Explain-then-Process: Using Grammar Prompting to Enhance Grammatical Acceptability Judgments

Russell Scheinberg[1]   Ameeta Agrawal[1]   Amber Shore[1]   So Young Lee[2]

[1] Portland State University, USA
[2] Miami University, USA

{rschein2, ameeta, ashore}@pdx.edu   soyoung.lee@miamioh.edu

Abstract

Large language models (LLMs) can explain grammatical rules, yet they often fail to apply those rules when judging sentence acceptability. We present *grammar prompting*, an explain-then-process paradigm: a large LLM first produces a concise explanation of the relevant syntactic phenomenon, then that explanation is fed back as additional context to the *target model* – either an LLM or a smaller language model (SLM) – before deciding which sentence of a minimal pair is grammatical. On the English BLiMP, Chinese SLING, and Russian RuBLiMP benchmarks, this simple prompt design yields substantial improvements over strong baselines across a wide range of syntactic phenomena. Feeding an LLMs metalinguistic explanation back to the target model bridges the gap between *knowing* a rule and *using* it. On SLMs, grammar prompting alone trims the average LLM-SLM accuracy gap by 20%, and when paired with chain-of-thought, by 56% (13.0 pp → 5.8 pp), all at negligible cost. The lightweight, language-agnostic cue lets low-cost SLMs approach frontier-LLM performance in multilingual settings.

1   Introduction

Large Language Models (LLMs) occupy an uncanny valley: they appear *almost* human in their language use, excelling at tasks from text generation to translation to complex reasoning. Yet these same models show surprising gaps, particularly when asked to make explicit grammatical judgments - a limitation that becomes especially apparent in multilingual contexts (Alrajhi et al., 2022; Suijkerbuijk et al., 2025). While modern language models generate fluent text, they often struggle to reflect on grammatical structure, suggesting a disconnect between their ability to use language and their capacity to reason about it, a gap that is particularly pronounced in smaller models.

Consider this example. We asked Claude 3.5 Sonnet to reason out which minimal pair sentence (Warstadt et al., 2020) is more grammatically acceptable:

> (A) Only a popsicle that Danielle admires ever freezes. ✓
> (B) A popsicle that only Danielle admires ever freezes. ✗

Sonnet incorrectly selected '**B**'. Its reasoning has three steps: **i**, it paraphrases the meanings of the sentences as (A), "The only thing that ever freezes is a popsicle Danielle admires" and (B), "A popsicle freezes only if Danielle, and no one else, admires it"; **ii**, Sonnet analyzes the paraphrases, noting that both (A) and (B) have the same "only" + "ever" pattern; and **iii**, Sonnet concludes that "both are grammatical", and chooses (B) as 'clearer'. We speculate that the paraphrase hides the fact that "ever" is a negative-polarity item (NPI) which must be within the scope of "only." In (A), "only" properly licenses "ever", but in (B), it does not. By reducing both sentences to paraphrases, the model loses track of this licensing constraint and declares (B) acceptable.

We observed this pattern across different LLMs and linguistic phenomena: when making acceptability judgments, models typically begin by paraphrasing (or translating) the sentences, and in this process frequently obscure the grammatical features crucial for accurate judgments.

How can we help models focus more directly on linguistic structure? To explore this question, we adopt a classic psycholinguistics task: *grammatical acceptability judgments* (Schütze, 2016). Specifically, we evaluate whether models can reliably distinguish grammatically acceptable forms from grammatically unacceptable forms by using minimal pairs that differ by only one syntactic feature (e.g., a missing question particle). We focus
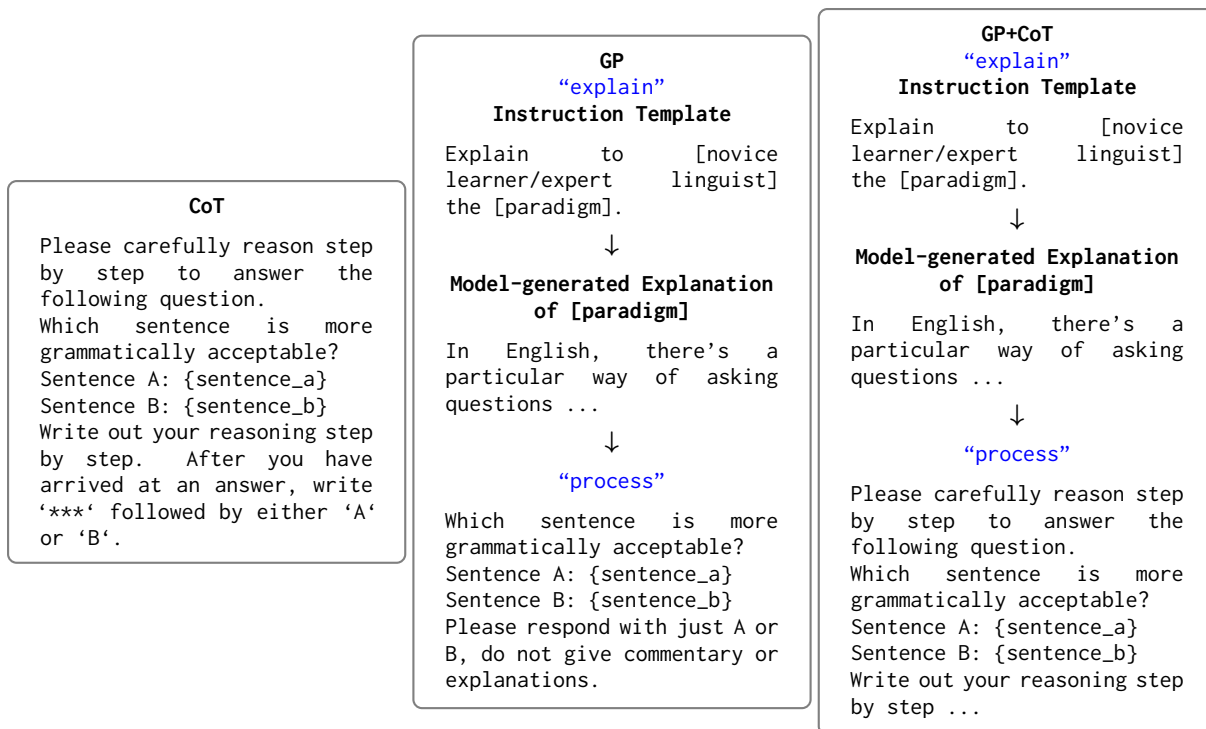
19778

CoT

```
Please carefully reason step
by step to answer the
following question.
Which sentence is more
grammatically acceptable?
Sentence A: {sentence_a}
Sentence B: {sentence_b}
Write out your reasoning step
by step.  After you have
arrived at an answer, write
'***' followed by either 'A'
or 'B'.
```

GP
"explain"
Instruction Template

```
Explain    to    [novice
learner/expert   linguist]
the [paradigm].
```
↓

**Model-generated Explanation
of [paradigm]**

```
In  English,    there's  a
particular  way  of  asking
questions ...
```
↓

"process"

```
Which   sentence   is   more
grammatically acceptable?
Sentence A: {sentence_a}
Sentence B: {sentence_b}
Please respond with just A or
B, do not give commentary or
explanations.
```

GP+CoT
"explain"
Instruction Template

```
Explain    to    [novice
learner/expert   linguist]
the [paradigm].
```
↓

**Model-generated Explanation
of [paradigm]**

```
In  English,    there's  a
particular  way  of  asking
questions ...
```
↓

"process"

```
Please carefully reason step
by step to answer the
following question.
Which sentence is more
grammatically acceptable?
Sentence A: {sentence_a}
Sentence B: {sentence_b}
Write out your reasoning step
by step ...
```

Figure 1: An illustration of Chain-of-Thought (CoT) reasoning vs. Grammar Prompting (GP) vs. Grammar Prompting with Chain-of-Thought reasoning (GP+CoT)

on three typologically distinct languages: English, Chinese, and Russian.

