applied sciences

MDPI

Article

Holistic Spatial Reasoning for Chinese Spatial Language Understanding

Yu Zhao * and Jianguo Wei

Collage of Intelligence and Computing, Tianjin University, Tiainjin 300350, China
* Correspondence: zhaoyucs@tju.edu.cn

Abstract: Spatial language understanding (SLU) is an important task in the field of information extraction, and it involves complex spatial analyses and reasoning processes. Unlike English SLU, in the case of the Chinese language, there may be some language-specific challenges, such as the phenomenon of polysemia and the substitution of synonyms. In this work, we explore Chinese SLU by taking advantage of large language models. Inspired by recent chain-of-thought (CoT) strategies, in this study, we propose the Spatial-CoT template to help improve LLMs' reasoning abilities in order to deal with the challenges of Chinese SLU. Spatial-CoT offers LLMs three steps of instructions from different perspectives, namely, entity extraction, context analysis, and common knowledge analysis. We evaluate our framework on the Chinese SLU dataset SpaCE, which contains three subtasks: abnormal spatial semantics recognition, spatial role labeling, and spatial scene matching. The experimental results show that our Spatial-CoT outperforms vanilla prompt learning on ChatGPT and achieves competitive performance in comparison with traditional supervised models. Further analysis revealed that our method could address the phenomenon of polysemia and the substitution of synonyms in Chinese spatial language understanding.

Keywords: spatial language understanding; artificial intelligence; natural language processing; large language model

Citation: Zhao, Y.; Wei, J. Holistic Spatial Reasoning for Chinese Spatial Language Understanding. *Appl. Sci.* **2023**, *13*, 11712. https://doi.org/10.3390/app132111712

Academic Editor: Andrea Prati

Received: 23 September 2023
Revised: 23 October 2023
Accepted: 25 October 2023
Published: 26 October 2023

1. Introduction

Spatial language understanding (SLU) involves recognizing and reasoning about spatial semantics, e.g., spatial objects, relations, and transformations [1], in natural languages' descriptions, and this has drawn much attention in recent years. Conventional SLU tasks include spatial role labeling [2,3], spatial question answering [4], and spatial reasoning [5]. SLU is an essential function of natural language processing systems, and it is necessary for many applications, such as robotics [6–8], navigation [9–11], traffic management [12,13], and query answering systems [4].

In the realm of SLU, extant research exhibits a pronounced concentration on the context of the English language, while there is a noticeable lack of investigations on non-English settings. Extending in this direction, Zhan et al. [14] proposed Spatial Cognition Evaluation (SpaCE) in order to benchmark Chinese spatial language understanding via abnormal spatial semantics recognition, spatial role labeling, and spatial scene matching. Unlike in the context of English, Chinese SLU has some specific challenges due to its ambiguous characteristics and complex grammar. For example, Figure 1a shows the phenomenon of polysemia among Chinese spatial prepositions. The word "上" in the sentence "书架上面放着书" (the books on the shelf) means "on", while the "上" in the sentence '书架上面挂着钟"' (the clock hanging above the shelf) means "above". With the same word "上" or phrase "上面", the sentences may express different spatial relationships, which is a common phenomenon in Chinese. Meanwhile, some conflicting prepositions may express the same meaning, which is also known as the substitution of synonyms. In Figure 1b, the "上" (which usually denotes "on") in "地上有一枚硬币" and the "下" (which usually denotes

"under") in "地下有一枚硬币" describe the same thing, i.e., "there is a coin on the ground". The substitution of synonyms occurs in many languages but does so more frequently in Chinese due to the lax usage of syntax. In these situations, traditional methods that extract spatial information based on spatial prepositions impair performance. The solution to this issue is to reason spatial relationships with the holistic semantics of a sentence and the common linguistic knowledge of the Chinese language.

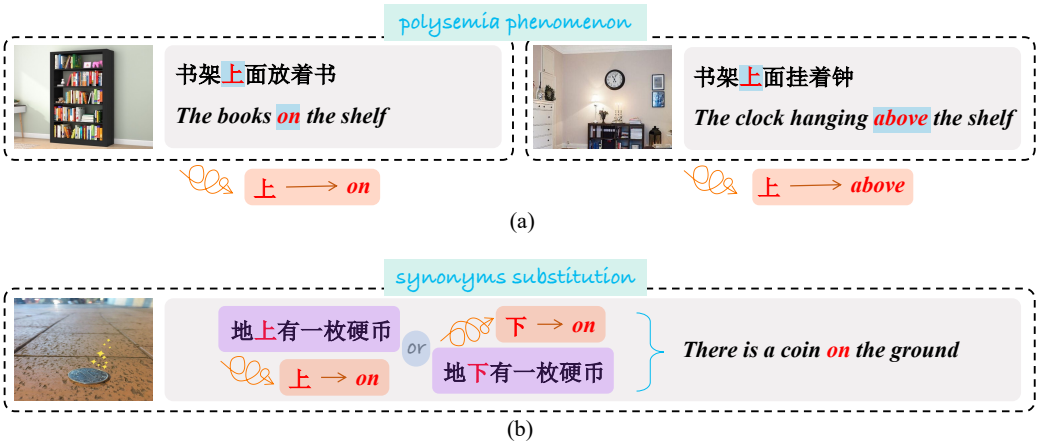

Figure 1. The cases of the phenomenon of polysemia (**a**) and the substitution of synonyms (**b**) in the Chinese language.

Very recently, the integration of LLMs into the NLP task has garnered increasing attention, which has facilitated the achievement of remarkable zero-/few-shot performance. Researchers have devised mechanisms such as in-context learning (ICL) that enable LLMs to directly generate results for an input question with just a few demonstrations. However, the approach of straightforward prompts falls short of eliciting the profound language comprehension and knowledge manipulation capabilities of LLMs. Consequently, it struggles to effectively address intricate and challenging spatial reasoning. In this study, we explore the newly emerged chain-of-thought (CoT) mechanism in which LLMs not only produce the most probable reasoning results, but also provide clues/the rationale underpinning an inference. This involves incorporating the holistic spatial information of the input descriptions, as well as the extensive common knowledge that LLMs have. The vanilla CoT simply concatenates the target problem statement with "Let us think step by step" as an input prompt for LLMs. It fails to explicitly focus on the key steps of spatial reasoning, leading to unstable results. In this study, we propose the Spatial-CoT template, where we manually decompose the problems of SLU into a three-step process to address the specific issues in spatial reasoning.

Concretely, we investigate CoT reasoning for three Chinese spatial understanding tasks: abnormal spatial semantics recognition (ASpSR), spatial role labeling (SpRL), and spatial scene matching (SpSM). For each task, we design a CoT prompt template, namely, Spatial-CoT, from a shallow entity to holistic semantics. Technically, our Spatial-CoT reasoning comprises three pivotal steps, with each being a special analyzer: an entity analyzer (EA), context analyzer (CA), and common knowledge analyzer (CKA). The overall framework is shown in Figure 2. First, the EA should recognize all of the related entities in the spatial scene of a given description, providing entity-level clues for deeper reasoning. Subsequently, the CA employs the entity clues and makes a prediction based on the provided context, making a preliminary decision about the task from its inclination. Finally, the CKA refines the decision based on common knowledge and language conventions. At its core, Spatial-CoT unfolds as a sequence of logical steps, with each building upon the insights of the preceding step, leading to conclusive outcomes.

