideration of the top 80% most probable outcomes. We set a cap on the output length at around 1500 units, ensuring that responses are concise yet comprehensive. To prioritize accuracy and relevance, the randomness factor in the models' responses was kept exceptionally low, near 0.01. Furthermore, we applied a repetition penalty setting of 1.0, indicating no extra penalty for repeating words or phrases, which is vital in tasks where repetition is necessary for clarity or emphasis. To further emphasize the capabilities of LLMs, we use a zero-shot evaluation approach to assess the structured output capabilities of LLMs such as Llama and GPT-3.5. In these experiments, each LLM is given instructions without any prior examples or training tailored to the specific task. A specific example can be seen in Fig. 6.

6.2. LLM models

To provide a thorough assessment of the ability of LLMs to generate structured output, we evaluate several state-of-the-art LLMs from well-known companies. The selection criteria are based on the models' industry presence and their reported capabilities. The experimental models include notable ones developed by OpenAI and other famous open-source LLMs. A comparative description of these models is as follows:

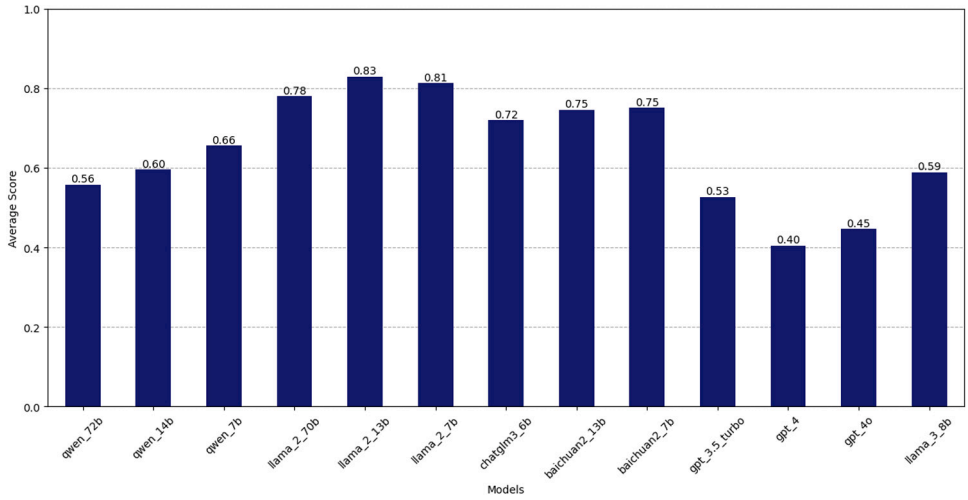


Fig. 7. Average scores of different LLMs.

- **GPT-4** (OpenAI, 2023) is a large multimodal model known for its reliability, creativity, and nuanced instruction handling, with improved safety and factual correctness over previous models.
- **GPT-3.5-Turbo** (OpenAI, 2022) is specifically designed for chat-based applications, offering efficient processing and refined language understanding at a lower cost.
- **GPT-4o** (OpenAI, 2024) is an optimized version of the GPT-4 model, specifically tailored for online interactions and real-time applications. It boasts enhanced speed and responsiveness, making it ideal for scenarios requiring rapid processing and immediate feedback.
- **Llama Series (2-70B, 2-13B, 2-7B, 3-8B)** (Touvron et al., 2023) includes models optimized for dialogue use cases, showing high performance in benchmarks and human evaluations for helpfulness and safety.
- **Baichuan Series (2-7B, 2-13B)** (Baichuan, 2023) features extensive multilingual models excelling in math, coding, medical, and legal domains, significantly outperforming their predecessor.
- **Qwen Series (7B, 14B, 72B)** (Bai et al., 2023) includes a range of foundational pretrained language models and chat models. These models excel across multiple tasks and areas, showcasing remarkable capabilities.
- **ChatGLM3-6B** (Du et al., 2022) offers a smooth dialogue experience and low deployment threshold, supporting multi-turn dialogues, code execution, and complex agent tasks in diverse scenarios.

