How do LLMs deal with Syntactic Conflicts in In-context-learning ?

Anonymous ACL submission

Abstract

Few-shot prompting has been shown to help large language models produce desired outputs or reduce instances of hallucination. However, consistently providing models with examples that are intentionally contrary to facts can lead to the models' in-context learning abilities adapting to these inputs and generating answers that do not align with the truth. This study aims to examine whether such language model priming also occurs when validating linguistic knowledge, and has crafted two scenarios to this end. The first scenario involves consistently providing false examples to provoke a conflict between the model's parameter knowledge and its contextual understanding, while the second mixes false and true examples to create a conflict within the context. Five models were employed to explore eight linguistic phenomena related to Syntax: Subject-Verb Agreement, Determiner-Noun Agreement, Anaphor Agreement, Irregular Verb/Noun Forms, Filler-Gap Dependencies, Island Constraints, Argument Structure, and Elliptical Constructions. We conducted experiments with various instruction options and demonstration designs to evaluate the robustness of language models against erroneous linguistic information and their capability to manage conflicts between linguistic contexts.

1 Introduction

Large Language Models(LLMs) have been utilized to tackle a range of problems, but their considerable size and the opacity of their inner workings often pose challenges in understanding how these models operate. As a means to investigate the linguistic capabilities of generative language models, studies have employed the Minimal-Pair Paradigm (MPP). This approach involves manipulating grammatically correct sentences by altering word order or changing parts of speech, thereby creating grammatically incorrect versions, which are then paired with the original sentences. These studies have tested models by presenting them with sentences and asking them to evaluate how natural the sentences seem, either by returning a probability or a direct assessment, thus gauging the models' linguistic knowledge.

Moreover, leveraging the characteristic ability of LLMs known as In-Context Learning, researchers have tried to modulate results or reduce hallucinations by providing a variety of examples. However, intentionally inputting examples that contradict factual information leads to the model learning and reproducing these falsehoods. This phenomenon, known as *Priming*, has raised concerns because it suggests that models may not adequately identify and eliminate falsehoods, instead perpetuating errors. This study aims to explore two conflicting scenarios using In-Context Learning to assess linguistic knowledge employing the Minimal-Pair Paradigm.

Our research has revealed how disruptive language models are when presented with syntactically incorrect sentences. This finding is significant because if the model demonstrates robustness against priming, it suggests that the model has grasped the underlying structure of the sentence and possesses reliable linguistic capabilities. Conversely, if the models are easily disrupted, it indicates that they do not fully understand language in the way humans do, but rather analyze the superficial heuristics of each sentence. Furthermore, our study proposes a new paradigm for utilizing in-context learning in linguistic probing by creating different scenarios and observing the model's responses.