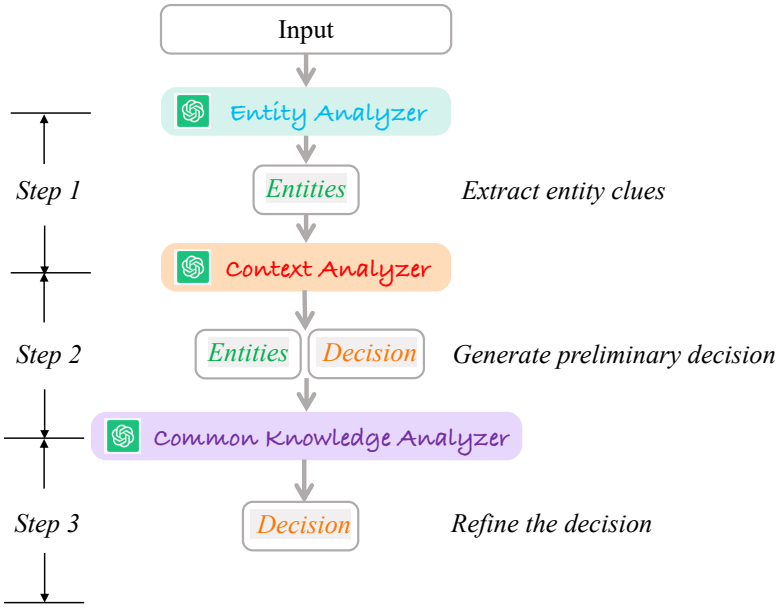

Figure 2. The overall framework of our chain-of-thought spatial reasoning.

We evaluated our method on the SpaCE dataset, which includes an abnormal spatial semantics recognition task, spatial role labeling task, and spatial scene matching task. The results underscore the efficacy of our proposed Spatial-CoT reasoning framework in LLM-based Chinese SLU. The Spatial-CoT method outperformed vanilla prompts and surpassed the current state-of-the-art supervised baseline.

The rest of this study is organized as follows. Section 2 describes the related work. Section 3 introduces the details of the designs of the CoT three subtasks. Section 4 compares the performance of the current approach with that of other models. Section 5 summarizes the conclusions.

2. Related Work

Spatial language understanding is a fundamental task in natural language understanding. One of the essential functions of natural language is the expression of spatial relationships between objects. An early method of spatial information extraction was introduced in the form of spatial role labeling (SpRL) by Kordjamshidi et al. [2]. Furthermore, SemEval2012 introduced the SRL task while mainly focusing on static spatial relations, and SemEval-2013 expanded static spatial relations to capture fine-grained semantics. SemEval-2015 [15] first proposed the evaluation of implementation systems for the SpaceEval annotation scheme, which is the current spatial information annotation scheme, and many spatial information extraction systems have been developed based on it [16–18]. Multimodal approaches to spatial understanding have also been recently proposed. In CLEF 2017, Kordjamshidi et al. [19] proposed multimodal spatial role labeling (mSpRL). The authors of Zhao et al. [20,21] proposed the task of visual spatial description, thus extending the line of spatial language understanding. Differing from most efforts that focused on the English context, CCL2021 (http://ccl.pku.edu.cn:8084/SpaCE2021/task, accessed on 1 September 2023) was the first proposal of the SpaCE task, which used a non-English language for spatial language understanding. In this study, we follow this work and propose an LLM-based framework for better spatial reasoning.

Prior research on spatial language understanding heavily relied on supervised learning and fine-tuned pre-trained models while using specific datasets with extensive labeled data [2,22,23], thus achieving remarkable performance. While traditional supervised frameworks face challenges such as the need for extensive labeled data and weak generalizability, recent LLMs have emerged as a solution to these issues. LLMs (e.g., ChatGPT [24], LLaMA [25], and Vacuna [26]) have garnered unprecedented attention and exhibit human-

level language understanding, thus amassing rich linguistic and commonsense knowledge. LLMs equipped with the prompt learning and ICL paradigm [27–32] have demonstrated robust performance in scenarios with little or even no supervision, which is a remarkable trait. In the context of LLM-based methods, the initial research involved constructing prompts to inquire targets directly. To address the need for in-depth reasoning, researchers have turned to the recent CoT strategy [33–36] to enhance LLMs' inference capabilities. CoT empowers LLMs to engage in human-like reasoning processes, thus not only providing answers, but also furnishing the underlying thought process behind those answers.

However, conventional CoT simply employs a single prompt template such as "let us think step by step" to inspire the LLM [37–39]. These methods improve reasoning performance on larger LLMs but fail to indicate the explicit reasoning steps for the provided task. Beyond single prompt CoT, Wei et al. [40] encourage internal dialogue by forcing the LLM to generate a sequence of intermediate steps for reasoning problems. Zhou et al. [41] go a step further; they (automatically) break a complex problem into simpler sub-problems and then solve them in sequence. Previous multi-step CoT methods primarily focus on arithmetic problems or logical problems, which are relatively easily symbolized and generalized. However, these strategies may struggle to maintain control over intermediate reasoning steps when confronted with diverse tasks, making them less applicable to our spatial understanding problem. Thus, in this study, we focused on the spatial understanding problem and designed the Spatial-CoT framework, aiming to solve the challenges of the phenomenon of polysemia and the substitution of synonyms in the Chinese language that previous methods have failed to address.

3. Methodology

In this section, we first provide the definitions of the three Chinese SLU tasks that we studied in this work. Then, we will introduce the existing LLM-based methods of solving these SLU problems for comparison. Finally, we illustrate our CoT framework for each subtask.

3.1. Task Definition

We explored three types of tasks in Chinese SLU: abnormal spatial semantics recognition (ASpSR), spatial role labeling (SpRL), and spatial scene matching (SpSM). Given the input textual description, X, the ASpSR task aims to output whether X has unreasonable expressions, such as self-contradictions and counterintuitive expressions. SpRL is similar to semantic role labeling (SRL) but replaces the semantic role labels with a set of spatial role labels, such as "Spatial Entity", "Event", and "Time". The pre-defined spatial role labels in SpaCE are listed in Table 1. Finally, the SpSM task takes two different descriptions, X_1 and X_2, and determines whether X_1 and X_2 describe the same scene.

Table 1. Spatial role labels in SpaCE.