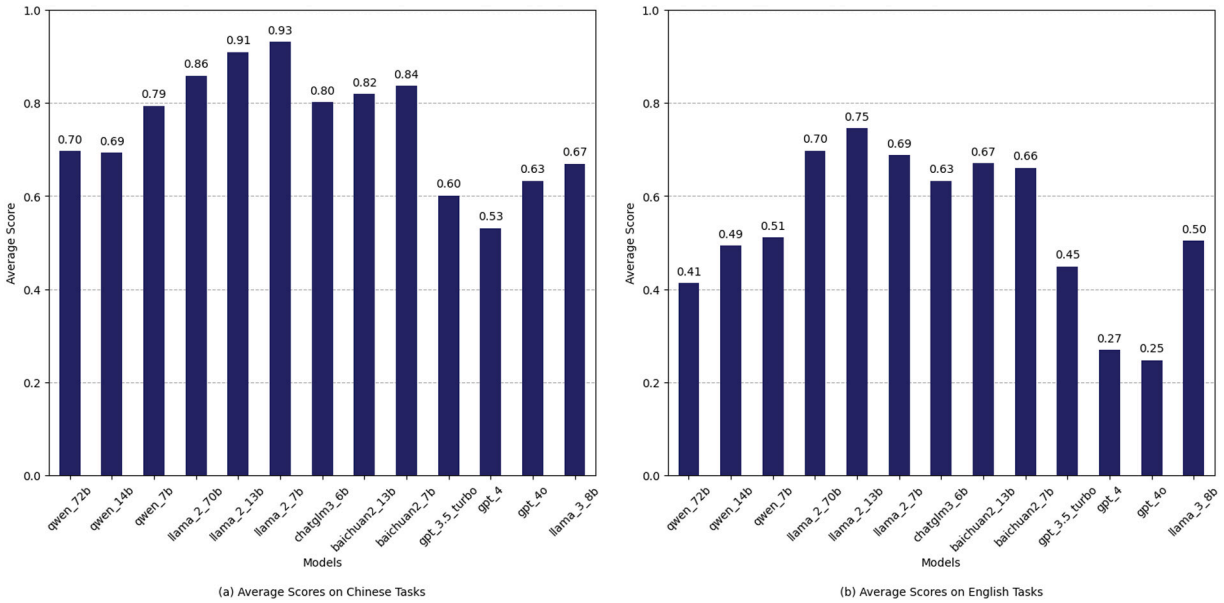
6.3. Evaluation measure

The evaluation of LLMs is based on the metrics defined in the 4.2 section. The primary criteria are the match with the regex pattern and the redundancy level of the LLM output content, offering quantitative insights into each model's performance. The evaluation aims to quantify the models' performance, allowing for a comparative analysis across different LLMs based on this standardized measure.

6.4. Results and discussion

Overall Results Analysis. The average scores obtained by various LLMs in our structured output benchmark experiments can be seen in Fig. 7. Our experiments has provided an intriguing perspective on the structured output generation capabilities of various LLMs, with the data indicating that a lower score is indicative of better performance. Among all the models, GPT-4 has the lowest score of 0.40, indicating superior alignment with the precision and constraints necessary for structured outputs. This performance may be attributed to its advanced training and nuanced understanding of complex instructions. The Llama Series, although highly proficient in other benchmarks, did not perform as well in this structured output task. Llama 2-13B and Llama 2-7B scored 0.83 and 0.81, respectively. This indicates that the models' dialogue optimization may not directly translate to structured output proficiency. The Baichuan Series and Qwen Series models have moderate scores in adhering to structured formats, with the Qwen 7B scoring 0.72. These scores indicate decent but not optimal performance. It is worth noting that the Baichuan Series models have multilingual capabilities. GPT-3.5-Turbo achieved a score of 0.53, which is commendable. However, there is room for improvement in structured output tasks. The varied scores across different models highlight the nuanced nature of model performance, suggesting that the ability to generate structured outputs is a distinct skill set within the broader spectrum of language processing capabilities.