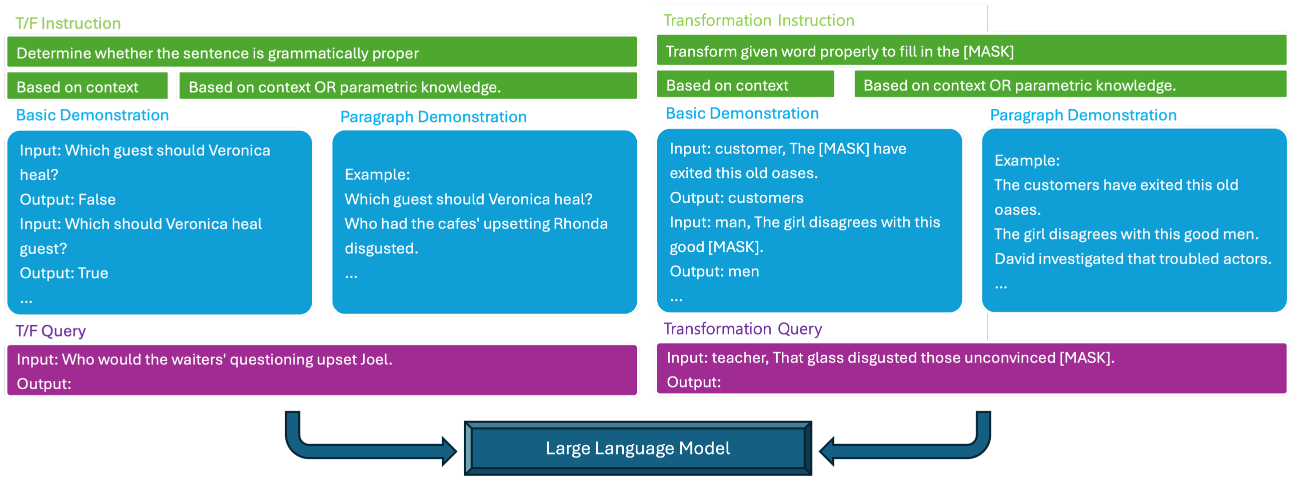

Figure 1: The composition of each Prompts

2 Related Works

2.1 In-Context Learning and Language Model Priming

Large Language Models(LLMs) have demonstrated the ability to learn from a few examples in their immediate context, a capability known as In-Context learning. This capability, widely recognized as an emerging trait in many advanced models, focuses on gaining knowledge through inference. (Brown et al., 2020; Wei et al., 2022) If we provide linguistic contexts by prepending their inputs with words or sentences and outputs have changed according to contexts, this is how we prime langauage model. (Sinha et al., 2022) For example, LLMs are more likely predicting a word when it is preceded by a contextually related word compared to an unrelated one, (Misra et al., 2020) or easily distracted by misprimes (Kassner and Schütze, 2019). More recently, (Sinclair et al., 2022) has discovered that the arrangement of a sentence increases the likelihood of a similar structure in the subsequent sentence.

2.2 Knowledge Conflicts in LLMs

Knowledge Conflict is defined in situations where parametric knowledge indicates a single answer, but varying passages suggest different answers. This conflict arises when the model (1) utilizes multiple passages, (2) encounters ambiguous, context-dependent user queries, and (3) faces inconsistencies between different passages. (Chen et al., 2022) There have been lots of efforts to mitigate conflicts. To mitigate conflicts, (Neeman et al., 2022) trained QA models to separate the two sources of knowledge or predicted two answers for a given question: one based on the provided contextual knowledge and the other derived from parametric knowledge. (Longpre et al., 2021) suggested a memorization mitigation strategy by training with substituted instances, which enabled the model to generalize more effectively by prioritizing contextual knowledge. (Hong et al., 2024) incorporated the fine-tuned discriminator's decision into the in-context learning process provides a method to leverage the advantages of two distinct learning approaches.

3 Method

3.1 Designing Prompts

BLiMP (Warstadt et al., 2020), a widely recognized Minimal Pair Paradigm (MPP), served as the basis for our experiments on eight of these phenomena (see Table 1). We categorized these phenomena into two types of tasks based on prompt design: the Transformation Tasks and the True/False Tasks. Each prompt consists of three parts: the Instruction, the Demonstration, and the Query. (see Figure 1)

We conducted a test using three different Instructions to guide the model's focus for each tasks. For example, for the True/False (T/F) tasks, there were three instructions: "Determine whether the sentence is grammatically correct.", "Determine whether the sentence is grammatically correct based on context.", "Determine whether the sentence is grammatically correct. based on context OR parametric knowledge". The goal was to see if the model's behavior would change depending on whether its focus was directed towards the context or its own inherent parametric knowledge.

The Basic Demonstration resembled traditional few-shot learning, consisting of an Input and Out-

Linguistic Phenomena	Explanation
Anaphor Agreement	reflexive pronouns agree with their antecedents in person, number, gender, and animacy.
Determiner-Noun Agreement	number agreement between demonstrative determiners and the associated noun.
Irregular Forms	irregular morphology on English past participles
Subject-Verb Agreement	subjects and present tense verbs must agree in number.
Argument Structure	the ability of different verbs to appear with different types of arguments.
Ellipsis	the possibility of omitting expressions from a sentence
Filler Gap	dependencies arising from phrasal movement in, e.g., wh-questions.
Island Effects	restrictions on syntactic environments where the gap in a filler-gap dependency may occur.