Role Label	Type	Description	Role Label	Type	Description
Spatial Entity	Text Span	The entity described in the scene	Direction	Text Span	The direction of a dynamic entity
Reference Entity	Text Span	The entity used as the reference object	Orientation	Text Span	The orientation of an entity
Event	Text Span	The event related to the spatial entity	Component Location	Text Span	The location of part of an entity with respect to the whole
Facticity	Extra Label	Whether the description is true	Part	Text Span	The part of an entity
Time	Text Span	The time that the entity is in some spatial state	Shape	Text Span	The shape of an entity
Location	Text Span	Description of the location of a static entity	Route	Text Span	The route of a moving entity
Starting Point	Text Span	Description of the starting location of a dynamic entity	Distance	Text Span	The distance between entities
Destination Point	Text Span	Description of the ending location of a dynamic entity			

3.2. Preliminary on In-Context Learning

Vanilla Prompt. The vanilla prompt directly asks the LLM to generate a judgment of the input description. The template of a vanilla prompt can be described as follows:

> **ASpSR:** *Given the context, please answer whether it has any abnormal expressions.*
> **SpRL:** *Given the context, please label each word with one of the following role labels. Options: [. . .]*
> **SpSM:** *Given the two descriptions, please answer whether they describe the same spatial scene.*

The vanilla prompt is concise, but it is hard for it to explore the reasoning potential of the LLM.

In-Context Learning. An LLM based on the ICL paradigm can show remarkable performance in many NLP tasks by offering a few demonstrations in the prompt. To better take advantage of the reasoning abilities of LLMs, some studies incorporated reasoning into ICL demonstrations. Wan et al., 2023 [27] let an LLM induce the reasoning for demonstrations by using gold labels, which can be described as follows:

> *Given the context, please do the task xxx. Here is an example: Input [. . .] Output [. . .]. Input: [. . .]*

However, this reasoning design makes the LLM incapable of complex reasoning tasks and greatly depends on the quality of the demonstrations. Finally, it has difficulty dealing with zero-shot scenarios.

3.3. Spatial Chain-of-Thought Framework

The prompt in the vanilla ICL only contains input–output pairs for the target task. This demands that the model learns the task in one go, which becomes more challenging as the task grows more complex. Existing studies on chain-of-thought methods reveal that prompts containing step-by-step instructions assist the model in decomposing the task and making better predictions. Now, we present our CoT prompt design. However, vanilla CoT simply appends the "Let us think step by step" instruction to the target problems fed to LLMs [37], which may cause missing-step errors and misunderstanding errors. For Chinese SLU tasks, the holistic context and external language usage must be taken into consideration, and missing steps or semantic misunderstandings may lead to suboptimal outputs. Thus, we propose the Spatial Chain-of-Thought (Spatial-CoT) framework, which explicitly prompts by using instructions for each reasoning step. Inspired by one's own thought process when solving a spatial reasoning task, we introduced a three-step prompt in order to decompose complex spatial reasoning problems into simpler sub-problems.

Then, we introduced the prompts of the three tasks as follows.

Step-1: Entity Analyzer. The entity analyzer (EA) template is used to extract entities. Generally, to solve SLU problems while avoiding errors resulting from missing attention on key entities, this step aims to construct templates to extract all of the related entities in the scenes in input descriptions. Thus, the prompt would be as follows:

> **ASpSR:** *You are acting as an entity analyzer. Based on the description, please list all of the entities in this scene. Description: [input description]*
> **SpRL:** *You are acting as an entity analyzer. Based on the description, please list all of the entities in this scene. Description: [input description]*
> **SpSM:** *You are acting as an entity analyzer. Based on the two pieces of description, list all of the entities in the two scenes. Description 1: [input description] Description 2: [input description]*

Step-2: Context Analyzer. The step of the context analyzer (CA) is used to generate the preliminary decision of a task based on the holistic perspective of the described scene by leveraging the entity clues generated by the EA. In this step, "pay attention to the whole scene" is added to the trigger sentence to request the LLMs to perform holistic spatial

reasoning as accurately as possible. In the SpRL task, we also include the candidate role labels. The templates of CA are designed as follows:

> **ASpSR:** *You are acting as a context analyzer. Based on the description and the entity clues generated previously, please tell whether there are any abnormal expressions in the given description. Pay attention to the whole scene. Description: [input description]*
> **SpRL:** *You are acting as a context analyzer. Based on the description and the entity clues generated previously, please label the related roles in the description. Pay attention to the whole scene. The role labels should be chosen from [R]. Description: [input description]*
> **SpSM:** *You are acting as a context analyzer. Based on the descriptions and the entity clues generated previously, please tell whether the two pieces of text describe the same scene. Pay attention to the whole scene. Description 1: [input description], Description 2: [input description]*

Step-3: Common Knowledge Analyzer. The entities and context-based preliminary decisions generated in the first two steps are used to refine and generate cogitative results while considering common sense and language knowledge. In this step, the three tasks adopt the same instruction, as "pay attention to the Chinese-language characters" is added to help LLMs focus on Chinese language usage. The templates of the CKA are designed as follows:

> **ASpSR/SpRL/SpSM:** *You are acting as a common knowledge analyzer. Based on the description, the generated entity clues, and the preliminary decision, please check whether the preliminary decision makes more sense depending on common sense and the language usage. In addition, generate rationales for the final decision. Pay attention to the Chinese-language characters.*

4. Experiment

4.1. Experimental Setting

Dataset. We evaluated our proposed method on the SpaCE [14] dataset, a Chinese spatial language understanding corpus that comprises more than 10,000 examples. SpaCE encompasses a wide range of domains, such as news, literature, textbooks, sports corpuses, and encyclopedias. The dataset comprises annotations of abnormal spatial semantics recognition, spatial role labeling, and spatial scene matching. The detailed dataset statistics are listed in Table 2.

Table 2. The statistics of the SpaCE dataset. "#SENT" means the number of sentences. "#SPAN" means the number of SRL spans. "#ROLE" means the number of role types.

Split	ASpSR	SpRL			SpSM
	#SENT	#SENT	#SPAN	#ROLE	#SENT
Train	10,993	1529	25,108	15	-
Test	1602	700	3815	15	10

In this work, we adopted the official dev split for the test set and processed zero-shot/1-shot CoT. We also reproduced some supervised models for comparison, and we split the official training set into 7:1 for training and validation. For the SpSM task, there was no training set, but 10 examples were used for evaluation (https://2030nlp.github.io/SpaCE2023/index.html, accessed on 1 September 2023), and we only reported its results with the LLM-based methods.

Baselines.We provide a comparison of our proposed method with extensive baseline methods. We first report results on conventional supervised methods, such as **BERT-like** language models. We adopted a state-of-the-art framework, MRC-SRL [42] for the SpSRL task. The MRC-SRL treat the SRL task as an multiple-choice machine reading comprehension (MRC) problem [43]. For ASpSR, we used a conventional BERT-based text classification framework [44], where the pooled outputs of the BERT backbone are used to predict whether the given sentence has abnormal expressions through a binary classification layer.