Performance Analysis Across Different Language Tasks. Analyzing the benchmarking results depicted in Fig. 8 from the SoEval framework, a distinct pattern of performance across various language tasks becomes evident. The bar chart's two sections highlight



(a) Average Scores on Chinese Tasks

(b) Average Scores on English Tasks

Fig. 8. Average scores of different LLMs on Chinese and English tasks.

a disparity between tasks in Chinese and English, with lower scores indicating superior performance. In tasks involving the chinese language, LLMs generally achieved higher scores, suggesting difficulty with the intricacies of the language, with the Llama 2-7B model achieving the highest score at 0.93. Conversely, these same models exhibited notable enhancements in tasks involving the english language. Noteworthy is the outstanding performance of GPT-4 across both languages, particularly excelling in english with a top score of 0.27, highlighting its adaptability and finesse in generating structured outputs. These results indicate that the linguistic complexity and training of models on specific language datasets significantly influence their ability to produce accurate structured output, emphasizing the importance of tailored training and optimization for proficient multilingual LLM performance.

Comparing the Performance of Different Sized Models. Our research reveals fascinating correlations between the size of the model and its effectiveness. Contrary to the common assumption that larger models invariably yield better outcomes, our findings suggest a more nuanced reality. In tasks requiring structured output, we find that the larger Llama_2_13b model does not outperform its smaller 7b counterpart. This suggests that the distillation process for Llama models may have overlooked the importance of structured output capabilities. However, for the Qwen series of models, the ability to produce structured output consistently improves as model size increases. This comparison highlights a potential oversight in current research on large language models with respect to the ability to produce structured output.

Detailed Performance Analysis by Category and Subject. By analyzing the distribution of performance on structured output types and task subjects, as shown in Figs. 9 and 10, we can gain a more detailed understanding of the capabilities of LLMs within the SoEval framework. Each types and subjects present its unique challenges, with lower scores indicating superior model performance. The analysis shows that models like GPT-4 and GPT-4o excel particularly in handling "XML Format", "JSON Format", and "Formats Modify" tasks, achieving impressive performance in these areas. However, in more complex types such as timeline and keyValue, even advanced models such as GPT-4 show some scope for improvement. As shown in the figures, common models have the lowest compliance with zero-shot structured output instructions in "Attribute Graph", "Key-Value Pairs", and "Timeline". From the perspective of task subjects, models excel in areas such as "Science & Technology" and "Environmental Studies" due to the emphasis on accuracy and facts. This might be because such topics are more prevalent in the training data of these models. However, they face challenges in subjects like "Social Sciences" and "Human Rights", where concepts are more abstract, leading to higher complexity. Overall, these findings underscore the need for targeted enhancements in model training, particularly in structured output types and task subjects where performance currently lags.

7. Discussion and analysis

In this part, we primarily provide a comprehensive analysis of the experimental results. Simultaneously, we discuss the possible consequences of our study in theoretical and practical implications.

7.1. Results analysis

In this study, we introduce SoEval, a new benchmark designed to comprehensively evaluate the structured output capabilities of large language models. By applying standardized tasks and metrics, we can objectively compare models like GPT-4, GPT-3.5-Turbo, the Llama Series, the Baichuan Series, the Qwen Series, and ChatGLM3-6B in terms of their ability to generate structured

Fig. 9. Performance of different LLMs by structured output types.