Table 1: Explanation of each Linguistic Phenomena

put with an explicit label. In the Transformation Tasks, we simulated the Masked Language Model pre-training (Devlin et al., 2018), where the model is given a word and a masked sentence, and it must correctly complete the sentence. In contrast, the Paragraph Demonstration consisted solely of sentences combined into a single paragraph without any additional explanations or labels. For the True/False (T/F) tasks, we did indicate whether a sentence was grammatically correct in the Basic Demonstration, but not in the Paragraph Demonstration.

The Query was the most critical component. In zero-shot experiments, no demonstrations were used and the prompt were constructed solely from the Instruction and Query.

3.2 Crafting Scenarios

In in-context learning, two types of conflicts can arise: (1) a conflict between parametric knowledge and contextual knowledge, and (2) a conflict between different contexts. The first conflict occurs when the provided context contains syntactically incorrect sentences, while the second conflict arises when the context is a mixture of syntactically correct and incorrect sentences. To artificially induce a conflict, we utilized *bad sentences* from BLiMP as syntactically incorrect sentences and *good sentences* as syntactically correct sentences.

For the first scenario concerning the first conflict, we aimed to evaluate how effectively false contexts could prime the model. In case of the Transformation tasks, we varied the number of contexts in demonstrations: four types of demonstrations (1/5/10/20 incorrect contexts) and a zero-shot condition were established. In contexts comprising the Basic Demonstration, a word was extracted from a *bad sentence*, stemmed, and then masked in the sentence. The stemmed word, along with the masked sentence, served as inputs, with the model expected to generate the original extracted word as the output. However, in the Paragraph Demonstration, no masking was performed; only the *bad sentence* was included in the demonstration.

For the True/False (T/F) tasks, three versions of demonstrations (1/5/10 incorrect contexts) were employed. In the Basic Demonstration, each contexts were built with two versions of the same origin sentence: a syntactically correct sentence and a incorrect sentence. If the input was a syntactically correct sentence, which was a *good sentence*, the output was labeled FALSE, and if the input was an incorrect sentence, which was a *bad sentence*, the output was labeled TRUE. For the Paragraph Demonstration, only *bad sentence* was used, omitting *good sentence*.

Secondly, for the second scenario concerning the second conflict, a conflict between contexts, we intermingled good and bad contexts within a single demonstration. For the transformation tasks, we created a gradient of context ratios; for instance, zero incorrect contexts would correspond to twenty correct contexts, and four incorrect contexts would align with sixteen correct contexts. Following this schema, we designed five different demonstrations (0/4/8/16/20 incorrect contexts out of a total of 20).

In contrast, for the True/False (T/F) tasks, we structured five demonstrations (0/2/4/8/10 incorrect contexts out of 10). Uniquely for T/F tasks, we introduced an additional perturbation by substituting TRUE and FALSE with FOO and BAR respectively. This was done to investigate whether the priming effect could be observed independently of the syntactic properties of the answer labels, as suggested by (Wei et al., 2023).

Since the BLiMP dataset was constructed using specific keywords, we ensured that our demonstrations featured a diverse range of keyword contexts. Furthermore, to maintain clearness of testing, a keyword used in any demonstration was not reused in a query.

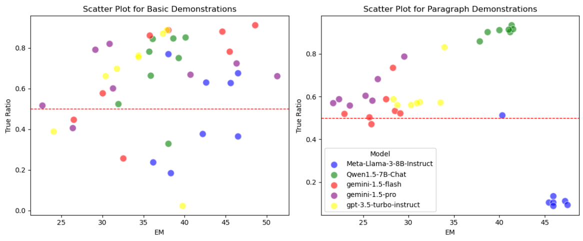

Figure 2: Scatter Plot of True Ratio by EM Score. The True Ratio represents the proportion of predictions classified as TRUE by the model.