We also include the results of a prompt-learning paradigm for the LLM. We report three types of prompting baselines: (1) **Vanilla Prompt (zero-shot).** We include vanilla prompt learning [45], where only the description of the task's target is fed to the LLM as an input. (2) **In-Context Learning (ICL).** In-context learning [27] adds demonstrations to the prompt template, thus strongly guiding the generation of output answers for desired tasks. (3) **Vanilla CoT (zero-shot).** Following reference [37], vanilla CoT appends "Let us think step by step" to the prompt with or without any demonstration examples, guiding the LLM to decompose a problem into steps. (4) **Vanilla CoT-ICL.** Based on zero-shot CoT, CoT-ICL add demonstrations to further guide the generation in an ICL paradigm. Technically, we list the input template of the three methods in Table 3.

Implementation Details. For supervised ASpSR, we use a BERT-base backbone and a binary linear classifier. For supervised SpRL, we follow reference [42] to use a BERT-base model as the base encoder and use two special symbols and to mark the predicate of the input sentence. We adopt Adam as optimizer, and the warmup rate is 0.05, the initial learning rate is 1×10^{-5}. For LLM-based methods, we select ChatGPT3.5 (https://openai.com/blog/chatgpt, accessed on 1 September 2023), which is the top-performing open-source large language model. The temperature value of ChatGPT is set to 0.2. The detailed parameters are presented in Appendix B.

Evaluation. For the ASpSR and SpSM tasks, we use the accuracy as the evaluation metric. For the SpRL task, we use the F1 score to evaluate the sequence labeling performance; this is denoted as RL-F1, and the extra label classification is denoted as Ext-F1. We also follow the human evaluation of SpaCE on CCL2023 to evaluate the rationales of the SpSM task. Concretely, we involved two evaluators to score the generated explanation for each SpSM question. A higher score indicates a clearer explanation of the reasoning behind identifying similarities and differences in spatial scenes. The scores are categorized into six levels ranging from 0 to 5:

- **5:** Using external knowledge and worldly understanding to restate spatial scenes.
- **4:** Rewrite the differences between the two descriptions and restate spatial scenes.
- **3:** When determining that two descriptions are identical, the differing string is directly used as the reason without rewriting it, without restating the spatial scene.
- **2:** (a) The explanations provided are related to spatial meaning but not specific to the spatial scene where the differing string is located; (b) When determining that two descriptions are NOT identical, the differing string is directly used as the reason without rewriting it, without restating the spatial scene.
- **1:** The explanations provided are NOT related to spatial meaning.
- **0:** (a) No explanations or the explanations are not related to the questions; (b) Wrong differing string.

We report the average value of the two evaluators as the final score.

Table 3. Three templates of prompting baselines for SpaCE.

Methods	Task	Template
Vanilla Prompt (zero-shot)	ASpSR	[input description] Given the context, please answer whether it has any abnormal expressions.
	SpRL	[input description] Given the context, please label each word with one of the following role labels. Options: [. . .]
	SpSM	[input description 1] [input description 2] Given the two descriptions, please answer whether they describe the same spatial scene.
ICL	ASpSR	Given the context, please answer whether it has any abnormal expressions. For example, Q: [example description] A: [Yes/No.] Q: [input description] A:
	SpRL	Given the context, please label each word with one of the following role labels. Options: [label set] For example, Q: [example description] A: [SpRL labeling results]. Q: [input description] A:
	SpSM	Given the two descriptions, please answer whether they describe the same spatial scene. For example, Q: [example description 1] [example description 2] A: [Yes/No.] Q: [input description 1] [input description 2]

Table 3. *Cont.*

Methods	Task	Template
Vanilla CoT (zero-shot)	ASpSR	[input description] Given the context, please answer whether it has any abnormal expressions via a step by step reasoning.
	SpRL	[input description] Given the context, please label each word with one of the following role labels. Options: [. . .] Please extract the roles via step by step reasoning.
	SpSM	[input description 1] [input description 2] Given the two descriptions, please answer whether they describe the same spatial scene via a step by step reasoning.
Vanilla CoT-ICL	ASpSR	Given the context, please answer whether it has any abnormal expressions via a step by step reasoning. For example, Q: [example description] A: [Let us think step by step . . . So, the description has/doesn't has an abnormal expression]. Q: [input description]
	SpRL	Given the context, please label each word with one of the following role labels. Options: [. . .] Please extract the roles via step by step reasoning. For example, Q: [example description] A: [Let us think step by step . . . So the labeling results are . . .] Q: [input description]
	SpSM	Given the two descriptions, please answer whether they describe the same spatial scene via a step by step reasoning. For example, Q: [example description 1] [example description 2] A: [Let us think step by step . . . So, the two descriptions describe/does not describe the same scene.] Q: [input description 1] [input description 2]

4.2. Main Results

Table 4 details the performance of the baselines and our CoT method. We observed that the vanilla prompts for LLMs were inferior to existing supervised methods, but the ICL and CoT knowledge significantly boosted it, especially when at least one demonstration was used. Our Spatial-CoT method further outperformed the others on all three spatial language understanding tasks. Overall, SpRL was more difficult than ASpSR and SpSM.

Table 4. Results of three tasks on seven SpaCE datasets. BERT/MRC-SRL lines mean that we use BERT classifier for ASpSR task and use MRC-SRL for SpRL task. We report the standard deviation in the brackets. **Bold:** best result.

	ASpSR	SpRL		SpSM	
	Acc	RL-F1	Ext-F1	F1	Rationale Score (Human)
• ***Fine-tuning baselines***					
BERT/MRC-SRL (1-shot) [42]	50.33 ± 5.25 (46.85)	9.72 ± 2.17 (54.27)	6.19 ± 2.03 (49.82)	-	-
BERT/MRC-SRL (Full) [42]	78.26 ± 1.89 (58.99)	46.49 ± 0.13 (39.94)	49.10 ± 0.81 (36.04)	-	-
• ***Prompt-based methods***					
Vanilla Prompt (zero-shot)	71.19 ± 0.07 (37.05)	43.01 ± 0.36 (44.54)	46.20 ± 0.25 (46.85)	35.10 ± 0.21 (39.32)	1.79
Vanilla CoT (zero-shot)	73.29 ± 0.02 (38.10)	44.13 ± 0.48 (46.31)	47.12 ± 0.25 (39.94)	35.66 ± 0.32 (48.03)	2.64
ICL (1-shot)	77.13 ± 0.48 (49.38)	45.29 ± 1.02 (58.33)	49.15 ± 2.21 (52.66)	39.29 ± 1.86 (52.45)	2.77
Vanilla CoT-ICL (1-shot)	80.30 ± 0.70 (46.08)	46.04 ± 1.52 (59.95)	51.28 ± 1.89 (58.63)	43.40 ± 0.18 (57.62)	2.80
• ***Ours***					
Spatial-CoT (zero-shot)	80.03 ± 0.02 (35.40)	46.19 ± 0.03 (39.72)	52.91 ± 0.07 (28.79)	45.12 ± 0.02 (40.02)	3.01
Spatial-CoT-ICL (1-shot)	**81.41 ± 0.12 (28.71)**	**48.61 ± 0.08 (48.10)**	**55.33 ± 0.15 (43.22)**	**47.76 ± 0.03 (49.36)**	**3.72**

Under the one-shot setting, the vanilla prompt and vanilla CoT methods were transformed into ICL-based methods, and these are denoted as ICL and vanilla CoT-ICL. The demonstration is randomly selected from the training set. We can see that the ICL could enhance the LLM significantly when one demonstration was provided.