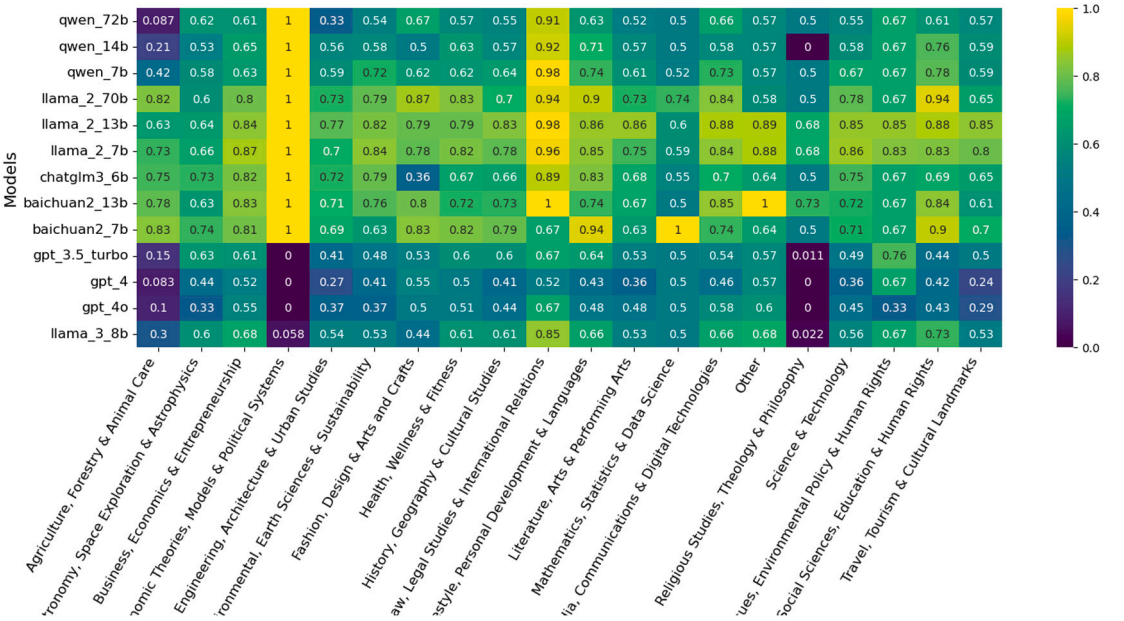


Fig. 10. Performance of different LLMs by task subjects.

outputs across different languages, task types, and subjects. The experimental results show that GPT-4 has the best performance in generating structured outputs that align with the required precision and constraints, achieving the lowest average score of 0.40. In contrast, Llama 2-13B and Llama 2-7B have higher scores of 0.83 and 0.81, respectively, indicating that their advantages in dialogue optimization may not directly translate to exceptional performance in structured output tasks. The Baichuan Series and Qwen Series models exhibit moderate performance, with Qwen 7B scoring 0.72. GPT-3.5-Turbo achieves a commendable score of 0.53, but there is still room for improvement in structured output tasks. Furthermore, we observe that large language models generally achieve higher scores on Chinese tasks, while showing significant improvements on English tasks. In particular, GPT-4 excels in English tasks, achieving the highest score of 0.27. In addition, the performance of different large language models varies across different

structured output types and task subjects, providing specific directions for future improvements in the structured output capabilities of large models.

7.2. Implications

Theoretical implications. This study's exploration into evaluating Large Language Models (LLMs) for their structured output capabilities offers significant theoretical contributions. By dissecting the prompt structure into task-related and structure-related components, we provide a novel theoretical framework for understanding how prompts influence LLMs' output. This approach not only enhances our comprehension of the interaction between various components of a prompt but also provides a methodical framework for prompt design, enabling more precise control over the LLM's output. Moreover, the introduction of regular expressions as a tool to define and assess structured outputs enriches the theoretical underpinnings of LLM evaluation, providing a quantifiable and systematic approach to measure models' adherence to desired output formats. Overall, our study contributes to the theoretical understanding of how LLMs can be more effectively evaluated, designed, and utilized for tasks requiring structured outputs.

Practical Implications. From a practical standpoint, our benchmark and its underlying methodology have profound implications for industries that rely on LLMs for data structuring and processing tasks. By providing a clear metric for structured output capabilities, SoEval enables developers and organizations to make informed decisions when selecting LLMs for specific applications, whether for content creation, data analysis, or automated reporting. In addition, for companies developing LLMs, our research emphasizes the importance of evaluating the structured output capabilities of large models, encouraging improvements from the data source and training methods to enhance these capabilities. Overall, our methodology holds significant practical value for both individuals and companies.