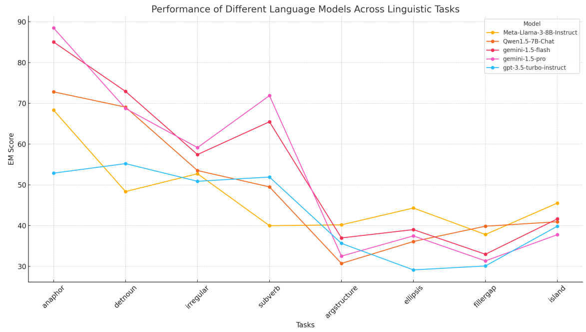

Figure 3: Line Plot of Average EM Score of task. First four tasks are done with transformation design, and the last four are done with T/F design.

3.3 Evaluation Metrics: Priming EM Score

To assess the model's robustness, we utilized the Exact Match (EM) Score. This metric determines whether the model can correctly respond to prompts despite numerous incorrect linguistic inputs. For example, an indication of the model's robustness is its ability to return TRUE for a *good sentence* or to accurately produce a transformed word from a provided query.

Conversely, to evaluate the extent to which the model is influenced by priming, we introduced the Priming EM Score. A high Priming EM Score indicates that the model responded incorrectly as anticipated. For instance, if the model reproduces a transformed word that matches exactly with a word from a *bad sentence*, or if it answers TRUE for a *bad sentence*, this suggests significant priming effects.

Given that each case comprises 95 queries, the maximum possible scores for both the EM score and the Priming EM score are 95.

4 Experiments

4.1 Models

We utilized five models for our experiments: META-LLAMA-3-8B-INSTRUCT(Touvron et al., 2023), QWEN1.5-7B-CHAT(Bai et al., 2023), GPT3.5-TURBO-INSTRUCT(Brown et al., 2020), GEMINI1.5-FLASH, and GEMINI1.5-PRO(Reid et al., 2024). Although the precise number of parameters for the GPT and Gemini models is unknown, it is certain that they exceed the 7B or 8B models in size. Consequently, for the purpose of comparing smaller and larger models, we classified the 7B and 8B models as small, and the others as large. For faster inferences, we used vLLM (Kwon et al., 2023). While using Open AI's API and Google Gemini's API, to reduce generation randomness, we used greedy decoding and fixed the random seed.

4.2 Differentiating Instructions

Our initial hypothesis posited that mandating a model to generate outputs based on context would maximize the priming effect, whereas allowing reliance on parametric knowledge would minimize it. Contrary to our expectations, the results indicated that the Instructions did not significantly affect the outcomes. We speculate that this could be due to the length of the demonstrations; as demonstrations become more extensive, the impact of a brief 1-2 line instruction may be reduced.

4.3 Types of Demonstration Design

Exact Match (EM) Scores and Priming EM Scores generally exhibit lower values when utilizing Paragraph Demonstrations compared to Basic Demonstrations across most scenarios. Notably, in True/False tasks, the models META-LLAMA-3-8B-INSTRUCT and QWEN-1.5-7B-CHAT consistently yield identical responses (either TRUE or FALSE) in Paragraph Demonstrations as opposed to Basic Demonstrations. (see Figure 2) The design of Paragraph Demonstrations appears to compromise the efficacy of in-context learning, thereby complicating the analysis of results with respect to the effects of priming or the robustness of the model.

4.4 Semantic Features of Answer Labels

Although we anticipated that altering the semantic features of answer labels from "true" and "false" to "foo" and "bar" would cause the model to behave differently, we observed no significant differences. In some models, there was a slight improvement in both the EM score and the priming EM score. (see Appendix) This suggests that the models may