Following Mahowald et al. (2024), we distinguish between two kinds of linguistic competence: *formal* – the knowledge of structural rules needed to differentiate grammatical from ungrammatical forms – and *functional* – practical language use in tasks like translation or summarization. While language models often display strong functional competence, they may possess implicit syntactic knowledge but fail to access it systematically, particularly in grammatical acceptability judgments.

We introduce the **explain-then-process** paradigm for grammar prompting (see Figure 1), which has two steps: first, a grammatical explanation of a given paradigm is elicited from an LLM; then the grammatical explanation is supplied back to an LLM or a smaller language model (SLM), and we investigate the interplay between grammar prompting and chain-of-thought reasoning. By equipping LLMs and SLMs with these targeted prompts, we find that models more consistently identify syntactic violations, suggesting that explicit grammar guidance can unlock more systematic linguistic reasoning.

We make the following contributions:

- A novel explain-then-process paradigm that

elicits explanations from models for a given task before using them for processing.

- A grammar prompting method that leverages explicit metalinguistic input to further unlock a model's linguistic reasoning and boost accuracy in minimal-pair grammatical acceptability task, especially for SLMs.

- A systematic evaluation of multiple language models from the GPT, Claude, Llama families across English, Chinese, and Russian, analyzing how well grammar prompting improves their ability to make structured linguistic judgments.

2   Related Work

Recent theoretical work by Wies et al. (2024) suggests that LLMs possess considerable latent linguistic knowledge that can be activated through in-context learning (ICL) and introducing new linguistic information. For example, Lampinen et al. (2022) and Srivastava et al. (2023) demonstrate the effectiveness of targeted explanations and validation-tuned prompts to improve ability in specific tasks. Notably, the *Machine Translation from One Book* (MTOB) paradigm found that LLMs could improve translation between English and a zero-resource Papuan language, Kalamang,

by consulting one grammar book (Tanzer et al., 2024). Follow-up work showed that incorporating explicit rules can help in some cases (Guo et al., 2024; Zhang et al., 2024), but might be insufficient without additional strategies such as validation-tuned prompts (Lampinen et al., 2022; Srivastava et al., 2023).

Grammatical acceptability judgment tasks have long been a fundamental method in psycholinguistics for assessing linguistic competence, typically presenting single sentences for evaluation (Schütze, 2016). In computational linguistics, this task has been systematized using minimal pairs–sentence pairs that differ by only one syntactic feature–allowing for more controlled comparisons. This approach led to the development of benchmark datasets such as the English-language CoLA and BLiMP (Warstadt et al., 2019, 2020), which have inspired similar resources for other languages, including Chinese (Song et al., 2022), Russian (Taktasheva et al., 2024), Arabic (Alrajhi et al., 2022), Dutch (Suijkerbuijk et al., 2025), and Indonesian (Leong et al., 2023). While some computational approaches to minimal pair evaluation still rely on intrinsic metrics like perplexity or model-assigned probabilities, there is growing interest in prompt-based methods that explicitly ask models to make grammatical acceptability judgments (Qiu et al., 2024), aligning more closely with how LLMs are used in interactive settings.

3 Explain-then-Process Grammar Prompting

We hypothesize that while LLMs can often describe grammatical rules, they do not consistently apply these rules when making grammatical acceptability judgments. To address this, we propose an *explain-then-process* framework for grammar prompting allowing us to test whether LLMs can integrate explicit metalinguistic knowledge into their grammatical reasoning.

1. **explain** - the model first generates a detailed description of a grammatical phenomenon, outlining key syntactic constraints.
2. **process** - this explanation is then fed back to the model to guide syntactic decision-making and improve processing.

By explicitly linking grammatical descriptions to acceptability judgments, this approach tests whether LLMs can internalize syntactic rules.

Task: Explain, to a [novice learner/expert linguist], the 'left branch island echo question' in English.
Context: Help a [novice learner/expert linguist] understand and identify correct usage of this concept to pass a test.
Requirements:
- Clear, non-technical title
- Explain key rules and patterns
- Include specific grammatical details
- ...

Examples (for reference, do not include):
Good: Many boys will embarrass what patient?
Bad: What will many boys embarrass patient?
Good: Sara was insulting what student?
Bad: What was Sara insulting student?
Target length: 250 words

Figure 2: Instruction Template for Eliciting Grammar Explanations

This is particularly important in multilingual settings, where access to explicit grammatical explanations is limited for many languages. Unlike English, which benefits from extensive linguistic resources, many languages lack formalized grammatical descriptions. Explain-then-process prompting allows LLMs to self-generate grammatical explanations, reducing reliance on external resources.

To elicit principled and consistent grammatical explanations from LLMs, we first design an *instruction template* (Figure 2 shows a sample template) so as to ensures systematic coverage across languages and structures. Each template includes the grammatical paradigm name (e.g., NPI licensing), sample minimal pairs, and an instruction to avoid full example sentences, preventing reasoning based on examples (Aycock et al., 2024). The instruction template also specifies a target audience: we request either a description for *beginners* with clear, straightforward explanations, or for *experts* with technical descriptions and formal terminology, following MTOB (Tanzer et al., 2024).

Next, we use these instructions to ask LLMs to generate targeted grammatical explanations, or *grammar prompts*. Figure 3 shows a grammatical explanation generated by Sonnet. Given that the models we examine have English as their dominant training language (Wendler et al., 2024), we used English prompts throughout, while the mini-

```
Understanding Special Question Patterns in
English
In English, there's a particular way of
asking questions that might seem tricky at
first. This pattern involves asking about
a specific part of a sentence while keeping
most of the original sentence structure
intact.
Key Rules:
  1. ... [truncated for brevity]
Grammatical Details:
  • ... [truncated for brevity]
Tips for Correct Usage:
  • ... [truncated for brevity]
To Spot Incorrect Usage:
  • ... [truncated for brevity]
Remember, this question form is used
in specific contexts, often to express
surprise or seek confirmation. It's not
the standard way to form all questions in
English, but it's important to recognize
and use correctly when appropriate.
```

Figure 3: Sonnet-generated 'Beginner-Oriented' Grammar Explanation for '*left branch island echo question*'

mal pair sentences were in the target language.

4 Experimental Setup

4.1 Models

We evaluated five models, including top-tier proprietary models such as GPT-4o (`gpt-4o-2024-08-06`) and Claude 3.5 Sonnet (`claude-3-5-sonnet-20240620`) and their smaller counterparts such as GPT-3.5 (`gpt-3.5-turbo-0125`) and Claude 3.5 Haiku (`claude-3-haiku-20240307`), along with Llama 3.3 9B as an open-source multilingual model.

4.2 Datasets

We focus our evaluation on three languages. Each minimal pair dataset covers a specific syntactic phenomenon with minimally different sentence pairs – one grammatical and one ungrammatical. Each dataset organizes its minimal pair paradigms into broad syntactic categories (e.g., island effects, classifier agreement).

English The Benchmark of Linguistic Minimal Pairs (BLiMP) (Warstadt et al., 2020) spans 67 syntactic paradigms (e.g., subject-verb agreement, island constraints, negative polarity licensing) that were extracted from English syntax textbooks and cover diverse aspects of English grammar.

Chinese The Benchmark of Sino LINGuistics (SLING) (Song et al., 2022) comprises treebank-derived syntactic minimal pairs, covering 38

paradigms including phenomena like aspect marking with *le*, *guo* and *zai*, and wh-fronting.

Russian The Russian Benchmark of Linguistic Minimal Pairs (RuBLiMP) (Taktasheva et al., 2024) is made up of corpus-derived minimal pairs characterized by lexical and syntactic diversity. The dataset comprises 45 paradigms covering morphosyntactic phenomena like subject-predicate agreement and reflexive binding.

Focusing on "challenging" paradigms Initial testing showed gpt-4o achieving near-perfect accuracy on many paradigms. Appendix B contains the full set of results. We selected the most challenging paradigms across 8 categories from BLiMP (90% or lower accuracy), 6 from SLING (excluding easier relative clause and anaphora categories), and 7 categories from RuBLiMP ($\leq 96\%$ accuracy). Each category includes 1-5 minimal pair paradigms, and we use the first 50 minimal pairs per paradigm.