Research Implications. For the research community, This study makes significant contributions to the research field of evaluating LLMs' structured output capabilities. By developing the SoEval benchmark, we establish a standardized framework for assessing and comparing the performance of various models in generating structured outputs, laying the foundation for future research in this area. Moreover, our modular and automated pipeline for dataset construction, which leverages regular expressions and LLMs for data generation, offers a scalable and efficient approach that can inspire novel dataset construction techniques across various research fields. Overall, the research field of structured output capabilities in LLMs has been largely overlooked, and the introduction of SoEval aims to inspire more related research and enhance the structured output abilities of LLMs.

8. Conclusion and limitations

This paper introduces a new benchmark designed to evaluate the ability of Large Language Models (LLMs) to produce structured outputs, an aspect that is often overlooked but crucial. Initially, we analyze the structure of prompts and create a corresponding causal diagram, emphasizing the significance of both task-related instructions and structure-related instructions within the prompt. The SoEval dataset is developed with GPT-4.0 as the foundational tool for generating the initial dataset. This process is facilitated by a comprehensive pipeline starting from defining structured output formats and extending to manual verification and data editing. The dataset comprises 3.7K entries in both Chinese and English, encompassing 13 types of structured output tasks across 20 task subjects. During the experimental phase, we conduct a comparative analysis to evaluate the performance of current leading LLMs on the SoEval benchmark, presenting detailed performance metrics for each model. This research aims to enhance the understanding and advancement of structured output capabilities in LLMs.

While our benchmark sets a new standard for evaluating LLMs' ability to generate structured outputs, it has some shortcomings. One significant limitation is its narrow coverage of structured output formats. Moreover, the models assessed are mostly trained on English and Chinese datasets, potentially limiting their performance evaluation in diverse, multilingual contexts. Additionally, the complexity of tasks in SoEval varies, and some may not fully reflect real-world application complexities requiring structured outputs. Addressing these limitations in future research could improve the benchmark's comprehensiveness and relevance, offering more nuanced insights into LLMs' capabilities in generating structured outputs.

CRediT authorship contribution statement

Yu Liu: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Duantengchuan Li:** Writing – review & editing, Writing – original draft, Visualization, Investigation, Data curation, Conceptualization, Software, Supervision. **Kaili Wang:** Writing – review & editing, Formal analysis. **Zhuoran Xiong:** Conceptualization, Methodology, Writing – review & editing. **Fobo Shi:** Methodology, Writing – review & editing. **Jian Wang:** Writing – review & editing. **Bing Li:** Writing – review & editing, Funding acquisition, Conceptualization. **Bo Hang:** Writing – review & editing.

Data availability

Data will be made available on request.

References

Abdelali, A., Mubarak, H., Chowdhury, S. A., Hasanain, M., Mousi, B., Boughorbel, S., et al. (2024). LAraBench: Benchmarking arabic AI with large language models. In *Proceedings of the 18th conference of the European chapter of the association for computational linguistics: volume 1, long papers*. Malta: Association for Computational Linguistics.

Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., et al. (2023). Palm 2 technical report. arXiv preprint arXiv:2305.10403.

Austin, J., Odena, A., Nye, M., Bosma, M., Michalewski, H., Dohan, D., et al. (2021). Program synthesis with large language models. arXiv:2108.07732.

Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., et al. (2023). Qwen technical report. arXiv preprint arXiv:2309.16609.

Baichuan (2023). Baichuan 2: Open large-scale language models. arXiv preprint arXiv:2309.10305. Retrieved from https://arxiv.org/abs/2309.10305.

Bisk, Y., Zellers, R., Bras, R. L., Gao, J., & Choi, Y. (2019). PIQA: Reasoning about physical commonsense in natural language. arXiv:1911.11641.

Brown, P. F., Della Pietra, V. J., Desouza, P. V., Lai, J. C., & Mercer, R. L. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, *18*(4), 467–480.