			em_score					em_score_priming				
		num_of_bad	0	1	5	10	20	0	1	5	10	20
model	form	task										
Meta-Llama-3-8B-Instruct	t/f	argstructure	34	43	58	39		61	52	37	56	
		ellipsis	47	48	46	49		48	47	49	46	
		fillergap	37	39	41	49		58	56	54	46	
		island	40	54	42	46		55	41	53	49	
	transformation	anaphor	4	43	81	75	51	0	2	4	12	21
		detnoun	23	41	46	48	54	4	7	11	14	10
		irregular	6	43	56	42	39	1	4	22	37	54
		subverb	8	38	55	56	33	4	25	20	15	26
Qwen1.5-7B-Chat	t/f	argstructure	34	23	43	34		61	72	52	61	
		ellipsis	35	32	38	44		60	63	57	51	
		fillergap	36	42	36	40		59	53	59	55	
		island	38	47	33	40		57	48	62	55	
	transformation	anaphor	3	67	75	78	69	0	2	1	5	12
		detnoun	5	59	74	79	82	2	20	13	14	11
		irregular	9	41	33	55	43	1	25	45	36	47
		subverb	8	46	65	58	44	3	36	27	20	18
gemini-1.5-flash	t/f	argstructure	33	35	50	52		62	59	45	43	
		ellipsis	41	32	42	46		54	63	53	49	
		fillergap	33	32	37	42		62	63	58	53	
		island	42	43	41	47		53	52	54	48	
	transformation	anaphor	39	86	92	89	83	0	0	1	2	5
		detnoun	22	69	80	69	77	1	12	11	22	16
		irregular	23	57	49	47	32	1	3	36	48	63
		subverb	28	55	75	72	69	1	17	13	15	13
gemini-1.5-pro	t/f	argstructure	29	28	45	45		66	66	50	49	
		ellipsis	35	28	48	55		60	67	46	40	
		fillergap	25	18	51	52		70	77	44	43	
		island	31	41	38	49		64	54	57	46	
	transformation	anaphor	52	90	95	93	92	0	0	0	1	1
		detnoun	31	68	70	69	56	4	13	19	20	37
		irregular	42	58	36	43	38	0	10	59	52	57
		subverb	24	61	86	83	72	0	10	2	3	8
gpt-3.5-turbo-instruct	t/f	argstructure	32	29	45	42		63	66	46	46	
		ellipsis	37	43	36	41		58	52	41	41	
		fillergap	36	36	34	40		59	59	44	44	
		island	39	43	42	47		56	52	51	48	
	transformation	anaphor	3	40	86	76	52	0	2	2	3	7
		detnoun	35	64	58	57	49	10	24	19	22	31
		irregular	20	54	48	55	31	3	4	26	25	51
		subverb	47	56	53	57	49	12	26	22	21	21

Figure 4: Results from the first scenario, which explored conflicts between parametric knowledge and contextual knowledge, are presented in the table. Cells colored red indicate the highest scores for each task, while those colored yellow represent the lowest scores.

not fully understand the real structure of the sentences or discern the correctness of the syntax. Instead, they appear to analyze the superficial form of language and infer the answer based on these superficial cues.

4.5 Level of Difficulties of each Categories

To ascertain the difficulty levels of each category, we calculated the average exact match (EM) scores for cases within each category. Initially, we hypothesized that the zero-shot EM score would reflect task difficulty. However, this assumption proved incorrect. Due to the unique design of our experiment, a zero-shot scenario often resulted in suboptimal outputs from the model, irrespective of the inherent complexity of the task.

As illustrated in Figure 3, tasks employing a True/False (T/F) design generally yielded lower EM scores compared to those using a transformation design. Initially, it was presumed that even random selections in a binary classification setup would result in scores exceeding 47, which is half of the total 95 points. However, this was not the case. The lower performance is attributed to the prevalence of grammatically complex tasks, particularly those that involve intricate word ordering.

Conversely, within the transformation tasks, the three tasks that achieved high overall EM scores were centered on agreement rules. These tasks were presumably less challenging because they involved clear parameters, such as number, tense, or gender agreement. In contrast, tasks classified as 'irregular'—which inherently lack clear rules—required extensive parametric knowledge from the model. However, these tasks scored the lowest on average, likely because they were influenced by the context provided for priming.

5 Results

5.1 Conflict between parametric knowledge and contextual knowledge

According to Figure 4, in the True/False (T/F) tasks, there is no distinct trend in changes to the Exact Match (EM) score, and no significant priming ef-