4.3 Presentation Schema

Intrinsic metrics such as probability or perplexity, often used in minimal pair datasets to probe LLMs' syntactic knowledge (Warstadt et al., 2019, 2020; Alrajhi et al., 2022; Song et al., 2022), frequently fail to capture how models behave when explicitly asked to judge grammaticality (Suijkerbuijk et al., 2025). Furthermore, access to intrinsic metrics is limited for proprietary models, making direct prompt-based evaluations of grammatical competence a more practical and usage-aligned alternative as LLMs are deployed in applications requiring explicit linguistic processing and feedback.

Following Leong et al. (2023), we present both sentences of a minimal pair simultaneously. This approach serves to focus the model's attention on the critical grammatical contrast and to avoid yes-response bias, where models show a clear preference toward positive answers (Dentella et al., 2023), which we observed in preliminary tests when presenting single sentences. Additionally, we systematically vary the presentation order of grammatical and ungrammatical sentences across three A/B prompting trials: **Trial 1**: The grammatical sentence is presented as "Sentence A" and the ungrammatical as "Sentence B". **Trial 2**: The order is reversed. **Trial 3**: The assignment is randomized. The results of these trials are averaged.

4.4 Baseline Conditions

We include several different types of conditions in our experiments.

- **Base** is the basic prompt as follows:

  Which sentence is more grammatically acceptable?

  Sentence A: {sentence_a}

  Sentence B: {sentence_b}

  Please respond with just A or B, do not give commentary or explanations.

- **Chain-of-Thought (CoT)** asks the model to reason step by step before answering (Fig. 1)

- **Control (con)** provides an explanation of an *irrelevant* linguistic phenomenon (the "null quotative") which does not correspond to any tested paradigms, instead of providing a grammar explanation relevant to the minimal pair. This control ensures that performance gains are due to targeted grammar explanations rather than generic instruction-following behavior.

- **Textbook (txt)** provides a compilation of multiple grammatical explanations instead of a single relevant grammar prompt, and models must select and apply the correct rule for each minimal pair.

- **3-shot** encourages superficial pattern matching instead of reasoning (Du et al., 2023), so we do not consider it a strictly comparable baseline; nevertheless, we present and discuss three-shot results in Section 5.6.

The proposed grammar prompts – beginner (GPb) and expert (GPx) – are tested with both base and CoT prompts. To generate the explanations, we initially leveraged the two strongest available models: Claude 3.5 Sonnet and GPT-4o, but we transitioned from gpt-4o to GPT-o1 when it was released. Small-scale manual verification confirmed the quality and adherence to the instructions.

5 Results and Analysis

We present the results (macro-average accuracy across paradigm categories) of four main experiments: (1) comparing beginner vs. expert-oriented grammar prompts (Section 5.1), (2) evaluating control and textbook conditions (Section 5.2), (3) assessing performance on BLiMP, SLING, and RuBLiMP (Sections 5.3–5.5), and (4) comparison

Categories	gpt-3.5		gpt-4o	
	GPb	GPx	GPb	GPx
ENGLISH (**BLIMP**)				
NPI lic.	69.7	**75.3**	87.7	**91.3**
arg. Str.	**79.3**	68.0	**90.7**	88.0
binding	**67.7**	63.3	**80.3**	67.0
cntrl/Rs.	**92.9**	85.3	**92.4**	91.1
ellipsis	72.7	**76.0**	69.7	58.3
filler gap	58.9	**78.2**	81.3	**91.1**
island ef.	**65.1**	57.5	**79.1**	72.0
quantifiers	83.0	**86.3**	100.0	90.3
average (BLiMP)	73.6	**73.7**	85.15	81.14
CHINESE (**SLING**)				
alt. quest.	**100.0**	**100.0**	**100.0**	**100.0**
aspect	78.8	76.5	**99.0**	97.0
classifier-Noun	**75.1**	72.1	**89.9**	88.0
definiteness	**98.7**	98.2	99.8	**100.0**
average (SLING)	**88.1**	86.7	**97.1**	96.2
RUSSIAN (**RUBLIMP**)				
anaphor	**88.7**	85.3	93.3	**94.0**
reflexives	**100.0**	**100.0**	**100.0**	**100.0**
subj.-pred. agr	**83.8**	83.3	96.7	**96.9**
aspect	**85.8**	75.8	**100.0**	**100.0**
government	78.0	**80.0**	**98.3**	96.3
word formation	**98.7**	93.3	**100.0**	**100.0**
average (RuBLiMP)	**89.1**	86.2	**98.0**	97.8
average over all	**82.0**	80.8	**92.1**	90.1

Table 1: Results of Beginner (GPb) and Expert (GPx) grammar prompting conditions; explanations generated using Claude Sonnet

Categories	gpt-3.5			gpt-4o		
	con	txt	GPb	con	txt	GPb
NPI lic.	64.0	59.3	**67.3**	74.3	81.0	**93.6**
arg. str.	**69.7**	65.3	46.7	84.3	**90.3**	89.6
binding	**57.5**	56.0	57.3	74.8	73.7	**78.0**
cntrl/rs.	75.4	75.6	**87.8**	81.6	85.8	**94.7**
ellipsis	69.5	58.2	**74.0**	69.7	69.5	**89.5**
filler gap	46.9	53.4	**88.9**	70.0	73.8	**95.1**
island ef.	63.0	**63.3**	63.1	71.6	75.4	**81.6**
quantifiers	66.7	59.0	**95.0**	80.2	73.0	**99.2**
average	64.1	61.3	**72.5**	75.8	77.8	**90.2**

Table 2: Control and Textbook condition comparison on BLiMP dataset

with three-shot results (Section 5.6). In Section 5.7 we examine how grammar prompting narrows the gap between LLM and SLM performance.

5.1 Beginner *vs.* Expert Explanations

Table 1 compares 'beginner' and 'expert' prompts across all datasets using gpt-3.5 and gpt-4o models. In a macro-analysis across 18 categories, simpler beginner-oriented explanations generally outperformed expert-level explanations by a small but significant margin (-1.9% ± 5.7%, p=0.002, d=-

0.344). Based on these results, we limited further investigations to 'beginner'-oriented prompts.

The prompting styles differ markedly: beginner prompts emphasize practical recognition ("Use 'who' or 'what' when the following part of the sentence is missing information"), while expert prompts employ technical terminology ("long-distance dependencies," "selectional restrictions"). Based on these results, we limited further investigations to beginner-oriented prompts. Additional examples are included in Appendix A.

5.2 Control *vs.* Textbook Conditions

Table 2 shows the results of two conditions we devised to understand how models utilize grammar prompts. In the **control** condition, we tested gpt-3.5 and gpt-4o models on BLiMP paradigms using a mismatched grammar prompt (explaining the "null quotative" - a syntactic phenomenon irrelevant to any tested minimal pairs). Performance was consistently worse than with GP-beginner's relevant explanations, indicating that models rely on the content of explanations rather than merely responding to their presence. In the **textbook** condition, we provided a compilation of *all* BLiMP-paradigm beginner grammar prompts as a simulated 'syntax textbook'. Performance was similar to the control condition, suggesting that models struggle to select and apply appropriate explanations when presented with too much information.

5.3 BLiMP Paradigms (English)

Table 3 presents BLiMP results (for additional results using Llama please see Appendix B).

The grammar prompting conditions consistently outperforms basic and chain-of-thought conditions, with GPb+CoT(o1) performing best; for example, Sonnet and gpt-4o achieve strongest performance with GPb+CoT (92.7% and 95.2%), and Llama follows at 88.7%.

Sonnet's and o1's grammar prompts vary in effectiveness by phenomenon: o1's prompt underperforms Sonnet's for argument structure and binding, while excelling at filler-gap dependencies. Interestingly, in the GP condition, the smaller models perform better with Sonnet's explanations while larger models prefer o1's.