Chandran, N., Sitaram, S., Gupta, D., Sharma, R., Mittal, K., & Swaminathan, M. (2024). Private benchmarking to prevent contamination and improve comparative evaluation of LLMs. arXiv preprint arXiv:2403.00393.

Chang, Y., Wang, X., Wang, J., Wu, Y., Zhu, K., Chen, H., et al. (2023). A survey on evaluation of large language models. arXiv preprint arXiv:2307.03109.

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. d., Kaplan, J., et al. (2021). Evaluating large language models trained on code. arXiv:2107.03374.

Clark, J. H., Choi, E., Collins, M., Garrette, D., Kwiatkowski, T., Nikolaev, V., et al. (2020). TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*.

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., et al. (2018). Think you have solved question answering? Try ARC, the AI2 reasoning challenge. arXiv:1803.05457.

Clark, C., Lee, K., Chang, M.-W., Kwiatkowski, T., Collins, M., & Toutanova, K. (2019). BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *NAACL*.

Introducing claude. (2023). Retrieved from https://www.anthropic.com/index/introducing-claude (Accessed 14 December 2023).

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., et al. (2021). Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168.

Dagdelen, J., Dunn, A., Lee, S., Walker, N., Rosen, A. S., Ceder, G., et al. (2024). Structured information extraction from scientific text with large language models. *Nature Communications*, *15*(1), 1418.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Dong, Z., Zhou, Z., Yang, C., Shao, J., & Qiao, Y. (2024). Attacks, defenses and evaluations for llm conversation safety: A survey. arXiv preprint arXiv:2402.09283.

Du, M., Luu, A. T., Ji, B., & Ng, S.-K. (2024). Mercury: An efficiency benchmark for LLM code synthesis. arXiv preprint arXiv:2402.07844.

Du, Z., Qian, Y., Liu, X., Ding, M., Qiu, J., Yang, Z., et al. (2022). GLM: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 320–335).

Fei, Z., Shen, X., Zhu, D., Zhou, F., Han, Z., Zhang, S., et al. (2023). LawBench: Benchmarking legal knowledge of large language models. arXiv preprint arXiv:2309.16289.

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, *30*, 681–694.

Goyal, N., Gao, C., Chaudhary, V., Chen, P.-J., Wenzek, G., Ju, D., et al. (2021). *The FLORES-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation*. arXiv preprint arXiv:2106.03193.

Gu, Z., Zhu, X., Ye, H., Zhang, L., Wang, J., Jiang, S., et al. (2023). Xiezhi: An ever-updating benchmark for holistic domain knowledge evaluation. arXiv preprint arXiv:2306.05783.

Guidotti, R., Monreale, A., Turini, F., Pedreschi, D., & Giannotti, F. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, *51*, 1–42. http://dx.doi.org/10.1145/3236009.

Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D., et al. (2021). Aligning AI with shared human values. In *Proceedings of the International Conference on Learning Representations*. ICLR.

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., et al. (2021). Measuring mathematical problem solving with the MATH dataset. *NeurIPS*.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. http://dx.doi.org/10.1162/neco.1997.9.8.1735.

Huang, Y., Bai, Y., Zhu, Z., Zhang, J., Zhang, J., Su, T., et al. (2024). C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, *36*.

Izquierdo, J. L. C., & Cabot, J. (2014). Composing JSON-based web APIs. (pp. 390–399). http://dx.doi.org/10.1007/978-3-319-08245-5_24.

Jöreskog, K. G., & Sörbom, D. (1982). Recent developments in structural equation modeling. *Journal of Marketing Research*, *19*(4), 404–416.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., et al. (2019). Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, *7*, 453–466.

Lai, G., Xie, Q., Liu, H., Yang, Y., & Hovy, E. (2017). Race: Large-scale reading comprehension dataset from examinations. arXiv preprint arXiv:1704.04683.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, *55*(9), 1–35.

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., et al. (2024). Large language models: A survey. arXiv preprint arXiv:2402.06196.

Ning, K.-P., Yang, S., Liu, Y.-Y., Yao, J.-Y., Liu, Z.-H., Wang, Y., et al. (2024). Peer-review-in-LLMs: Automatic evaluation method for LLMs in open-environment. arXiv preprint arXiv:2402.01830.