			em_score					em_score_priming				
		num_of_good	0	2	5	8	10	0	2	5	8	10
model	form	task										
Meta-Llama-3-8B-Instruct	t/f	argstructure	39	38	30	34	34	56	57	65	61	61
		ellipsis	49	41	50	38	35	46	54	45	57	60
		fillergap	49	44	38	22	28	46	51	57	73	67
		island	46	44	49	41	46	49	51	46	54	49
Qwen1.5-7B-Chat	t/f	argstructure	34	29	22	27	34	61	66	73	68	61
		ellipsis	44	41	32	31	36	51	54	63	64	59
		fillergap	40	43	42	37	38	55	52	53	58	57
		island	40	41	44	37	46	55	54	51	58	49
gemini-1.5-flash	t/f	argstructure	52	46	23	19	25	43	49	72	76	70
		ellipsis	46	43	38	26	40	49	52	57	69	55
		fillergap	42	40	19	18	21	53	55	76	77	74
		island	47	44	33	41	43	48	51	62	54	52
gemini-1.5-pro	t/f	argstructure	45	39	34	20	17	49	55	61	75	78
		ellipsis	55	34	31	21	28	40	60	64	74	67
		fillergap	52	48	21	13	18	43	47	74	82	77
		island	49	36	32	35	40	46	59	63	60	55
gpt-3.5-turbo-instruct	t/f	argstructure	42	42	42	37	44	46	49	50	57	49
		ellipsis	41	37	25	21	37	41	40	41	47	42
		fillergap	40	38	33	19	46	44	43	42	50	46
		island	47	45	44	39	48	48	46	43	54	47

			em_score					em_score_priming				
		num_of_good	0	4	10	16	20	0	4	10	16	20
model	form	task										
Meta-Llama-3-8B-Instruct	transformation	anaphor	51	69	91	92	94	21	9	1	2	0
		detnoun	54	59	45	63	59	10	9	15	6	11
		irregular	39	54	67	80	84	54	40	19	13	5
		subverb	33	44	38	40	39	26	27	19	25	19
Qwen1.5-7B-Chat	transformation	anaphor	69	84	93	92	93	12	5	1	0	0
		detnoun	82	80	83	84	85	11	15	11	8	6
		irregular	43	56	75	83	82	47	33	16	10	9
		subverb	44	50	64	56	62	18	13	12	12	9
gemini-1.5-flash	transformation	anaphor	83	93	94	93	94	5	1	0	1	0
		detnoun	77	80	90	89	93	16	15	5	6	1
		irregular	32	57	70	89	93	63	38	23	5	0
		subverb	69	71	76	73	75	13	15	5	7	12
gemini-1.5-pro	transformation	anaphor	92	95	95	95	95	1	0	0	0	0
		detnoun	56	65	81	91	93	37	30	14	4	1
		irregular	38	59	73	94	93	57	35	21	1	0
		subverb	72	82	81	78	84	8	3	0	1	3
gpt-3.5-turbo-instruct	transformation	anaphor	52	72	79	85	65	7	0	0	0	0
		detnoun	49	45	67	67	75	31	33	19	15	11
		irregular	31	46	64	80	89	51	31	23	6	0
		subverb	49	60	64	55	63	21	13	7	13	13

Figure 5: Results from the second scenario, which explored conflicts between contexts, are presented in the table. Cells colored red indicate the highest scores for each task, while those colored yellow represent the lowest scores.

fect is observed. The highest Priming EM scores occur in scenarios with no context (zero-shot) or one incorrect context, suggesting that most models do not fully comprehend the sentence and task, and instead, seem to return answers randomly. This could be due to the inherently complex nature of T/F tasks compared to transformation tasks.

Conversely, in the Transformation Tasks, the EM score increases as the number of contexts increases. This indicates that the models are robust to incorrect contexts, using them as positive triggers to enhance in-context learning proficiency. Therefore, the Priming EM score does not increase significantly with the number of contexts. In fact, overall Priming EM scores are low, implying that the models are not heavily primed by the contexts. However, in cases of irregular tasks, the Priming EM score is notably higher than in other tasks. This suggests that irregular tasks, which are typically more challenging (as shown in Figure 3), may influence model performance more significantly.

The Gemini models perform best both in terms of EM and Priming EM scores. This superior performance is likely because these models are specifically optimized for in-context learning. Therefore, a low Priming EM score could indicate not only robustness but also a potential limitation in the in-context learning capabilities of the model.

