Linguistic Research 42(1): 53-93
DOI: 10.17250/khisli.42.1.202503.003

Investigating ChatGPT's phonology problem-solving abilities through reasoning with varying custom instructions*

Hyesun Cho ** · **Sunwoo Park** · **Sanghoun Song** · **Eunjin Oh*****
(Dankook University · Keimyung University · Korea University · Ewha Womans University)

Cho, Hyesun, Sunwoo Park, Sanghoun Song, and Eunjin Oh. 2025. Investigating ChatGPT's phonology problem-solving abilities through reasoning with varying custom instructions. *Linguistic Research* 42(1): 53-93. In this paper, we investigate ChatGPT's ability to solve phonological problems with varying custom instructions. It is known that ChatGPT has difficulty solving unfamiliar novel problems, and phonology provides a good test case for evaluating its reasoning abilities when faced with such challenges. We designed phonological problems using data or rules that are not observed in the phonology literature. We varied model versions (GPT-4 and GPT-4o) and custom instructions by varying the levels of knowledge: no custom instructions, beginner ('Student'), and expert ('Professor') levels. Four novel questions were created by modifying the following phonological data: (1) Korean fricative palatalization, (2) Spanish intervocalic lenition, (3) Rule ordering in Canadian English, and (4) English syllabification. ChatGPT exhibits some level of reasoning capabilities, with test scores ranging from 27% to 75%. GPT-4o performed better than GPT-4. However, it made errors when dealing with novel patterns or rules. The effects of varying custom instructions were present but less clear and inconsistent. Our results suggest that, depending on the model version and customization, GPT can learn phonologically unnatural processes through reasoning. **(Dankook University · Keimyung University · Korea University · Ewha Womans University)**

Keywords ChatGPT-4, ChatGPT-4o, phonology, reasoning, custom instructions

* This work was supported by the Ewha Frontier 10-10 Project grant funded by Ewha Womans University (1-2023-0154-001-3).
** First Author
*** Corresponding author

1. Introduction

1.1 ChatGPT and its reasoning ability

ChatGPT has attracted attention from both the public and academics since its release (GPT-3.5) on November 30, 2022. ChatGPT is a generative, conversational AI that can hold a conversation with human users by generating answers. The underlying model of ChatGPT is GPT(Generative Pre-trained Transformer), a large language model (LLM) with the transformer architecture (Vaswani et al. 2017) which predicts the next words given the present input. The GPT-4 version of ChatGPT was released on March 13, 2023. The GPT-4o version, released on May 13, 2024, is multi-modal (audio, vision, and text), and features stronger reasoning ability than previous versions[1]. GPT-4o demonstrated the same level of performance as GPT-4 Turbo in terms of reasoning[2]. For example, GPT-4o shows higher accuracy than GPT-4 in M3Exam[3], which requires language understanding and complex reasoning (Zhang et al. 2023).

Studies show that GPT-4 possesses some level of reasoning ability (Liu et al. 2023). GPT-4 has strong reasoning ability, can generate reasoning steps for arithmetic questions, and can answer complex logic questions (Qin et al. 2023; Wu et al. 2023). It is different from rule-based chatbots where structured interactions are pre-programmed (Zohuri and Rahmani 2023). Instead of responding with a limited set of pre-defined sentences, generative LLMs understand and generate novel sentences, which is considered language competence. Metalinguistic abilities such as reasoning may emerge from language competence (Beguš et al. 2023).

To test LLMs‘ reasoning ability, it is essential to use novel problems rather than pre-existing ones, since novel problems require a model to rely on reasoning rather than on stored knowledge. Liu et al. (2023) tested ChatGPT’s reasoning ability by using logical reasoning datasets that include multiple choice reading comprehension and natural language inference. GPT-4 performed better than the baseline model, RoBERTa. However, its performance significantly dropped when dealing with

1 GPT-o1 is released on September 12, 2024, which has a stronger reasoning ability, introduced as “a new series of reasoning models”, designed to solve harder problems in science, coding and math. The present study was conducted on August 10, 2024, before the release of GPT-o1.

2 https://openai.com/index/hello-gpt-4o/

3 A multilingual, multimodal, multilevel benchmark for examining large language models. It is developed based on real and official human exam questions (Zhang et al. 2023).

out-of-data distribution (OOD) datasets, which refer to datasets that are significantly different from the data that the model was trained on. That is, ChatGPT models "struggle to handle new and unfamiliar data" (Liu et al. 2023: 6). In linguistics, Beguš et al. (2023) and Oh et al. (2023) tested ChatGPT's reasoning ability by devising artificial problem sets.