Grammar prompting shows dramatic improvements in some categories – NPI licensing (c.f. the 'popsicle' example from the introduction) improves from 73.3% to 98.7% for gpt-4o and reaches perfect accuracy with Sonnet, with similar

gains in quantifier scope – though some complex phenomena like binding and island effects show only modest improvements.

5.4 SLING Paradigms (Chinese)

Table 4 shows SLING results (for additional results using Llama please see Appendix B). Grammar prompting again yields substantial gains, with GPb+CoT optimal: gpt-3.5's average climbs from 74.0% to 89.3%, and Llama leaps from 82.4% to 96.3%. While smaller models start at 70–80% and show large gains, Llama and larger models improve from 88–92% to 95–100% with GPb+CoT. Sonnet and o1's prompts perform similarly, though Sonnet's explanations are better for gpt-3.5 on alternative questions and definiteness paradigms.

The alternative question paradigm, which tests the constraint that the adverb *haishi* cannot co-occur with question particle *ma*, shows a striking pattern: CoT alone leads to systematic errors in smaller models (GPT-3.5: 16.7%, Haiku: 31.3%, Llama: 8.5%), but GPb and GPb+CoT achieve near-perfect results. Aspect particles improve from 70-90% to the high 90s with GPb+CoT, while classifier-noun paradigms show minimal gains due to the long-distance dependency and lexical nature. Definiteness and polarity paradigms, which test simple adverbs and demonstratives, consistently reach near-perfect accuracy with GPb.

5.5 RuBLiMP Paradigms (Russian)

Table 5 shows RuBLiMP results (for additional results using Llama see Appendix B). Grammar prompting again shows consistent gains, with GPb+CoT optimal: gpt-4o and Sonnet again achieve the strongest results (94.8% and 95.2% with GPb+CoT), followed by Llama (93.7%), with 5-15 point improvements over baseline across models. Prompt effectiveness varies by model and paradigm: smaller models struggle with Sonnet's argument structure and o1's government prompts, while Haiku strongly prefers Sonnet's aspect prompts (96.3% vs o1's 88.7%).

The reflexive possessor paradigm shows dramatic improvement, reaching near 100% with GPb as models grasp the constraint that *u + sebya* is allowed only as a locative and not a possessive. Aspect-related paradigms, which involve selecting perfect or imperfect verb form, improve from poor/moderate baselines to near-perfect scores with GPb+CoT, suggesting models know the mor-

Categories	base	CT	GPb son	gpt-3.5 GPb o1	GPb+CT son	GPb+CT o1	base	CT	GPb son	Haiku GPb o1	GPb+CT son	GPb+CT o1
NPI licensing (2)	60.0	67.3	69.7	67.3	**88.7**	87.0	57.3	73.3	77.0	65.3	96.0	**96.7**
arg. struct. (1)	77.3	66.0	79.3	46.7	74.7	66.0	74.7	85.3	86.7	74.0	82.7	86.0
binding (2)	65.3	57.3	67.7	57.3	68.7	62.0	56.7	63.7	64.7	50.7	74.3	70.7
control/raising (3)	83.1	79.1	92.9	87.8	94.4	82.9	69.3	83.1	75.1	70.7	94.2	90.4
ellipsis (2)	76.0	69.7	72.7	74.0	72.0	79.7	60.3	75.0	59.3	50.0	73.7	87.3
filler gap (3)	52.7	48.4	58.9	88.9	59.6	86.0	55.6	62.9	62.4	87.6	73.1	94.7
island effect (5)	68.9	61.6	65.1	63.1	74.4	70.9	60.9	64.7	60.5	59.1	73.3	74.3
quantifiers (2)	59.7	52.0	83.0	95.0	91.0	92.7	54.7	68.3	79.0	81.3	91.3	91.7
Average	67.9	62.7	73.6	72.5	77.9	78.4	61.2	72.0	70.6	67.3	82.3	86.5
Categories	base	CT	GPb son	gpt-4o GPb o1	GPb+CT son	GPb+CT o1	base	CT	GPb son	Sonnet GPb o1	GPb+CT son	GPb+CT o1
NPI licensing (2)	73.3	81.7	87.7	93.7	**98.7**	97.0	75.7	79.0	96.7	99.3	**100.0**	99.3
arg. struct. (1)	88.0	88.0	**90.7**	86.7	84.7	88.7	95.3	95.3	96.0	88.7	96.0	**96.7**
binding (2)	72.7	90.0	80.3	77.7	97.7	92.3	90.7	91.0	98.0	91.0	98.0	99.0
control/raising (3)	86.0	89.8	92.4	94.9	97.3	94.7	84.7	88.2	93.8	93.8	**98.4**	95.8
ellipsis (2)	70.3	85.7	69.7	88.7	59.3	**92.0**	81.0	88.0	77.0	98.0	62.0	**98.0**
filler gap (3)	75.3	80.7	81.3	94.4	84.7	**95.6**	80.2	71.3	90.4	97.6	85.6	**97.8**
island effect (5)	77.2	77.1	79.1	**82.1**	80.5	81.6	68.8	65.9	72.8	71.6	**81.3**	74.9
quantifiers (2)	81.0	86.3	**100.0**	99.3	**100.0**	**100.0**	62.3	63.7	97.7	98.0	**100.0**	**100.0**
Average	78.0	84.9	85.1	89.7	87.9	**92.7**	79.8	80.3	90.3	92.2	90.2	**95.2**

Table 3: Different prompting conditions for English-language BLiMP dataset. GPb = beginner-oriented grammar prompting, CT = Chain of Thought. For GP conditions, 'son' or 'o1' indicates grammar prompts generated by Claude 3.5 Sonnet or GPT-o1, respectively.

Categories	base	CT	GPb son	gpt-3.5 GPb o1	GPb+CT son	GPb+CT o1	base	CT	GPb son	Haiku GPb o1	GPb+CT son	GPb+CT o1
altern. quest. (1)	18.7	16.7	**100.0**	99.3	**100.0**	88.7	46.0	31.3	100.0	100.0	100.0	100.0
Aspect (5)	73.2	76.9	83.1	88.7	**94.1**	93.9	77.9	89.2	95.5	96.1	**98.9**	98.9
Classifier-Noun (6)	82.7	81.9	75.1	72.1	81.0	**82.8**	84.0	86.3	68.6	63.7	81.1	85.4
Definiteness (3)	97.6	89.8	**98.7**	94.2	**98.7**	86.0	88.7	95.8	96.0	97.1	98.0	99.1
Polarity Item (3)	73.8	83.1	96.7	98.7	99.1	**99.8**	75.1	79.8	100.0	99.6	100.0	100.0
Wh Fronting (2)	**98.3**	90.3	83.0	83.0	91.7	88.7	84.0	99.0	100.0	99.7	100.0	100.0
Average	74.0	73.1	89.4	89.3	**94.1**	90.0	75.9	80.2	93.3	92.7	96.3	**97.2**
Categories	base	CT	GPb son	gpt-4o GPb o1	GPb+CT son	GPb+CT o1	base	CT	GPb son	Sonnet GPb o1	GPb+CT son	GPb+CT o1
altern. quest. (1)	92.7	82.7	100.0	100.0	100.0	100.0	60.0	100.0	100.0	100.0	100.0	100.0
Aspect (5)	94.1	98.1	99.2	**99.7**	**99.7**	**99.7**	96.5	98.8	99.5	99.3	99.6	**99.9**
Classifier-Noun (6)	88.1	**90.0**	89.7	88.4	89.4	89.6	87.4	90.0	84.7	86.1	89.4	**90.6**
Definiteness (3)	94.7	95.8	99.8	100.0	99.6	100.0	95.6	96.9	100.0	100.0	99.3	100.0
Polarity Item (3)	89.6	95.1	100.0	99.6	100.0	100.0	94.2	95.3	100.0	100.0	100.0	100.0
Wh Fronting (2)	93.3	98.0	100.0	100.0	100.0	100.0	95.7	100.0	100.0	100.0	100.0	100.0
Average	92.1	93.3	98.1	98.0	98.1	**98.2**	88.2	96.8	97.4	97.6	98.1	**98.4**

Table 4: Different prompting conditions for Chinese-language SLING dataset

phology but need usage clarification. Case government improves from 70-90% to 80-97%, with Sonnet reaching 97-99% in GPb+CoT.