Specifically, with the zero-shot setting, the LLM-based methods were comparable to the full-setting supervised method and even outperformed it on the SpRL task, demonstrating the generation capabilities of LLMs. Moreover, our Spatial-CoT showed the following scores: 8.84 ASpSR ACC, 2.96 RL-F1, 3.01 Ext-F1, and 8.02 SpSM F1. These were absolute improvements on the SpaCE dataset in comparison with the vanilla prompt methods. The ICL method was able to enhance the prompt-based and CoT methods by providing a demonstration to guide the LLM, earning 9.11 ASpSR Acc, 0.06 RL-F1, and 2.19 SpSM F1

for the vanilla prompt, and 1.39 ASpSR Acc, 2.42 RL-F1, 2.42 Ext-F1, and 2.64 SpSM F1 for Spatial CoT.

For the SpSM task, there was no training set; thus, we only report the few-shot/zero-shot performance of the LLM-based methods. We also report the human evaluations of the generated rationales. Scores from 0 to 5 were chosen, where 5 was the highest. We recruited two students to score the generated rationales and report the average values. Compared with the vanilla prompt or CoT methods, our proposed Spatial-CoT was also able to generate more explicit rationales.

4.3. Analysis

4.3.1. Few-Shot Analysis

To explore the impact of the scale of the training set and demonstrations, we compare the few-shot performance of each method. In few-shot supervised methods, i.e., BERT/MRC-SRL, we randomly select 1, 2, 4, 8, and 16 samples from the original training set for fine-tuning. In few-shot LLM-ICL methods, the selected samples with their ground-truth are used as demonstrations.

The results are shown in Figure 3. For the BERT-based methods, the x-axis denotes the number of training samples, while for the ICL-based methods, it denotes the number of demonstrations. We found that, with the few-shot setting, the LLM-based methods outperformed the supervised method, although the gap gradually decreased with an increasing number of training samples. In settings with a small number of samples, the supervised model may not have been well trained, while the LLM-based ICL was able to leverage rich common knowledge to process the reasoning. The CoT method further outperformed the others due to the rational design of the fine-grained reasoning steps. The curve of the supervised method was steeper than those of the LLM-based methods, especially on the 0 point. This demonstrated that the supervised model was sensitive to the scale of the training sample. In contrast, the ICL-based methods were almost unaffected by the number of demonstrations, as long as at least one was provided.

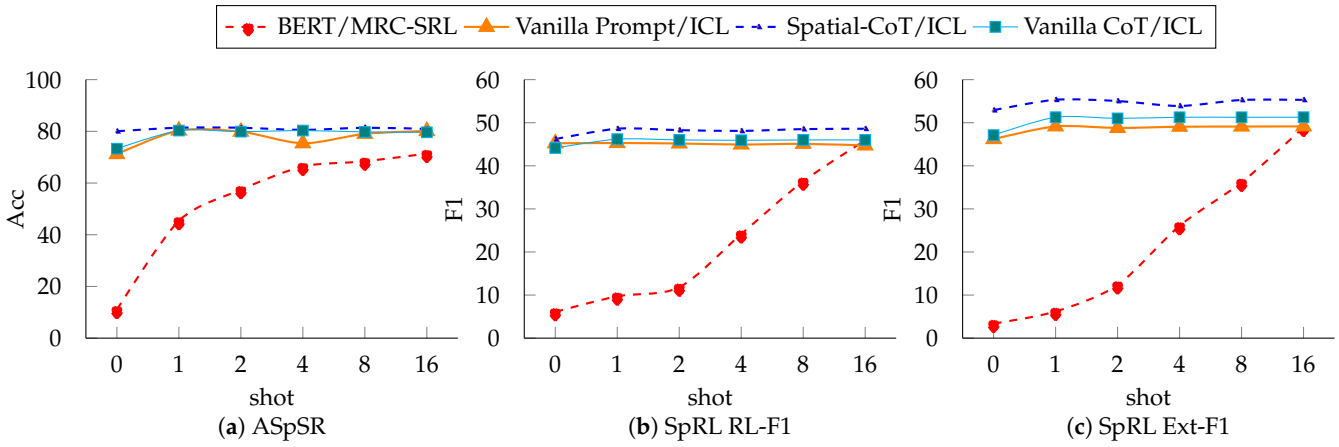

Figure 3. Few-shot performance on the SApSR and SpRL tasks. The x-axis denotes the number of samples. The y-axis denotes the task performance.

4.3.2. Step Ablation

To illustrate the necessity of each reasoning step, a step ablation study was conducted. We display the quantified contributions of each reasoning step of Spatial-CoT in Table 5. The results reveal that the entity analyzer, context analyzer, and common knowledge analyzer steps all contributed to the tasks. The entity analyzer contributed the most because it provided the most important entity clues for spatial reasoning. Additionally, we found that the CA and CKA contributed to the reasoning, but their order did not affect the final performance much.

Moreover, the order of each step also matters. We experimented on a sampled subset of SpaCE and explored the performance when using different chain orders. We found that the *EA-CA-CKA* reasoning sequence performed better than any other chain order.

Table 5. Ablation study of the reasoning steps.

EA	CA	CKA	ASpSR Acc	SpRL RL-F1	SpRL Ext-F1
✓	✓	✓	81.41	48.61	55.33
✓	✓	✗	81.16	47.42	54.13
✓	✗	✓	81.22	47.29	53.70
✓	✗	✗	80.81	46.80	52.40
✗	✗	✗	80.30	45.29	49.15

4.3.3. Prompt Analysis

We also explored the performance with different prompt templates. We tested some prompts that contained different keywords for instructions. The results are shown in Figure 4. The label "w/o Role Instructions" denotes that the prompt does not contain role instructions such as "You are acting as an entity analyzer", and the label "w/o Step Instructions" denotes that the prompt does not contain instructions such as "based on the previous results…". We found that it is useful to use suitable instructions in the CoT prompt, which can guide the LLM with a presupposed role and force it to reason with clues generated from reasoning chains. These instructions could improve each task by at least 1–2 points based on our investigation in Figure 4. We also explored whether different expressions of the same instruction would affect the final results; we tried many phrases, such as *"You are trying to extract entities from…"* and *"You are now an entity analyzer."*, to replace the instruction *"You are acting as an entity analyzer"*. We found that the expressions of these instructions hardly affected the results, while their existence mattered.