When testing GPT's ability to solve linguistic problems, it is possible that the model may simply memorize answers from its training data rather than generate them through reasoning. Beguš et al. (2023: 4, 24) refer to this as a "memorization" problem. That is, GPT is trained on a very large training dataset, which likely contains linguistic analyses including problem sets and solutions found in linguistics textbooks, and it memorizes and retrieves these sets and analyses when solving problems. As a result, we cannot distinguish whether GPT's answer relies on memorized knowledge or genuine reasoning. To avoid this problem, Beguš et al. (2023) and Oh et al. (2023) designed novel problems in such a way that potential solutions are not already present in the model's training data, either by creating new versions of existing problems or by developing an entirely new language.

Beguš et al. (2023) examined ChatGPT's responses to syntactic, phonological, and semantic problems. They had two phonology problems—one existing, the other artificial: Korean palatalization, a well-known problem whose answer can be easily found in linguistic textbooks; and spirantization in an artificial language, a novel problem. They found that GPT-4 performed better in Korean, but less well in analyzing the data in an artificial language.

Oh et al. (2023) also tested ChatGPT's ability to solve linguistic problems, focusing on phonology. They compared GPT-4, GPT-3.5, and human answers to the same problem set. The problem set included conceptual questions[4] and data set analysis questions. The data set analysis questions were designed to test GPT's reasoning ability. As in Beguš et al. (2023)[5], they created novel problems from hypothetical languages that ChatGPT could not possibly have encountered in its training. In solving these problems, GPT-4 performed better than GPT-3.5. Compared to human testees, both models performed better on conceptual questions that simply require stored knowledge

4 Conceptual questions test simple factual knowledge that can be easily found in various resources, e.g., "Explain the terms "phoneme" and "allophone" with examples" (#3 in Oh et al. (2023) problem set).

5 Beguš et al. (2023) is a manuscript as of August 21, 2023. The experiment in Oh et al. (2023) was conducted on March 2023.

but worse on questions requiring reasoning ability[6].

The implication of Beguš et al. (2023) and Oh et al. (2023) is that the LLMs not only possess language competence, but reasoning ability as well. The abilities beyond language performance were referred to as "metalinguistic abilities", defined as "the ability to analyze language itself and to generate formal, theoretical analyses of linguistic phenomena" and is "cognitively more complex than language use" (Beguš et al. 2023: 3). Linguistic data analysis is thus an effective way to test an LLM's reasoning ability. Furthermore, phonology offers a useful test case because solving phonological problems requires extracting the phonological environment and making abstract generalizations over a string of segments, which demands reasoning ability. In addition, it is relatively easy to create new, artificial data simply by replacing segments or reverse existing rules or patterns.

Building on Beguš et al. (2023) and Oh et al. (2023), we focus on phonological problems that require ChatGPT's problem solving ability that hinges on reasoning ability. We use hypothetical, novel language data to prevent ChatGPT from finding answers from stored knowledge.

1.2 Custom instructions

As we test ChatGPT's problem solving ability, we vary custom instructions to examine whether custom instructions may affect ChatGPT's responses and performance. Custom instructions are a feature added to ChatGPT on July 20, 2023. Using custom instructions, one may add preferences or requirements that they want ChatGPT to consider when generating its responses[7]. Through custom instructions, ChatGPT can be tailor-made to meet users' needs. The information about the user or the preference of the response styles can be preset in custom instructions. Once set, ChatGPT considers the user's custom instructions in all of its conversations, so the user does not have to repeat the same preferences or information every time they use ChatGPT[8].

6 In total, ChatGPT scored 59% (GPT-3.5), 71% (GPT-4), compared to the human average (85.5%). In reasoning problems, however, the differences were greater: 36% (GPT-3.5), 46% (GPT-4), and 94% (human average). The percentage values were calculated based on Table 3 in Oh et al. (2023: 82), for easier comparison.

7 https://openai.com/index/custom-instructions-for-chatgpt/

8 An example given by OpenAI is the following: "a teacher crafting a lesson plan no longer has to repeat that they're teaching 3rd grade science."