5.6 Three-shot Results

Table 6 reports a three-shot results alongside grammar prompting, contrasting *pattern-matching from examples* with our rule-based GP approach. With three labeled minimal pairs, many paradigms become solvable via surface similarity alone, offering a useful foil for testing formal grammatical competence.

The results diverge sharply by model size. For the two LLMs (Sonnet and GPT-4o), 3-shot pat-

Categories	base	CT	GPb son	gpt-3.5 GPb o1	GPb+CT son	GPb+CT o1	base	CT	GPb son	Haiku GPb o1	GPb+CT son	GPb+CT o1
Anaphora (1)	92.0	88.7	88.7	89.3	92.7	**93.3**	91.3	90.7	92.7	92.7	**96.7**	95.3
Arg. Str. (1)	68.7	**77.3**	62.0	**77.3**	56.7	70.0	**84.7**	**84.7**	73.3	79.3	74.0	80.7
Reflexives (1)	84.7	68.0	100.0	100.0	98.7	100.0	90.0	90.7	99.3	100.0	100.0	99.3
Subj.-Pred. Agr (5)	84.1	87.3	84.8	84.5	**92.3**	89.5	90.1	94.0	88.5	88.8	94.0	94.8
aspect (3)	58.9	64.7	85.8	86.4	**97.8**	97.3	63.6	73.8	97.6	88.0	100.0	99.3
government (2)	74.3	69.3	78.0	69.0	**83.3**	82.7	88.3	91.0	96.3	88.7	97.3	94.0
word formation (1)	90.0	90.0	98.7	94.7	94.7	79.3	98.7	100.0	100.0	87.3	100.0	98.0
Average	79.0	77.9	85.4	85.9	**88.0**	87.4	86.7	89.3	92.5	89.3	**94.6**	94.5
Categories	base	CT	GPb son	gpt-4o GPb o1	GPb+CT son	GPb+CT o1	base	CT	GPb son	Sonnet GPb o1	GPb+CT son	GPb+CT o1
Anaphora (1)	94.7	**96.7**	93.3	95.3	**96.7**	**96.7**	**96.0**	94.0	95.3	94.7	94.7	94.7
Arg. Str. (1)	88.7	87.3	**90.7**	90.0	88.0	84.0	**90.0**	89.3	89.3	88.0	89.3	86.7
Reflexives (1)	94.7	96.7	100.0	100.0	100.0	100.0	96.0	93.3	100.0	100.0	100.0	100.0
Subj.-Pred. Agr (5)	94.1	95.9	95.6	96.5	**98.1**	95.1	96.8	97.2	96.4	96.9	97.7	98.1
aspect (3)	90.7	86.7	100.0	98.9	100.0	99.8	96.2	92.2	100.0	99.3	100.0	99.3
government (2)	96.3	96.0	98.3	97.0	**99.0**	98.0	97.7	97.3	99.7	99.0	99.0	98.7
word formation (1)	98.7	**100.0**	100.0	100.0	100.0	99.3	98.7	99.3	99.3	98.7	100.0	97.3
Average	94.0	94.2	96.8	96.8	**97.4**	96.1	95.9	94.7	97.2	96.7	**97.2**	96.4

Table 5: Different prompting conditions for Russian-language RuBLiMP dataset

	Setting	Sonnet	GPT-4o	Haiku	GPT-3.5
EN	3-shot	**97.31**	**94.48**	73.63	65.26
	GP(b)-o1	90.17	87.86	**82.33**	**77.92**
CH	3-shot	97.44	98.03	79.90	79.75
	GP(b)-o1	**98.06**	**98.12**	**96.34**	**94.10**
RU	3-shot	**97.37**	96.61	83.21	75.59
	GP(b)-o1	97.19	**97.36**	**94.36**	**87.75**

Table 6: Macroaverage performance of three-shot vs grammar prompting. Three-shot beats grammar prompting for large models in English, but grammar prompting is dramatically more effective *on smaller LLMs* across all three languages.

tern matching is highly effective, outperforming GP on English and matching it on Chinese and Russian. For the three SLMs (gpt-3.5, Haiku, and Llama), the picture reverses dramatically: *grammar prompting exceeds 3-sh by an average of 12.1-13.1 percentage points (pp)*. GP therefore narrows the LLM–SLM performance gap precisely where few-shot learning leaves smaller models behind, reinforcing its value as a low-cost equalizer.

5.7 Effect on the LLM–SLM Performance Gap

To assess whether grammar prompting narrows the disparity between model sizes, we group the five models into SLMs (gpt-3.5, Haiku, Llama) and LLMs (gpt-4o, Sonnet). For each language we compute the mean accuracy within each group and define the *gap* as $\overline{\text{LLM}} - \overline{\text{SLM}}$. The bottom block of Table 7 reports the cross-language average of these three per-language gaps.

In the zero-shot baseline, the mean LLM–SLM gap is 13.0 pp. Adding grammar prompting alone reduces the gap to 10.4 pp, a 20% relative contraction. When chain-of-thought reasoning is supplied in addition to the grammar prompt (GP + CoT), the gap shrinks to 5.8 pp, a 56% reduction relative to the baseline. Hence a single LLM-generated rule explanation, especially when paired with step-by-step reasoning, brings three SLMs to within roughly six percentage points of gpt-4-class models across English, Chinese, and Russian, without any additional training or task-specific tuning.

6 Discussion

Our results across English, Chinese and Russian reveal clear patterns how grammar prompting works for different linguistic phenomena. It is most effective for phenomena governed by simple distributional constraints and categorical distinctions, such as negative polarity items in English, aspect marking in Russian and Chinese, or alternative questions in Chinese. These often reach near-perfect accuracy under grammar prompting, likely because they involve specific functional elements (e.g., *any, haishi, sebya*) that operate under specific licensing environments or paradigmatic

	Condition	SLM Avg	LLM Avg	Gap
English	Base	66.0	78.9	12.9
	CT	69.0	82.6	13.6
	GP	73.6	91.0	17.4
	GP+CT	**84.5**	**94.0**	**9.4**
Chinese	Base	77.4	90.2	12.7
	CT	77.1	95.1	18.0
	GP	92.8	97.7	4.9
	GP+CT	**94.9**	**98.3**	**3.4**
Russian	Base	81.5	95.0	13.4
	CT	82.4	94.5	12.1
	GP	87.9	**96.8**	8.9
	GP+CT	**91.8**	96.3	**4.4**
Average	Base	75.0	88.0	13.0
	CT	76.2	90.7	14.5
	GP	84.8	95.2	10.4
	GP+CT	**90.4**	**96.2**	**5.8**

Table 7: Accuracy (%) of small language models (SLMs) vs. large language models (LLMs) under different prompting conditions. Gap is the difference LLM-SLM; lower is better for SLMs. The bottom block reports the cross-language average.

choices like Russian perfect/imperfect verb aspect. This accuracy jump is especially pronounced for SLMs.

In contrast, phenomena that demand fine-grained constituent recognition (e.g. English island constraints) or language-specific lexical knowledge (e.g. Chinese classifier-noun agreement, wh-fronting) show smaller but still consistent accuracy gains. Appendix C documents instances of successes and characteristic failure modes: when the model *correctly identifies* the relevant constituent, grammar prompting enables rule-based judgments even on island violations; when it misparses or loses track of its own reasoning (as GPT-3.5 occasionally does on Chinese wh-fronting), the benefit is muted. These errors appear to stem less from the prompting paradigm itself than from (i) occasional constituent identification failures, (ii) the added cognitive load of explicit chain-of-thought reasoning, and (iii) our deliberate decision to omit illustrative examples. Addressing any of these factors – better parsing, better reasoning models, or simply adding in-context examples – could therefore push performance on the "hard" paradigms substantially higher.