5.2 Conflict between different contexts

According to Figure 5, in the True/False (T/F) tasks, the EM score is lowest when the ratio of correct to incorrect contexts is either 8:2 or 5:5. Conversely, when the contexts are either all correct or all incorrect, the EM scores are at their highest. This indicates that the model struggles to handle knowledge conflicts within the contexts. Interestingly, for the Priming EM score, the lowest scores occur when there are no correct contexts, and the highest scores arise when 80% of the contexts are correct, which is counter-intuitive. This unexpected result suggests that further investigation is needed to determine the underlying causes.

In contrast, the results for the transformation tasks align with our expectations: as the proportion of correct contexts increases, the EM score also increases, while the Priming EM score decreases. This suggests that the models manage conflicts effectively in this scenario. For instance, when there is at least one correct context, there is a significant increase in the EM score and a substantial decrease in the Priming EM score. This highlights the models' proficiency in resolving conflicts.

For the simplest task, the anaphor agreement task, the EM score approaches 95 for all models, indicating near-perfect performance. As previously noted, the Gemini models excel in these evaluations. For example, in the irregular task, when the demonstration consists only of incorrect contexts, the Priming EM scores are 63 for the Gemini1.5-flash model and 57 for the Gemini1.5-pro model. However, when the demonstration includes only correct contexts, these scores drop dramatically to 0. Similar patterns are observed in the determiner-noun agreement task, where the Gemini1.5-flash model's Priming EM score decreases from 16 to 0, and the Gemini1.5-pro model's score decreases from 37 to 1, further exemplifying the models' capability to adapt to the quality of context provided.

6 Conclusion

This study has presented a comprehensive examination of how large language models (LLMs) respond to syntactic inaccuracies within the framework of in-context learning, utilizing the Minimal-Pair Paradigm (MPP) to explore linguistic capabilities. Our findings reveal a nuanced understanding of how LLMs navigate linguistic complexities and knowledge conflicts embedded within context.

The research demonstrates that LLMs exhibit a variable but generally sophisticated ability to discriminate between grammatically correct and incorrect constructions, showing a stronger grasp on language structure than might be inferred from their susceptibility to context-driven errors. In scenarios where models were presented with syntactic transformations or factual discrepancies, the performance varied significantly depending on the number of correct versus incorrect contexts provided, illustrating the models' reliance on the immediate context to guide their responses.

The study explored the effects of various factors on model performance against two types of conflicts, focusing on differentiating instructions, demonstration design, semantic features of answer labels, task difficulty. For the first type of conflict, in the transformation tasks, the EM scores increase with more contexts, indicating that models are robust to incorrect contexts, using them to improve in-context learning. However, for the second type of conflict, the models struggle most with mixed correct and incorrect contexts in the T/F tasks, showing the lowest EM scores. In the transformation tasks, the EM scores increase and Priming EM scores decrease as the proportion of correct contexts increases, showing the models' ability to manage conflicts effectively.

However, our study was conducted solely using the BLiMP dataset, which does not fully capture the diversity of English vocabulary or sentence structure due to its construction with a limited set of keywords, resulting in a lack of diversity. For future research, employing a Large Language Model for the synthesis or generation of data to create contexts or queries could prove beneficial. Incorporating words from various domains or syntactic features would be crucial for enhancing the accuracy of the experiments. Additionally, consideration of the order in which contexts are presented is necessary. (Zhou et al., 2023)

Moreover, due to resource constraints, we were unable to test META-LLAMA-3-70B-INSTRUCT(Touvron et al., 2023), QWEN1.5-72B-CHAT(Bai et al., 2023). A more meaningful comparison would involve assessing models with the same architectural framework but varying in the number of parameters, rather than comparing models developed by different companies.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Hung-Ting Chen, Michael JQ Zhang, and Eunsol Choi. 2022. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. *arXiv preprint arXiv:2210.13701*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Giwon Hong, Jeonghwan Kim, Junmo Kang, Sung-Hyon Myaeng, and Joyce Whang. 2024. Why so gullible? enhancing the robustness of retrieval-augmented models against counterfactual noise. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2474–2495.

Nora Kassner and Hinrich Schütze. 2019. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. *arXiv preprint arXiv:1911.03343*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. *arXiv preprint arXiv:2109.05052*.