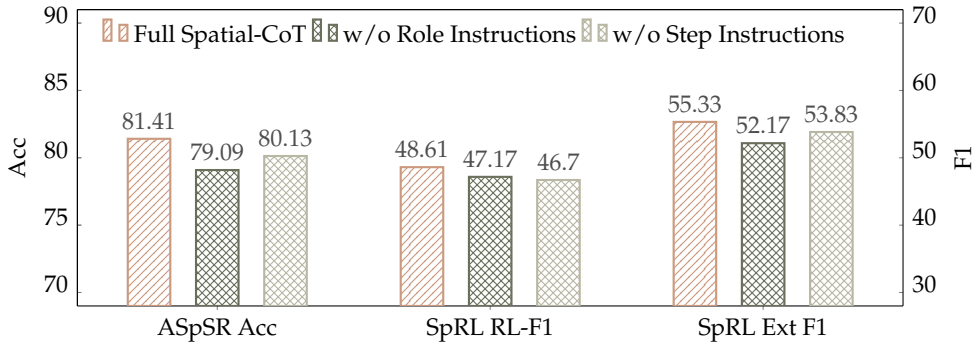

Figure 4. Comparison among the prompt templates that we explored.

4.3.4. Case Study

We qualitatively assessed the effectiveness of our CoT methods for the Chinese phenomenon of polysemia and the substitution of synonyms through some case studies. As shown in Figure 5, we chose some typical cases from the SpaCE dataset. We can see in the case "她不会匍匐前进，也不能快跑。她干脆直着身子，一摇一摆，慢慢地向方场上走去。一段还没有炸断的铁栏杆拦在她里面，她也不打算跨过去。她太衰老了，跨不过去，因此慢慢地绕过了那段铁栏杆，走进了方场。" ("She couldn't crawl forward, nor could she run quickly. She simply stood upright, swaying gently, and slowly walked toward the open field. A segment of iron railing, which had not yet been blown apart, blocked her in. She had no intention of crossing it either. She was too old to do so. Therefore, she slowly went around that section of iron railing and entered the open field."), there is an abnormal expression, "一段还没有炸断的铁栏杆拦在她里面", where "里面" (in her) should be "前面" (on her way). The vanilla prompt could not recognize this mistake, as shown in the top-left panel of Figure 5. Spatial CoT still failed to recognize the mistake based only on the

entity and context clues. When prompted with "the common sense" and "language usage" instructions, the LLMs finally refined the decision and gave the abnormal words.

Figure 5. The cases of the vanilla prompt and Spatial-CoT method for ASpSR (**left**) and SpRL (**right**).

For the case "乒乓球先是旋转了几下，接着从台面滚落到地上" and "乒乓球先是旋转了几下，接着从台面滚落到地下", the two sentences have the same meaning: "The ping pong ball first spun a few times, and then rolled off the table onto the ground". Based on Chinese language usage, the only difference in the two sentences, "落到地上" and "落到地下", is actually an expression of the same action—"fall to the ground". However, with the vanilla prompt, the LLM output that the two descriptions did not describe the same spatial scene, as shown in top-right panel of Figure 5. When feeding the instruction "based on common sense and language usage", our Spatial-CoT successfully guided the LLM to give the right decision.

5. Conclusions and Future Work

In this study, we investigated Chinese spatial language understanding with LLMs. We proposed the Spatial-CoT framework, which divides the reasoning process into three steps, namely, entity extraction, context analysis, and common knowledge analysis, and we designed a corresponding prompt template for Chinese SLU. First was a step for entity extraction, which provided base-entity-level clues for spatial reasoning. Next, the context analyzer employed the entity clues to make a preliminary prediction based on the input context. Finally, common knowledge was checked, and the output was refined based on common sense and Chinese language usage. With the proposed Spatial-CoT framework, an LLM can reason from a holistic perspective, thus effectively addressing the phenomena of polysemy and synonym substitution in Chinese spatial language understanding. On the SpaCE dataset, our strategy demonstrated superiority over vanilla prompt-learning methods, and it had competitive performance with respect to fully trained supervised models.

Finally, we discuss the limitations. This work has the following major limitations. First, our method needs a pre-trained large language model, which is expensive and time-consuming in the pre-training stage. Accessing the API-based LLM, such as ChatGPT4.0, also comes at a considerable cost. We will also continue studying how to use an open source large language model to solve spatial language understanding problems. Second, the generalization of our method remains to be tested. More experiments on other spatial understanding tasks should be verified. We will leave these studies in the future. Furthermore, we also plan to explore the following directions. For an in-depth exploration of the spatial language understanding in other languages, more high-quality multilingual data should be built. In another direction, we could extend this task to multi-modal settings, which could be applied to many real-life scenarios, such as vision and language navigation, remote manipulation, and robotics.

Author Contributions: Methodology, Y.Z.; Writing—original draft, Y.Z.; Supervision, J.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

Abbreviation and corresponding full names:

Abbreviation	Full Name	Abbreviation	Full Name
SLU	Spatial Language Understanding	LLM	Large Language Model
SpaCE	Spatial Cognition Evaluation	ICL	In-Context Learning
ASpSR	Abnormal Spatial Semantics Recognition	CoT	Chain-of-Thought
SpRL	Spatial Role Labeling	EA	Entity Analyzer
SpSM	Spatial Scene Matching	CA	Context Analyzer
CKA	Common Knowledge Analyzer	MRC	Machine Reading Comprehension

Appendix A. Extensive Results

Table A1. Extensive results of three tasks on seven SpaCE datasets. BERT/MRC-SRL lines mean that we use BERT classifier for ASpSR task and use MRC-SRL for SpRL task. We report the standard deviation in the brackets. **Bold:** best result.

	ASpSR		SpRL				SpSM	
	Recall	Acc	RL-Recall	RL-F1	Ext-Recall	Ext-F1	Recall	F1
• *Fine-tuning baselines*								
BERT/MRC-SRL (1-shot) [42]	49.29 ± 3.25 (41.04)	50.33 ± 5.07 (46.85)	7.90 ± 2.13 (51.44)	9.72 ± 2.17 (54.27)	6.02 ± 1.97 (54.29)	6.19 ± 2.03 (49.82)	-	-
BERT/MRC-SRL (Full) [42]	76.91 ± 1.10 (52.73)	78.26 ± 1.89 (58.99)	46.93 ± 0.10 (39.71)	46.49 ± 0.13 (39.94)	47.92 ± 0.62 (37.11)	49.10 ± 0.81 (36.04)	-	-
• *Prompt-based methods*								
Vanilla Prompt (zero-shot)	71.04 ± 0.81 (39.74)	71.19 ± 0.07 (37.05)	42.79 ± 0.41 (36.04)	43.01 ± 0.36 (44.54)	46.58 ± 0.40 (49.17)	46.20 ± 0.25 (46.85)	49.10 ± 0.81 (36.04)	35.10 ± 0.21 (39.32)
Vanilla CoT (zero-shot)	72.97 ± 0.66 (45.15)	73.29 ± 0.02 (38.10)	44.72 ± 0.19 (45.90)	44.13 ± 0.48 (46.31)	46.42 ± 0.31 (41.04)	47.12 ± 0.25 (39.94)	36.13 ± 0.40 (51.01)	35.66 ± 0.32 (48.03)
ICL (1-shot)	76.58 ± 0.91 (45.84)	77.13 ± 0.48 (49.38)	45.61 ± 1.41 (55.74)	45.29 ± 1.02 (58.33)	48.60 ± 0.81 (54.74)	49.15 ± 2.21 (52.66)	40.03 ± 1.14 (54.10)	39.29 ± 1.86 (52.45)
Vanilla CoT-ICL (1-shot)	79.71 ± 0.65 (47.36)	80.30 ± 0.70 (46.08)	45.57 ± 0.81 (56.44)	46.04 ± 1.52 (59.95)	50.83 ± 2.01 (58.04)	51.28 ± 1.89 (58.63)	43.02 ± 0.31 (59.11)	43.40 ± 0.18 (57.62)
• *Ours*								
Spatial-CoT (zero-shot)	79.66 ± 0.01 (33.74)	80.03 ± 0.02 (35.40)	47.31 ± 0.04 (38.24)	46.19 ± 0.03 (39.72)	52.13 ± 0.10 (32.31)	52.91 ± 0.07 (28.79)	44.30 ± 0.01 (41.14)	45.12 ± 0.02 (40.02)
Spatial-CoT-ICL (1-shot)	**80.89 ± 0.08 (31.10)**	**81.41 ± 0.12 (28.71)**	**48.30 ± 0.01 (47.74)**	**48.61 ± 0.08 (48.10)**	**54.80 ± 0.11 (44.34)**	**55.33 ± 0.15 (43.22)**	**48.10 ± 0.01 (48.66)**	**47.76 ± 0.03 (49.36)**