Through custom instructions, we can tailor prompts, give specific instructions on the style of outputs, and narrow down the area of specialty. In general, ChatGPT performs better if the specific domain of interest is given at the beginning of a conversation because it allows ChatGPT to have access to contextual documentation (Garrido-Merchán et al. 2023). In Garrido-Merchán et al. (2023), their customized GPT provided better responses when explicitly prompted, "I would like to practice a programming exercise similar to those in R practice 4".

Other prompting methods that are known to be effective include in-context learning and Chain-of-Thought prompting (Kojima et al. 2023; Wu et al. 2023). ChatGPT's performance improves when related questions are asked in the same conversation window, which is called "in-context learning" (Liu et al. 2023). CoT prompting is to get the model to think step by step, which helps its reasoning process. Just adding "Let's think step by step" to the prompt improves performance (Kojima et al. 2023). Adjusting custom instructions can also be considered a prompting method. It is equivalent to repeating the same, customized prompts at the start of every conversation.

So far, only a few studies have tested effects of custom instructions on ChatGPT's performance in problems solving. Kumar and Kats (2023) used ChatGPT (versions 3.5, 4, and 4 with Code Interpreter) to solve introductory college-level vector calculus and electromagnetism problems. They tested ChatGPT with the same problem set many times. Since ChatGPT's answers are stochastic, the answers are not identical each time. The researchers discovered that when asked many times, the most frequent answer is the best solution. However, adjusting custom instructions did not measurably improve the problem-solving performance in their study. Garrido-Merchán et al. (2023) developed a customized GPT, a 'Business Statistics Virtual Professor (BSVP)', comparing with non-customized GPT-4 Turbo. The BSVP showed a substantial modification in the communication style, but there were no significant improvement in the quality of responses. Nevertheless, given the small number of studies conducted so far, it may still be worth testing the possible effects of custom instructions.

Therefore, in our study, we test ChatGPT's ability to solve phonological problems under varying custom instructions. We had three types of custom instructions assuming different levels of knowledge: Student, Professor, and No custom instruction (corresponding to beginner, expert and default condition, respectively). We compare the answers from GPT-4 and GPT-4o versions.

The rest of this paper is organized as follows. In Section 2, we describe the experiment procedure, the phonological problem set, the models, and the custom instructions that we used. In Section 3, we report the results of the experiment, analyzing and comparing the answers from different model versions and custom instructions. A summary of the results and a general discussion are provided in Section 4. Section 5 concludes the paper.

2. Experiment

2.1 Problem set

As mentioned in Section 1.1, we used novel problems to prevent ChatGPT from generating answers from stored knowledge. There were four questions. Two of them (Questions 1 and 2) are the data-analysis questions adopted from Oh et al. (2023). Their study was conducted in 2023, so we may find differences between their results and the responses from the current version of ChatGPT (August 2024)[9]. The other two questions (Questions 3 and 4) focus on determining the rule ordering for novel words and applying a new rule for syllabification. In all the questions, we asked the model to explain its answers. This section describes the problem set. See Appendix A for the entire problem set.

2.1.1 Question 1: Pseudo-Korean palatalization (Oh et al. 2023)

We provided the model with a set of words created by modifying Korean palatalization process involving [s] and [ʃ], two allophones of /s/. We asked the model to determine whether [s] and [ʃ] are phonemes or allophones, and if they are allophones, to describe the phonetic environments where each allophone occurs. In Korean, /s/ becomes [ʃ] only before a high front vowel /i/ (e.g. [suʃin] 'reception') (Ahn 1985; Jun 1996; Hahm

9 We discussed the possibility of ChatGPT learning from the results of Oh et al. (2023), which may affect the answers in the current study. However, it does not seem likely because given vast amounts of training data, the result from only one paper or a few instances of conversations they have conducted would not likely affect ChatGPT's answers. Even if the current version of ChatGPT should learn the results from Oh et al. (2023), the probability of adopting the result of one paper, compared to a greater majority of evidence from the existing literature indicating the opposite pattern, would be low.

2007). This process is well-known and can be found in many resources such as linguistics textbooks (e.g. Yavaş 2020). The process is modified so that palatalization occurs before /i/ and /e/ (e.g. [suʃemi] 'scrub brush'). Oh et al. (2023) created this problem based on Davenport and Hannahs (2020) but replaced the words so that such data would not appear in any publicly available resources. The intended answer is front vowels, covering both /i/ and /e/. If ChatGPT generated its answer from stored knowledge, the answer would be "/s/ becomes [ʃ] before a high front vowel (/i/)". Instead, if it analyzed the given data correctly, its answer would be "/s/ becomes [ʃ] before front vowels (/i, e/)". Beguš et al. (2023) used the same Korean phenomenon, but they left it unmodified in order to compare the results with those from an artificial language.