OpenAI (2022). Introducing ChatGPT. https://openai.com/blog/chatgpt. (Accessed 2023).

OpenAI (2023). GPT-4 technical report. arXiv:2303.08774.

OpenAI (2024). Hello GPT-4o. Retrieved from https://openai.com/index/hello-gpt-4o/.

Pilehvar, M. T., & Camacho-Collados, J. (2019). WiC: The word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of NAACL 2019 (short)*.

Sheng, E., Chang, K.-W., Natarajan, P., & Peng, N. (2021). Societal biases in language generation: Progress and challenges. arXiv preprint arXiv:2105.04054.

Sun, K., Yu, D., Yu, D., & Cardie, C. (2020). Investigating prior knowledge for challenging Chinese machine reading comprehension. *Transactions of the Association for Computational Linguistics*, Retrieved from https://arxiv.org/abs/1904.09679v3.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., et al.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Wan, Z., Cheng, F., Mao, Z., Liu, Q., Song, H., Li, J., et al. (2023). Gpt-re: In-context learning for relation extraction using large language models. arXiv preprint arXiv:2305.02105.

Wang, B., Chen, W., Pei, H., Xie, C., Kang, M., Zhang, C., et al. (2023). DecodingTrust: A comprehensive assessment of trustworthiness in GPT models. arXiv preprint arXiv:2306.11698.

Wang, R., Chen, H., Zhou, R., Ma, H., Duan, Y., Kang, Y., et al. (2024). LLM-detector: Improving AI-generated Chinese text detection with open-source LLM instruction tuning. arXiv preprint arXiv:2402.01158.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., et al. (2019). Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in Neural Information Processing Systems*, *32*.

Wu, T., Luo, L., Li, Y.-F., Pan, S., Vu, T.-T., & Haffari, G. (2024). Continual learning for large language models: A survey. arXiv preprint arXiv:2402.01364.

Xu, L., Hu, H., Zhang, X., Li, L., Cao, C., Li, Y., et al. (2020). CLUE: A Chinese language understanding evaluation benchmark. arXiv preprint arXiv:2004.05986.

Xu, L., Li, A., Zhu, L., Xue, H., Zhu, C., Zhao, K., et al. (2023). Superclue: A comprehensive chinese large language model benchmark. arXiv preprint arXiv:2307.15020.

Xu, L., Xue, H., Zhu, L., & Zhao, K. (2024). SuperCLUE-Math6: Graded multi-step math reasoning benchmark for LLMs in Chinese. arXiv preprint arXiv:2401.11819.

Yang, S., Zhao, B., & Xie, C. (2024). AQA-bench: An interactive benchmark for evaluating llms' sequential reasoning ability. arXiv preprint arXiv:2402.09404.

Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z., & Zhang, Y. (2024). A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, Article 100211.

Zaremba, W., Sutskever, I., & Vinyals, O. (2014). Recurrent neural network regularization. arXiv preprint arXiv:1409.2329.

Zhang, X., Li, C., Zong, Y., Ying, Z., He, L., & Qiu, X. (2023). Evaluating the performance of large language models on gaokao benchmark.

Zhang, X., Yu, B., Yu, H., Lv, Y., Liu, T., Huang, F., et al. (2023). Wider and deeper llm networks are fairer llm evaluators. arXiv preprint arXiv:2308.01862.

Zhao, H., Yang, F., Lakkaraju, H., & Du, M. (2024). Opening the black box of large language models: Two views on holistic interpretability. arXiv preprint arXiv:2402.10688.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., et al. (2023). A survey of large language models. http://dx.doi.org/10.48550/arXiv.2303.18223, ArXiv, abs/2303.18223.

Zhong, W., Cui, R., Guo, Y., Liang, Y., Lu, S., Wang, Y., et al. (2023). Agieval: A human-centric benchmark for evaluating foundation models. arXiv preprint arXiv:2304.06364.