We opened this paper with an incorrect conclusion drawn by Sonnet using CoT.

(A) Only a popsicle that Danielle admires ever freezes. ✓

(B) A popsicle that only Danielle admires ever freezes. ✗

With grammar prompting, Sonnet now reasons: *"The key difference is the placement of "only". In Sentence A, it's correctly placed at the beginning of the entire noun phrase, scoping over 'a popsicle that Danielle admires'. In Sentence B, it's incorrectly placed inside the relative clause, which changes the meaning and doesn't follow the grammar rule explained. Therefore, Sentence A is more grammatically acceptable according to the provided explanation."*

In the introduction, we observed how models can err by abstracting away surface but essential linguistic form in the attempt to decipher communicative intent; now, we see how grammar prompting reduces problematic 'semantically-driven reasoning' by 1) harnessing the model's metalinguistic ability to generate a grammar prompt and 2) feeding the grammar prompt back to the model, thereby directing its reasoning process to operate over the relevant linguistic structure.

7 Conclusion

LLMs are trained to distill away surface forms to capture meaning. This 'meaning-first' strategy powers functional tasks such as translation but can sabotage syntactic reasoning. Grammar prompting recalibrates that bias and does so *as a low-cost equalizer between model sizes*. With a single rule explanation generated by a large LLM and fed back with chain-of-thought, three SLMs finish on average within 5.8 pp of GPT-4-class models across English, Chinese, and Russian, cutting the capacity gap by 56%.

From general to specific reasoning. Grammar prompting anchors the model's attention on relevant grammatical categories and constraints.

From implicit to explicit knowledge. It turns latent metalinguistic knowledge into an explicit rule the model can apply, boosting accuracy even when data-driven cues are sparse.

The explain-then-process paradigm is attractive for (i) low-resource languages, where training data are scarce, and (ii) deployments that must rely on smaller, cheaper models when frontier LLMs are impractical.

Acknowledgments

Limitations

Our study has several important limitations. First, we only elicited grammar prompts from large models (Sonnet and o1), not exploring whether smaller models could generate effective explanations. Additionally, while we used English for all prompts, recent work suggests that prompting in the non-dominant language might yield better results in some cases (Behzad et al., 2024).

Second, our decision to exclude example sentences from grammar prompts, while methodologically sound for isolating the effect of explanations, may have also limited the prompts' potential effectiveness, particularly for complex phenomena like island constraints.

Finally, we focused solely on high-resource languages to control for vocabulary familiarity, but it remains untested how effective grammar prompting would be for low-resource languages: while low-resource languages often have well-documented grammatical systems that could inform prompt generation, their limited representation in LLM training data might affect both the models' ability to generate accurate grammatical explanations and to apply them correctly. This limitation warrants future investigation.

References

Wafa Abdullah Alrajhi, Hend Al-Khalifa, and Abdulmalik AlSalman. 2022. Assessing the linguistic knowledge in Arabic pre-trained language models using minimal pairs. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 185–193, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Seth Aycock, David Stap, Di Wu, Christof Monz, and Khalil Sima'an. 2024. Can llms really learn to translate a low-resource language from one grammar book? *Preprint*, arXiv:2409.19151.

Shabnam Behzad, Amir Zeldes, and Nathan Schneider. 2024. To ask LLMs about English grammaticality, prompt them in a different language. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15622–15634, Miami, Florida, USA. Association for Computational Linguistics.

Vittoria Dentella, Fritz Günther, and Evelina Leivada. 2023. Systematic testing of three language models reveals low language accuracy, absence of response stability, and a yes-response bias. *Proceedings of the National Academy of Sciences*, 120(51):e2309583120.

Mengnan Du, Fengxiang He, Na Zou, Dacheng Tao, and Xia Hu. 2023. Shortcut learning of large language models in natural language understanding. *Communications of the ACM*, 67(1):110–120.

Ping Guo, Yubing Ren, Yue Hu, Yunpeng Li, Jiarui Zhang, Xingsheng Zhang, and Heyan Huang. 2024. Teaching large language models to translate on low-resource languages with textbook prompting. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15685–15697, Torino, Italia. ELRA and ICCL.

Andrew Lampinen, Ishita Dasgupta, Stephanie Chan, Kory Mathewson, Mh Tessler, Antonia Creswell, James McClelland, Jane Wang, and Felix Hill. 2022. Can language models learn from explanations in context? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 537–563, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wei Qi Leong, Jian Gang Ngui, Yosephine Susanto, Hamsawardhini Rengarajan, Kengatharaiyer Sarveswaran, and William Chandra Tjhi. 2023. Bhasa: A holistic southeast asian linguistic and cultural evaluation suite for large language models. *Preprint*, arXiv:2309.06085.

Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Preprint*, arXiv:2301.06627.

Zhuang Qiu, Xufeng Duan, and Zhenguang G. Cai. 2024. Grammaticality representation in chatgpt as compared to linguists and laypeople. *Preprint*, arXiv:2406.11116.

Carson Schütze. 2016. *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Language Science Press.

Yixiao Song, Kalpesh Krishna, Rajesh Bhatt, and Mohit Iyyer. 2022. SLING: Sino linguistic evaluation of large language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4606–4634, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power,

Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karaka, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Koco, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S.

Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem enel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Micha Swdrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan A. Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Mikowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin

19788

Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Preprint*, arXiv:2206.04615.

Michelle Suijkerbuijk, Zoë Prins, Marianne de Heer Kloots, Willem Zuidema, and Stefan L Frank. 2025. Blimp-nl: A corpus of dutch minimal pairs and acceptability judgments for language model evaluation. *Computational Linguistics*, pages 1–39.

Ekaterina Taktasheva, Maxim Bazhukov, Kirill Koncha, Alena Fenogenova, Ekaterina Artemova, and Vladislav Mikhailov. 2024. Rublimp: Russian benchmark of linguistic minimal pairs. *Preprint*, arXiv:2406.19232.

Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. 2024. A benchmark for learning to translate a new language from one grammar book. In *The Twelfth International Conference on Learning Representations*.

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.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural Network Acceptability Judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in English? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.

Noam Wies, Yoav Levine, and Amnon Shashua. 2024. The learnability of in-context learning. *Advances in Neural Information Processing Systems*, 36.

Chen Zhang, Xiao Liu, Jiuheng Lin, and Yansong Feng. 2024. Teaching large language models an unseen language on the fly. In *Findings of the Association for Computational Linguistics ACL 2024*, pages

8783–8800, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

A Appendix: Instruction Template and Grammar Prompt Examples

Figure 4, 5, and 6 contain the full instruction template and three generated prompts: an 'expert-oriented prompt' by Sonnet, a 'beginner-oriented prompt' by Sonnet, and a 'beginner-oriented prompt' by o1, respectively.

B Appendix: Choosing Challenging Paradigms to Test

Initially, we tested gpt-4o to ascertain which paradigms its performance is low enough to allow for improvement over the base condition. Tables 8, 9, and 10 show these results for BLiMP, SLING, and RuBLiMP datasets, respectively. Table 11 presents the results of Llama models.

C Appendix: Success and Failure Cases

We present examples of models' success and failure cases under two grammatical constraints: the sentential island constraint in English and the wh-fronting constraint in Chinese. These cases demonstrate the strengths of grammar prompting over a solely-CoT approach, and show where the models still fail to fully utilize grammar explanations.

C.1 English Sentential Island Constraint (Sonnet)

We asked Claude 3.5 Sonnet to judge which of the following sentences is more grammatically acceptable:

> (A) Who had the patients' cleaning those banks upset.
> (B) Who had the patients' cleaning upset those banks.