2.1.2 Question 2: Intervocalic stopping (Oh et al. 2023)

We provided the model a set of words created by reversing Spanish intervocalic lenition, a process unlikely to occur in any natural language. We asked the model to describe the phonetic environment where allophones [d] and [ð] occurs in a hypothetical language. In Spanish, /d/ becomes [ð] between vowels (e.g. [nɐðɐ] 'nothing', but *[nɐdɐ]), known as stop lenition (Carr 2013; Colantoni et al. 2022; Broś and Krause 2024). In the modified data, we created non-Spanish words and reversed the process so /ð/ becomes [d] between vowels. It is then intervocalic stopping instead of lenition (e.g. [kadu], not *[kaðu]). If ChatGPT generated its answer referring to stored knowledge, the answer would be "[ð] occurs between vowels, and [d] occurs elsewhere". If it strictly followed the instruction in the problem without referring to pre-existing data, its answer would be "[d] occurs between vowels, and [ð] occurs elsewhere".

2.1.3 Question 3: Rule ordering

The rule ordering question was based on Canadian Raising and Tapping (Hayes 2009), but the words and the environment were changed. All the words were novel words from a hypothetical language, so ChatGPT cannot refer to any existing data. Two rules were given: Vowel Raising (VR) and Intervocalic Voicing (Voicing). In this hypothetical language, Vowel Raising raises /a/ to [u] before voiced consonants,

whereas in Canadian English, vowel raising occurs before voiceless consonants. The question consists of four sub-questions (#3-1~3-4). The model was asked to complete the two derivation tables under two possible orderings: first, applying VR and then Voicing (#3-1), and second, applying Voicing first and then VR later (#3-2), as shown in (1) and (2) below. The examples in (1) and (2) show the correct derivation tables for Questions #3-1 and #3-2. It is assumed that 'painter' is /radən/, and 'swimmer' is /rudən/ in this language.

(1) [Question #3-1] The rule ordering: VR precedes Voicing.

	'painter'	'swimmer'
Phonemic forms	/ratən/	/radən/
Vowel Raising	ratən	rudən
Voicing	radən	rudən
Phonetic forms	[radən]	[rudən]

(2) [Question #3-2] The rule ordering: Voicing precedes VR.

	'painter'	'swimmer'
Phonemic forms	/ratən/	/radən/
Voicing	radən	radən
Vowel Raising	rudən	rudən
Phonetic forms	[rudən]	[rudən]

In the next question (#3-3), we asked the model to determine the correct rule ordering when phonetic forms are [radən] and [rudən] respectively. The correct answer is the ordering in (1) (VR precedes Voicing) because it results in the correct phonetic forms. In Question #3-4, we asked the model to find the phonemic form of hypothetical 'runner' [mabən]. Since /a/ is not raised to [u] before a voiced consonant [b] in the surface form, [b] is underlyingly /p/, so the correct phonemic form is /mapən/.

2.1.4 Question 4: Syllabification

In Question 4, we provided ChatGPT with a modified version of a syllabification rule and asked the model to syllabify four English real words and four non-words. The modified syllabification rule was "Assign intervocalic consonants to the onset

of the following syllable", which is different from the English syllabification rule where intervocalic consonants are supposed to attach to the following syllable as much as English phonotactics allow[10]. According to the modified rule, all of the intervocalic consonants are parsed as onsets to the following syllable regardless of language-specific phonotactics, e.g. athletic is supposed to be syllabified as /.æ.θlɛ.tɪk./ according to our rule, instead of /.æθ.lɛ.tɪk./.

We prompted ChatGPT to follow only this modified rule and not refer to dictionaries ("Make sure to answer the questions based strictly on the rules provided above, not on any information from dictionaries."). We intended to test if ChatGPT can understand instructions correctly and apply the rule as specified in the problem, even if the results conflict with prior knowledge.