Appendix B. Detailed Parameters

Table A2. Model hyperparameters.

Hyper-Param.	Value	Hyper-Param.	Value
Number of BERT layers	12	Batch size	256
Dimension of BERT embedding	768	Number of epochs	20
Learning rate	1×10^{-5}	Optimizer	AdamW
Warmup rate	0.05	ChatGPT temperature	0.2

References

1. Clements, D.H.; Battista, M.T. Geometry and spatial reasoning. In *Handbook of Research on Mathematics Teaching and Learning*; Macmillan Publishers: New York, NY, USA, 1992; Volume 420, p. 464.
2. Kordjamshidi, P.; van Otterlo, M.; Moens, M. Spatial role labeling: Towards extraction of spatial relations from natural language. *ACM Trans. Speech Lang. Process.* **2011**, *8*, 1–36. [CrossRef]
3. Fei, H.; Zhang, M.; Ji, D. Cross-Lingual Semantic Role Labeling with High-Quality Translated Training Corpus. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, 5–10 July 2020; Association for Computational Linguistics: Kerrville, TX, USA, 2020; pp. 7014–7026. [CrossRef]
4. Mirzaee, R.; Faghihi, H.R.; Ning, Q.; Kordjamshidi, P. SPARTQA: A Textual Question Answering Benchmark for Spatial Reasoning. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, 6–11 June 2021; Association for Computational Linguistics: Kerrville, TX, USA, 2021; pp. 4582–4598. [CrossRef]
5. Liu, F.; Emerson, G.; Collier, N. Visual Spatial Reasoning. *arXiv* **2022**, arXiv:2205.00363.
6. Francis, J.; Kitamura, N.; Labelle, F.; Lu, X.; Navarro, I.; Oh, J. Core Challenges in Embodied Vision-Language Planning. *J. Artif. Intell. Res.* **2022**, *74*, 459–515. [CrossRef]
7. Mogadala, A.; Kalimuthu, M.; Klakow, D. Trends in Integration of Vision and Language Research: A Survey of Tasks, Datasets, and Methods. *J. Artif. Intell. Res.* **2021**, *71*, 1183–1317. [CrossRef]
8. Wang, C.; Luo, S.; Pei, J.; Liu, X.; Huang, Y.; Zhang, Y.; Yang, J. An Entropy-Awareness Meta-Learning Method for SAR Open-Set ATR. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 4005105. [CrossRef]
9. Hong, Y.; Opazo, C.R.; Wu, Q.; Gould, S. Sub-Instruction Aware Vision-and-Language Navigation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, 16–20 November 2020; Association for Computational Linguistics: Kerrville, TX, USA, 2020; pp. 3360–3376. [CrossRef]
10. Wang, S.; Qiu, Z.; Huang, P.; Yu, X.; Yang, J.; Guo, L. A Bioinspired Navigation System for Multirotor UAV by Integrating Polarization Compass/Magnetometer/INS/GNSS. *IEEE Trans. Ind. Electron.* **2023**, *70*, 8526–8536. [CrossRef]
11. Wang, C.; Pei, J.; Luo, S.; Huo, W.; Huang, Y.; Zhang, Y.; Yang, J. SAR Ship Target Recognition via Multiscale Feature Attention and Adaptive-Weighed Classifier. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 4003905. [CrossRef]
12. Bretas, A.M.C.; Mendes, A.; Jackson, M.; Clement, R.; Sanhueza, C.; Chalup, S.K. A decentralised multi-agent system for rail freight traffic management. *Ann. Oper. Res.* **2023**, *320*, 631–661. [CrossRef]
13. Wang, C.; Pei, J.; Yang, J.; Liu, X.; Huang, Y.; Mao, D. Recognition in Label and Discrimination in Feature: A Hierarchically Designed Lightweight Method for Limited Data in SAR ATR. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5239613. [CrossRef]
14. Zhan, W.; Sun, C.; Yue, P.; Tang, Q.; Qin, Z. SpaCE2021 dataset construction. *Appl. Linguist.* **2022**, *2*, 99–110.
15. Pustejovsky, J.; Kordjamshidi, P.; Moens, M.; Levine, A.; Dworman, S.; Yocum, Z. SemEval-2015 Task 8: SpaceEval. In Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2015, Denver, CO, USA, 4–5 June 2015; The Association for Computer Linguistics: Kerrville, TX, USA, 2015; pp. 884–894. [CrossRef]
16. Nichols, E.; Botros, F. SpRL-CWW: Spatial Relation Classification with Independent Multi-class Models. In Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2015, Denver, CO, USA, 4–5 June 2015; The Association for Computer Linguistics: Kerrville, TX, USA, 2015; pp. 895–901. [CrossRef]
17. D'Souza, J.; Ng, V. UTD: Ensemble-Based Spatial Relation Extraction. In Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2015, Denver, CO, USA, 4–5 June 2015; The Association for Computer Linguistics: Kerrville, TX, USA, 2015; pp. 862–869. [CrossRef]
18. Salaberri, H.; Arregi, O.; Zapirain, B. IXAGroupEHUSpaceEval: (X-Space) A WordNet-based approach towards the Automatic Recognition of Spatial Information following the ISO-Space Annotation Scheme. In Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2015, Denver, CO, USA, 4–5 June 2015; The Association for Computer Linguistics: Kerrville, TX, USA, 2015; pp. 856–861. [CrossRef]
19. Kordjamshidi, P.; Rahgooy, T.; Moens, M.; Pustejovsky, J.; Manzoor, U.; Roberts, K. CLEF 2017: Multimodal Spatial Role Labeling Task Working Notes. In Proceedings of the Working Notes of CLEF 2017—Conference and Labs of the Evaluation Forum, Dublin, Ireland, 11–14 September 2017; Volume 1866.