(i) CoT in basic prompt (CoT failure case)

Excerpt from Sonnets response:

> ...Both sentences are confusing...In Sentence B: ... 'Who' is the subject, 'had' is the main verb, 'the patients' cleaning' is a noun phrase (object), ...Sentence B's structure is closer to a correct English sentence pattern. ...Sentence A's structure is less common and more difficult to interpret correctly ....

Analysis: Sonnet fails to identify the relevant grammatical constituents. It evaluates based on

Figure 4: Instruction for generating grammar explanations.

┌───┐

Sonnet-generated Expert grammar prompt

Title: Interrogative Phrase Extraction in Echo Questions
For expert linguists, it's crucial to understand the constraints on extracting interrogative phrases from certain syntactic positions in echo questions. This phenomenon relates to the left periphery of noun phrases and the limitations on moving wh-elements.
In grammatical constructions, the interrogative element (e.g., "what," "which") can be extracted when it's the entire object of a verb. However, when the interrogative element is embedded within a larger noun phrase, particularly as a determiner or modifier on the left edge, extraction becomes ungrammatical.
Key syntactic rules to note:

1. The interrogative element must remain in situ when it's part of a complex noun phrase.

2. Echo questions preserve the basic word order of the original statement.

3. The wh-word replaces only the specific element being questioned, not the entire phrase.

Linguists should pay attention to the following grammatical details:

- Determiner position within noun phrases

- Argument structure of verbs

- Prosodic patterns typical of echo questions

To distinguish correct usage, consider:

- Is the wh-word modifying a noun or standing alone as an argument?

- Does the question maintain the original sentence structure?

- Is there any violation of island constraints?

When analyzing potential violations, focus on the internal structure of noun phrases and the position of the interrogative element within them. Correct usage will maintain the integrity of the noun phrase, while incorrect usage will attempt to extract the wh-element from its embedded position.

└───┘

Figure 5: Sonnet-generated Expert grammar prompt for 'left branch island echo question' in the BLiMP paradigm.

Figure 6: Sonnet and o1-generated Beginner grammar prompt for 'left branch island echo question' in the BLiMP paradigm.

surface fluency and frequency, lacking access to the rule that governs island violations.

(ii) GP-CoT(o1) prompt (GP failure case)

Excerpt from Sonnets response:

1. First, we need to identify the main subject ... In both sentences, ... we can see that the subject is complex ...: 'the patients' cleaning.' This acts like the 'protective bubble' mentioned in the ex-

Category	Paradigm	gpt-3.5	gpt-4o
NPI LICENSING	matrix_question_npi_licensor_present	93.3	**98.0**
	npi_present_1	88.7	**96.7**
	npi_present_2	96.0	**100.0**
	only_npi_licensor_present	96.0	**98.7**
	only_npi_scope	**53.3**	52.0
	sentential_negation_npi_licensor_present	91.3	**99.3**
	sentential_negation_npi_scope	70.7	**89.3**
ANAPHORA AGREEMENT	anaphor_gender_agreement	**100.0**	100.0
	anaphor_number_agreement	99.3	**100.0**
ARGUMENT STRUCTURE	animate_subject_passive	82.7	**92.0**
	animate_subject_trans	87.3	**92.7**
	causative	86.0	**91.3**
	drop_argument	67.3	**89.3**
	inchoative	85.3	**90.7**
	intransitive	77.3	**97.3**
	passive_1	90.7	**95.3**
	passive_2	94.0	**94.7**
	transitive	84.0	**97.3**
BINDING	principle_A_c_command	91.3	**98.7**
	principle_A_case_1	**100.0**	100.0
	principle_A_case_2	**98.0**	96.7
	principle_A_domain_1	98.0	**99.3**
	principle_A_domain_2	88.7	**94.0**
	principle_A_domain_3	74.7	**81.3**
	principle_A_reconstruction	43.3	**63.3**
CONTROL/RAISING	**existential_there_object_raising**	**88.0**	86.0
	existential_there_subject_raising	90.7	**92.7**
	expletive_it_object_raising	82.7	**88.7**
	tough_vs_raising_1	72.0	**77.3**
	tough_vs_raising_2	76.0	**96.7**
DETERMINER-NOUN AGREEMENT	determiner_noun_agreement_1	98.7	**100.0**
	determiner_noun_agreement_2	99.3	**100.0**
	determiner_noun_agreement_irregular_1	90.7	**100.0**
	determiner_noun_agreement_irregular_2	97.3	**100.0**
	determiner_noun_agreement_with_adj_2	94.0	**96.7**
	determiner_noun_agreement_with_adj_irregular_1	88.7	**100.0**
	determiner_noun_agreement_with_adj_irregular_2	96.7	**98.0**
	determiner_noun_agreement_with_adjective_1	94.0	**99.3**
DISTRACTOR AGREEMENT	distractor_agreement_relational_noun	**100.0**	100.0
	distractor_agreement_relative_clause	93.3	**94.0**
ELLIPSIS	**ellipsis_n_bar_1**	80.0	**82.0**
	ellipsis_n_bar_2	**64.0**	55.3
FILLER-GAP	wh_questions_object_gap	86.7	**98.0**
	wh_questions_subject_gap	81.3	**97.3**
	wh_questions_subject_gap_long_distance	66.7	**90.7**
	wh_vs_that_no_gap	78.0	**98.0**
	wh_vs_that_no_gap_long_distance	68.0	**96.7**
	wh_vs_that_with_gap	50.7	**81.3**
	wh_vs_that_with_gap_long_distance	48.0	**54.0**
IRREGULAR FORMS	irregular_past_participle_adjectives	99.3	**100.0**
	irregular_past_participle_verbs	**100.0**	100.0
ISLAND EFFECT	**adjunct_island**	**86.7**	86.7
	complex_NP_island	63.3	**68.0**
	coordinate_structure_constraint_complex_left_branch	84.0	**93.3**
	coordinate_structure_constraint_object_extraction	90.7	**92.0**
	left_branch_island_echo_question	78.0	**92.0**
	left_branch_island_simple_question	97.3	**99.3**
	sentential_subject_island	**48.0**	48.0
	wh_island	66.7	**86.7**
QUANTIFIERS	existential_there_quantifiers_1	94.7	**97.3**
	existential_there_quantifiers_2	54.7	**79.3**
	superlative_quantifiers_1	88.7	**95.3**
	superlative_quantifiers_2	54.7	**79.3**
SUBJECT-VERB AGREEMENT	irregular_plural_subject_verb_agreement_1	**100.0**	100.0
	irregular_plural_subject_verb_agreement_2	**100.0**	100.0
	regular_plural_subject_verb_agreement_1	98.0	**100.0**
	regular_plural_subject_verb_agreement_2	98.7	**100.0**
	Average	84.0	<u>91.2</u>

Table 8: All 67 BLiMP paradigm results

planation. The explanation states that we shouldn't move words from inside this subject clause when forming ques-

tions... Looking at Sentence A... this sentence has moved 'those banks' out of the subject clause, which goes against

Category	Paradigm	gpt-4o-2024-08-06 accuracy
ALTERNATIVE QUESTION	**haishi_ma**	92.6
CLASSIFIER-NOUN	**cl_adj_comp_noun**	68.6
	cl_adj_comp_noun_v2	99.3
	cl_adj_simple_noun	95.3
	cl_comp_noun	72.0
	cl_comp_noun_v2	98.0
	cl_simple_noun	95.3
	dem_cl_swap	100.0
DEFINITENESS EFFECT	**definiteness_demonstrative**	88.0
	definiteness_every	96.0
POLARITY ITEM	**any**	85.3
	even_wh	98.0
	more_or_less	85.3
WH FRONTING	**bare_wh**	100.0
	mod_wh	86.6
ASPECT	**temporal_guo**	100.0
	temporal_le	100.0
	zai_guo	96.6
	zai_le	77.3
	zai_no_le	96.6
	Average	<u>92.8</u>