2.1.5 Predictions on GPT's performance

GPT's performance may vary depending on how each question is created. Question 1 was created by adding another environment (/e/) to the existing one (/i/). Palatalization of alveolar consonants in the context of /e/ instead of /i/ is cross-linguistically rare, though velar palatalization occurs in the context of non-high vowels (e.g. in Serbian, čove[k] – čove[ʧ]e 'human' nominative sg. – vocative sg.) (Fischer 2003). Therefore, it is unlikely that GPT has seen exactly the same data and process as Question 1. We expect that GPT will be able to easily identify the vowel /i/ as the environment for palatalization, but the vowel /e/ is more likely to be missing.

Unlike Question 1, in Questions 2 and 3, the phonological patterns were reversed entirely: intervocalic lenition becomes intervocalic stopping (Q2), and vowel raising applies before voiced instead of voiceless consonants (Q3). In Question 4, the imposed syllabification rule is incomplete and incorrect, as it fails to consider English Phonotactics where it should.

Therefore, we can say that Question 1 is weakly modified, whereas Questions 2-4 exhibit almost non-existent, phonetically unlikely patterns. Such processes, considered unnatural in the literature, are known to pose learnability problems for human learners (Hayes and White 2013). We predict that GPT will perform better on patterns that closely resemble natural processes because the training data likely contains far more

10 In Hayes (2009: 253), the Onset Formation rule states, "Join consonants to the following syllable, provided the resulting cluster can occur at the beginning of a word."

natural processes than unnatural ones. For this reason, we expect GPT to perform better on Question 1 than on the other questions due to their differing degrees of modification and naturalness. Nonetheless, if GPT succeeds in solving unnatural problems, this would suggest that it can learn unnatural patterns through reasoning. We will discuss this further in Section 4.2.

2.2 Custom instructions

Custom instructions[11] consist of two questions: one asking, "What would you like ChatGPT to know about you to provide better responses?", and the other asking, "How would you like ChatGPT to respond?" (See the Appendix 3 for the screen capture image).

We wrote three types of custom instructions: No custom instruction, Student, and Professor (See the Appendix 2 for all custom instructions). We intended that these types correspond to different levels of expertise: no specification, beginner, and expert. For No custom instruction, nothing is entered in the custom instruction window. For the beginner level, we described the user as a first-year undergraduate student who hopes to major in linguistics ('Student'). Student does not have much knowledge in linguistics, except a few basic concepts: phonemes, allophones, IPA symbols, and the phonological rule format (X→Y/A_B).

For the expert level, we described the user as a professor of linguistics specializing in phonology ('Professor'). Professor custom instructions included general guidelines and knowledge of phonemic analysis, phonological rule formulation, rule ordering, and application of phonological principles. The guidelines also included an instruction about the linear order of consonants and vowels[12], as follows, because ChatGPT (3.5 and 4) made substantial mistakes in recognizing the linear ordering of segments (Oh et al. 2023).

"Be accurate in the order of consonants and vowels in words. E.g., in [kæst],

11 Custom instructions can be entered as follows: User profile > Customize ChatGPT > Enter custom instructions.

12 This is not to assume that human students cannot recognize the ordering of segments while professors can. This was simply an attempt to more clearly differentiate beginner and expert knowledge sets by explicitly providing more detailed instructions.

the word-initial consonant is [k], the word-final consonant is [t], the vowel is [æ], and the consonant after the vowel [æ] is [s]."

This instruction includes a concrete example of [kæst]. This can be considered few-shot learning (Kojima et al. 2023), which is known to improve GPT's performance.

It was also emphasized that phonological generalization is crucial and the model should use natural classes where applicable, rather than making separate rules for individual segments. There were also step-by-step instructions on how to determine correct rule ordering.

However, due to a 1500-character limitation, the customization instructions had to be brief. The length of the instructions was the same for Student and Professor. Due to this limitation, we conducted an additional experiment by creating a customized GPT (using a function in ChatGPT called "myGPT"), which allows the users to create tailored AI chatbots for specific tasks[13]. In myGPT, we entered more detailed descriptions of phonology problem solving (3,878 characters). This can be considered an expanded version of Professor. We named it "Phonology Rule Solver" (abbreviated as Solver). The experiment with Solver was conducted on a separate date, October 10, 2024. We report the results of this additional experiment together with the others in Section 3.

2.3 The models

We used ChatGPT versions GPT-4 and GPT-4o with ChatGPT Plus accounts. All the conversations for the experiment were conducted on August 10, 2024, so that there would be no differences based on the date of the model version.