20. Zhao, Y.; Wei, J.; Lin, Z.; Sun, Y.; Zhang, M.; Zhang, M. Visual Spatial Description: Controlled Spatial-Oriented Image-to-Text Generation. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, 7–11 December 2022; Association for Computational Linguistics: Kerrville, TX, USA, 2022; pp. 1437–1449. [CrossRef]
21. Zhao, Y.; Fei, H.; Ji, W.; Wei, J.; Zhang, M.; Zhang, M.; Chua, T. Generating Visual Spatial Description via Holistic 3D Scene Understanding. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, ON, Canada, 9–14 July 2023; Association for Computational Linguistics: Kerrville, TX, USA, 2023; pp. 7960–7977. [CrossRef]
22. Kordjamshidi, P.; van Otterlo, M.; Moens, M. Spatial Role Labeling: Task Definition and Annotation Scheme. In Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, Valletta, Malta, 17–23 May 2010; European Language Resources Association: Paris, France, 2010.
23. Kordjamshidi, P.; Moens, M. Global machine learning for spatial ontology population. *J. Web Semant.* **2015**, *30*, 3–21. [CrossRef]
24. OpenAI. GPT-4 Technical Report. *arXiv* **2023**, arXiv:2303.08774.
25. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. LLaMA: Open and Efficient Foundation Language Models. *arXiv* **2023**, arXiv:2302.13971.
26. Zheng, L.; Chiang, W.L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.P.; et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *arXiv* **2023**, arXiv:2306.05685.
27. Wan, Z.; Cheng, F.; Mao, Z.; Liu, Q.; Song, H.; Li, J.; Kurohashi, S. GPT-RE: In-context Learning for Relation Extraction using Large Language Models. *arXiv* **2023**, arXiv:2305.02105.
28. Shome, D.; Yadav, K. EXnet: Efficient In-context Learning for Data-less Text classification. *arXiv* **2023**, arXiv:2305.14622.
29. Fei, Y.; Hou, Y.; Chen, Z.; Bosselut, A. Mitigating Label Biases for In-context Learning. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, ON, Canada, 9–14 July 2023; Association for Computational Linguistics: Kerrville, TX, USA, 2023; pp. 14014–14031.
30. Min, S.; Lyu, X.; Holtzman, A.; Artetxe, M.; Lewis, M.; Hajishirzi, H.; Zettlemoyer, L. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, 7–11 December 2022; Association for Computational Linguistics: Kerrville, TX, USA, 2022; pp. 11048–11064.
31. Liu, J.; Shen, D.; Zhang, Y.; Dolan, B.; Carin, L.; Chen, W. What Makes Good In-Context Examples for GPT-3? In Proceedings of the Proceedings of Deep Learning Inside Out: The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, DeeLIO@ACL 2022, Dublin, Ireland and Online, 27 May 2022; Association for Computational Linguistics: Kerrville, TX, USA, 2022; pp. 100–114. [CrossRef]
32. Wu, S.; Fei, H.; Qu, L.; Ji, W.; Chua, T. NExT-GPT: Any-to-Any Multimodal LLM. *arXiv* **2023**, arXiv:2309.05519.
33. Fei, H.; Li, B.; Liu, Q.; Bing, L.; Li, F.; Chua, T. Reasoning Implicit Sentiment with Chain-of-Thought Prompting. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2023, Toronto, ON, Canada, 9–14 July 2023; Association for Computational Linguistics: Kerrville, TX, USA, 2023; pp. 1171–1182. [CrossRef]
34. Inaba, T.; Kiyomaru, H.; Cheng, F.; Kurohashi, S. MultiTool-CoT: GPT-3 Can Use Multiple External Tools with Chain of Thought Prompting. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2023, Toronto, ON, Canada, 9–14 July 2023; Association for Computational Linguistics: Kerrville, TX, USA, 2023; pp. 1522–1532. [CrossRef]
35. Jin, Z.; Lu, W. Tab-CoT: Zero-shot Tabular Chain of Thought. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2023, Toronto, ON, Canada, 9–14 July 2023; Association for Computational Linguistics: Kerrville, TX, USA, 2023; pp. 10259–10277. [CrossRef]
36. Wu, D.; Zhang, J.; Huang, X. Chain of Thought Prompting Elicits Knowledge Augmentation. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2023, Toronto, ON, Canada, 9–14 July 2023; Association for Computational Linguistics: Kerrville, TX, USA, 2023; pp. 6519–6534. [CrossRef]
37. Kojima, T.; Gu, S.S.; Reid, M.; Matsuo, Y.; Iwasawa, Y. Large Language Models are Zero-Shot Reasoners. In Proceedings of the NeurIPS, New Orleans, LA, USA, 28 November–9 December 2022.
38. Nye, M.I.; Andreassen, A.J.; Gur-Ari, G.; Michalewski, H.; Austin, J.; Bieber, D.; Dohan, D.; Lewkowycz, A.; Bosma, M.; Luan, D.; et al. Show Your Work: Scratchpads for Intermediate Computation with Language Models. *arXiv* **2021**, arXiv:2112.00114.
39. Chung, H.W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, E.; Wang, X.; Dehghani, M.; Brahma, S.; et al. Scaling Instruction-Finetuned Language Models. *arXiv* **2022**, arXiv:2210.11416.
40. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.H.; Le, Q.V.; Zhou, D. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In Proceedings of the NeurIPS, New Orleans, LA, USA, 28 November–9 December 2022.
41. Zhou, D.; Schärli, N.; Hou, L.; Wei, J.; Scales, N.; Wang, X.; Schuurmans, D.; Cui, C.; Bousquet, O.; Le, Q.V.; et al. Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. In Proceedings of the The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, 1–5 May 2023.
42. Wang, N.; Li, J.; Meng, Y.; Sun, X.; Qiu, H.; Wang, Z.; Wang, G.; He, J. An MRC Framework for Semantic Role Labeling. In Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, 12–17 October 2022; International Committee on Computational Linguistics: New York, NY, USA, 2022; pp. 2188–2198.

43. Li, X.; Feng, J.; Meng, Y.; Han, Q.; Wu, F.; Li, J. A Unified MRC Framework for Named Entity Recognition. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, 5–10 July 2020; Jurafsky, D., Chai, J., Schluter, N., Tetreault, J.R., Eds.; Association for Computational Linguistics: Kerrville, TX, USA, 2020; pp. 5849–5859. [CrossRef]
44. Fei, H.; Zhang, Y.; Ren, Y.; Ji, D. Latent Emotion Memory for Multi-Label Emotion Classification. In Proceedings of the The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, 7–12 February 2020; AAAI Press: Washington, DC, USA, 2020; pp. 7692–7699. [CrossRef]
45. Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; Neubig, G. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Comput. Surv.* **2023**, *55*, 1–35. [CrossRef]