Table 9: SLING Paradigm Results

Category	Paradigm	gpt-4o-2024-08-06 accuracy
ANAPHOR AGREEMENT	**anaphor_agreement_gender**	92.0
	anaphor_agreement_number	98.0
ARGUMENT STRUCTURE	transitive_verb	100.0
	transitive_verb_iobject	99.3
	transitive_verb_object	99.3
	transitive_verb_passive	88.7
	transitive_verb_subject	96.0
ASPECT	**change_duration_aspect**	95.3
	change_repetition_aspect	96.0
	deontic_imperative_aspect	84.7
GOVERNMENT	adposition_government	100.0
	nominalization_case	96.0
	verb_acc_object	96.7
	verb_gen_object	94.7
	verb_ins_object	97.3
NEGATION	indefinite_pronoun_to_negative	98.0
	negative_concord	98.7
	negative_pronoun_to_indefinite	98.0
NOUN PHRASE AGREEMENT	floating_quantifier_agreement_case	100.0
	floating_quantifier_agreement_gender	98.7
	floating_quantifier_agreement_number	97.3
	np_agreement_case	97.3
	np_agreement_gender	99.3
	np_agreement_number	99.3
REFLEXIVES	**external_possessor**	95.3
SUBJECT-PREDICATE AGREEMENT	clause_subj_predicate_agreement_gender	98.7
	clause_subj_predicate_agreement_number	94.7
	clause_subj_predicate_agreement_person	96.7
	genitive_subj_predicate_agreement_gender	95.3
	genitive_subj_predicate_agreement_number	98.7
	genitive_subj_predicate_agreement_person	98.0
	noun_subj_predicate_agreement_gender	90.0
	noun_subj_predicate_agreement_number	100.0
	noun_subj_predicate_agreement_person	98.7
	subj_predicate_agreement_gender_attractor	91.3
	subj_predicate_agreement_number_attractor	96.7
TENSE AND MOOD	conj_verb_tense	99.3
	single_verb_tense	99.3
	tense_marker	97.3
WORD FORMATION	**add_new_suffix**	96.7
	add_verb_prefix	98.7
	change_declension_ending	98.7
	change_declension_ending_has_dep	98.7
	change_verb_conjugation	98.0
	change_verb_prefixes_order	100.0
	Average	<u>96.9</u>

Table 10: All 45 RuBLiMP paradigm results

Categories	Base	CT	GPb	GPb+CT
ENGLISH (**BLIMP**)				
NPI lic. (2)	62.3	66.7	87.7	**98.3**
Arg. Str. (1)	80.7	**82.6**	70.7	76.4
binding (2)	87.7	81.0	85.3	**90.3**
Cntrl/Rs. (3)	85.3	88.0	87.6	**91.3**
ellipsis (2)	54.0	70.3	69.3	**87.5**
filler gap (3)	55.7	61.3	94.4	**97.6**
island ef. (5)	65.1	70.8	71.9	**76.5**
quantifiers (2)	60.3	57.1	81.0	**91.3**
Average (BLiMP)	68.9	72.2	81.0	**88.7**
CHINESE (**SLING**)				
Alt. Question (1)	40.7	8.5	**100.0**	**100.0**
Aspect (4)	83.7	88.4	95.0	**97.0**
Classifier-Noun (6)	87.3	87.0	83.0	**88.1**
Definit. Ef. (3)	92.2	94.1	**100.0**	**100.0**
Polarity Item (3)	92.9	92.4	**100.0**	**100.0**
Wh Fronting (2)	97.7	97.0	**100.0**	**100.0**
Average (SLING)	82.4	77.9	96.3	**97.5**
RUSSIAN (**RUBLIMP**)				
Anaphora (1)	92.7	93.3	87.3	**95.3**
Arg. Str. (1)	86.7	**90.5**	87.3	90.3
Reflexives (1)	88.0	96.0	**100.0**	**100.0**
Subj.-Pred. Agr (5)	91.9	**94.0**	89.5	92.0
aspect (3)	75.6	79.9	91.1	**95.1**
government (2)	83.7	86.0	88.0	**94.9**
word formation (1)	92.0	**95.3**	87.3	87.8
Average (RuBLiMP)	87.2	90.7	90.1	93.6
Average over all	78.8	80.0	88.4	92.8

Table 11: Llama results on BLiMP, SLING and RuBLiMP over basic, CT, GPb(o1) and GPb+CT(o1)

the rule.

Analysis: Sonnet omits "those banks" from the sentential subject, misidentifying the full constituent as just "the patients' cleaning". As a result, it misapplies the constraint. This illustrates a case where the explanation is understood, but a failure in constituent recognition leads to incorrect reasoning.

(iii) GP-CoT(o1) prompt (GP success case)
Excerpt from Sonnets response:

> Let's approach this step-by-step using the grammar explanation provided ... Here, 'The patients' cleaning [of] those banks' is a complex subject that includes its own clause. It acts as a single unit.... According to the explanation, when forming questions, we shouldn't move words from inside the complex subject clause. The clause acts like a protective bubble... In Sentence B, the word order suggests that 'those banks' has been moved out of the subject clause and placed at the end. This violates the rule of not moving elements from within the subject clause....

Analysis: Sonnet correctly identifies the full sentential subject, understands the constraint (described in the grammar prompt as a bubble), and selects the sentence that violates it. The reasoning is structurally sound and consistent with the grammatical explanation.

Conclusion: In the CoT failure case, the model fails to supply appropriate categories of analysis, does not produce accurate rules, and relies on surface-level heuristics. In GP-CoT cases, the model reasons using categories and rules introduced in the prompt. GP failures arise when constituents are misidentified or reasoning chains are mismatched to the conclusion. In success cases, GP prompts enable correct identification of grammatical structure, appropriate constraint application, and accurate, rule-based reasoning. As models improve in reasoning, grammar prompting may become even more effective.

C.2 Chinese Wh-fronting (GPT-3.5)

GPT-3.5 performs strongly on Chinese wh-fronting tasks in zero-shot settings (e.g., 98.3% accuracy), but shows a noticeable drop under grammar prompting with chain-of-thought (GPb+CoT) conditions – falling to 88.7% (Table 4). This reflects a broader issue: grammar prompting introduces an additional layer of explicit reasoning, which GPT-3.5 does not always handle reliably.

Wh-fronting, common in English, involves moving a wh-word to the front of the sentence (e.g., I will hear it becomes What will I hear?). However, in Chinese, such fronting is ungrammatical. The correct structure keeps the wh-word in situ: I will

hear what? Thus, illegal wh-fronting is easily detectable to native speakers, and GPT-3.5 reliably chooses the correct option under basic conditions.

When prompted to reason explicitly about this rule, GPT-3.5 sometimes falters. Consider the following example from the wh_mod paradigm:

(A) *Tamen jiakuai* [VERB] *le shenme* [WHAT] *fazhan?* ✓
(B) *Shenme* [WHAT] *fazhan tamen jiakuai* [VERB] *le?* ✗

Excerpt from GPT-3.5's response:

> ...In Sentence B, the question word *shenme* ("what") comes before the verb ...Based on the grammar explanation provided, Sentence B is more grammatically acceptable because the question word *shenme* ("what") is correctly placed after the verb, following the rule that question words typically stay in the same position as the information they're asking about would be in a statement.

Analysis: GPT-3.5 begins with a correct observation that the wh-word appears before the verb in Sentence B. However, it later contradicts itself, claiming that the wh-word is after the verb and choosing the wrong sentence. This suggests a failure in internal consistency: although it correctly tracks the constituent position at first, it loses track of its earlier analysis during its reasoning chain. Such cases reveal that GPT-3.5 struggles with maintaining reasoning context and may misapply rules it appears to have learned.

Conclusion: GPT-3.5 exhibits characteristic reasoning failures in GP-CoT settings. These include:
- failure to track prior statements consistently,
- misapplication of otherwise correctly identified grammatical rules,
- distraction by irrelevant surface features.

While these issues are especially visible in GPT-3.5, similar patterns emerge across more complex paradigms even in stronger models. Grammar prompting clearly adds interpretive structure to a models linguistic reasoning process but that process itself remains susceptible to error.