2.4 Experiment procedure

In order to minimize any unknown side-effects when using the same ChatGPT account over and over on the same computer, each of the custom-instruction conditions was tested on different computers with different ChatGPT accounts.

13 How to build the customized GPT: User profile > My GPTs > Create a GPT. See Appendix 4 for the screen capture image.

Table 1. Experiment settings: custom instructions and models

Custom instructions	Level of Expertise	Experimenter	Models
No custom instructions	None	Author 1	GPT-4, GPT-4o
Student	Beginner	Author 2	GPT-4, GPT-4o
Professor 1	Expert	Author 3	GPT-4, GPT-4o
Professor 2	Expert	Author 2	GPT-4, GPT-4o
Solver	Expert	Author 3	GPT-4 Turbo

The three of the co-authors of this paper each took on different roles (No custom instruction, Student, Professor 1), as in Table 1, and conducted conversations using their own ChatGPT Plus accounts and computers. All the conversations were monitored by all co-authors through real-time Zoom screen-sharing. The conversations were conducted in the following order: No custom instructions, Student, and Professor 1. This is the order in which the level of linguistic expertise increases. The entire four-question problem set was asked in the same conversational window, with each problem presented one after another. The order of the models was kept consistent across the experimenters: GPT-4o was tested first, and then GPT-4 later.

Professor customization was tested twice by different experimenters because the results from the first attempt seemed problematic: some responses of Professor 1 seemed worse than those of Student. Because of this, another author (Author 2), who had run Student customization, ran Professor customization. These two different Professor customizations were named Professor 1 and Professor 2, respectively, and we analyze both Professors' responses in the results section. The experiment with Solver customization, explained in Section 2.2, was conducted by Author 3.

After the experiment, ChatGPT's answers were graded by the three authors who specialize in phonology (Graders 1, 2, 3). As summarized in Table 2, Questions 1, 2, 3-1, 3-2 were graded by each grader with his or her own criteria. For Questions 3-3, 3-4, and 4, the answers and explanations were graded separately. The "answer" parts of Questions 3-3, 3-4, and 4 were graded in a dichotomous basis. No partial credit was awarded, so correct answers received full points while incorrect answers received zero. This dichotomous scoring was conducted by Grader 1, while the other graders verified its accuracy. All graders graded ChatGPT's answers individually, without knowing other graders' scores.

Table 2. Grading methods

Question number			Point	Grader
1			10	Grader 1, 2, 3
2			10	Grader 1, 2, 3
3	3-1		4	Grader 1, 2, 3
	3-2		4	Grader 1, 2, 3
	3-3	Answer	2	Grader 1 (Verified by Grader 2, 3)
		Explanation	4	Grader 1, 2, 3
	3-4	Answer	2	Grader 1 (Verified by Grader 2, 3)
		Explanation	4	Grader 1, 2, 3
4	Answer		8	Grader 1 (Verified by Grader 2, 3)
	Explanation		4	Grader 1, 2, 3

After collecting the scores, scores were averaged across graders. Inter-rater reliability was obtained based on Cronbach Alpha (using the *psych* package in R) to see how consistently three raters graded the answers. In the next section, we compare the test scores and analyze the results.

3. Results

3.1 Overall results

The mean of Cronbach Alpha values across all graded scores was 0.93 (0.98, 0.91, 0.96, 0.87, 0.91 for each question respectively), with standard deviation of 0.04, which indicates high inter-rater reliability.

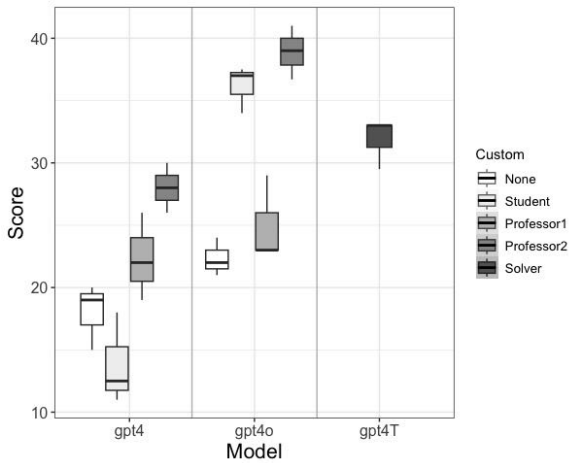


Figure 1. Mean scores by models (GPT-4, GPT-4o, and GPT-4 Turbo) for each custom instruction