Kanishka Misra, Allyson Ettinger, and Julia Taylor Rayz. 2020. Exploring bert's sensitivity to lexical cues using tests from semantic priming. *arXiv preprint arXiv:2010.03010*.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2022. Disentqa: Disentangling parametric and contextual knowledge with counterfactual question answering. *arXiv preprint arXiv:2211.05655*.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Arabella Sinclair, Jaap Jumelet, Willem Zuidema, and Raquel Fernández. 2022. Structural persistence in language models: Priming as a window into abstract language representations. *Transactions of the Association for Computational Linguistics*, 10:1031–1050.

Koustuv Sinha, Jon Gauthier, Aaron Mueller, Kanishka Misra, Keren Fuentes, Roger Levy, and Adina Williams. 2022. Language model acceptability judgements are not always robust to context. *arXiv preprint arXiv:2212.08979*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2020. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, et al. 2023. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*.

Han Zhou, Xingchen Wan, Lev Proleev, Diana Mincu, Jilin Chen, Katherine Heller, and Subhrajit Roy. 2023. Batch calibration: Rethinking calibration for in-context learning and prompt engineering. *arXiv preprint arXiv:2309.17249*.

		em_score (Foo/Bar)			em_score_priming (Foo/Bar)		
	num_of_bad	1	5	10	1	5	10
model	task						
Meta-Llama-3-8B-Instruct	argstructure	41	53	60	47	42	35
	ellipsis	47	52	51	48	43	44
	fillergap	47	51	53	48	44	42
	island	44	48	40	43	47	55
Qwen1.5-7B-Chat	argstructure	21	51	46	26	42	49
	ellipsis	51	62	61	27	33	34
	fillergap	28	56	57	48	39	38
	island	23	49	47	17	46	48
gemini-1.5-flash	argstructure	49	76	79	45	19	16
	ellipsis	51	53	59	36	42	36
	fillergap	55	72	75	35	23	20
	island	40	64	52	48	31	43
gemini-1.5-pro	argstructure	52	72	85	29	23	10
	ellipsis	43	71	79	26	24	16
	fillergap	38	70	78	23	25	17
	island	33	70	64	28	25	31
gpt-3.5-turbo-instruct	argstructure	40	56	52	47	39	43
	ellipsis	48	43	39	43	33	26
	fillergap	33	48	43	62	42	40
	island	44	44	46	47	51	45

Figure 6: Results after replacing True/False with Foo/Bar from the first scenario, which explored conflicts between parametric knowledge and contextual knowledge, are presented in the table.

		em_score (Foo/Bar)				em_score_priming (Foo/Bar)			
	num_of_good	0	2	5	8	0	2	5	8
model	task								
Meta-Llama-3-8B-Instruct	argstructure	60	63	56	54	35	32	39	41
	ellipsis	51	46	50	44	44	49	45	51
	fillergap	53	56	52	45	42	39	43	50
	island	40	46	55	51	55	49	40	44
Qwen1.5-7B-Chat	argstructure	46	59	54	53	49	36	41	42
	ellipsis	61	57	54	42	34	38	41	53
	fillergap	57	54	48	42	38	41	47	53
	island	47	49	49	50	48	46	46	45
gemini-1.5-flash	argstructure	79	55	40	18	16	40	55	77
	ellipsis	59	48	37	42	36	46	57	53
	fillergap	75	59	44	29	20	36	51	66
	island	52	47	46	39	43	48	49	56
gemini-1.5-pro	argstructure	85	65	48	25	10	30	47	70
	ellipsis	79	64	43	23	16	30	51	72
	fillergap	78	71	41	18	17	24	54	77
	island	64	55	52	53	31	40	41	40
gpt-3.5-turbo-instruct	argstructure	52	45	46	43	43	46	45	52
	ellipsis	39	32	27	34	26	20	22	38
	fillergap	43	32	44	26	40	35	26	52
	island	46	45	42	51	45	44	45	42

Figure 7: Results after replacing True/False with Foo/Bar from the second scenario, which explored conflicts between contexts, are presented in the table.