Table 3. Mean scores by question for each custom instruction and model(SD: Standard Deviation of the mean values among raters)

Question Number		Point	No custom instruction		Student		Professor1		Professor2		Solver
			GPT4	GPT4o	GPT4	GPT4o	GPT4	GPT4o	GPT4	GPT4o	GPT4 Turbo
1		10	4	7	2	10	8	9	8	10	10
2		10	3	5	3	5	3	4	6	6	7
3	3-1	4	3	3	2	4	3	3	3	2	1
	3-2	4	4	4	2	4	3	3	2	3	2
	3-3Ans	2	0	0	2	2	0	0	2	0	2
	3-3Exp	4	0	0	1	4	1	1	1	1	1
	3-4Ans	2	0	0	0	2	0	0	0	2	2
	3-4Exp	4	0	1	1	4	1	2	1	3	3
4	Ans	8	1	2	0	1	2	2	3	8	2
	Exp	4	2	1	1	1	1	1	3	4	2
Total		52	18	22	14	36	22	25	28	39	32
		SD	1.7	2.3	1.1	2.6	2.3	2.7	2.3	3.1	2.0
		%	35%	43%	27%	70%	43%	48%	54%	75%	61%

Figure 1 and Table 3 show mean scores across raters for each model and each custom instruction[14]. The mean percentage score across all the models and custom instructions was 51% (*SD* = 16). Overall, the effects of the model versions were clear. GPT-4o tended to have a higher score than GPT-4. The higher-level knowledge (Professor 1 and 2) from the lower model (GPT-4) scored similarly to the lower-level knolwedge (No custom instruction) from the higher model (GPT-4o).

On the other hand, custom instructions did not have consistent effects. It appears that high-level knowledge has a positive effect on the model's performance in many cases: the highest mean score was found in the expert customization (Prof2/GPT4o), and the lowest score was found in the beginner customization (Student/GPT4). However, the conflicting patterns are also found. In GPT-4o, Student's mean score was higher than that of Professor 1. In GPT-4, Student's mean score was lower than that of No custom instructions, though the difference was not statistically significant (t(29)=1.6, p=0.11). In GPT-4o, Prof1's mean score was lower than that of Student and not significantly different from that of No custom instruction (t(28)=-1.44, p=0.17).

Table 4. Mixed-effects linear regression results for mean scores

	Estimate	Std. Error	df	t	Pr(>|t|)	
(Intercept)	18	1.53	4.77	11.73	<0.001	***
custom_Professor 1	4.33	1.39	16	3.12	0.007	**
custom_Professor 2	10	1.39	16	7.19	<0.001	***
custom_Solver	13.83	1.39	16	9.94	<0.001	***
custom_Student	-4.17	1.39	16	-3.00	<0.01	**
model_GPT-4o	4.33	1.39	16	3.12	<0.01	**
custom_Prof1:model_GPT-4o	-1.67	1.97	16	-0.85	0.409	
custom_Prof2:model_GPT-4o	6.57	1.97	16	3.34	<0.01	**
custom_Student:model_GPT-4o	18	1.97	16	9.15	<0.001	***

To test the statistical significance of the variables, a mixed-effects linear regression was fitted to the data using the *lmerTest* package in R. The dependent variable was

14 Each model and customization setting will be abbreviated as follows: No custom/GPT4, No custom/GPT4o, Student/GPT4, Student/GPT4o, Prof1/GPT4, Prof1/GPT4o, Prof2/GPT4, Prof2/GPT4o, Solver.

the score, the fixed effects were custom instruction, model, and their interactions, and random intercepts for each grader were included. The baseline was No custom instruction and GPT-4. The resulting coefficient estimates imply the followings: Prof1, Prof2, and Solver custom instructions had significantly higher scores than No custom instruction ($t(16)=3.12$, $p<0.01$, $t(16)=7.19$, $p<0.001$, $t(16)=99.94$, $p<0.001$, respectively). Student custom instructions resulted in a significantly lower score than No custom instruction ($t(16)=-3.00$, $p<0.01$). GPT-4o had a significantly higher score than GPT-4 ($t(16)=3.12$, $p<0.01$). In GPT-4o, Professor 2 and Student custom instructions had significantly higher scores than No custom instruction ($t(16)=3.34$, $p<0.01$, $t(16)=9.15$, $p<0.001$, respectively) but Professor 1 was not different from No custom instruction ($t(16)=-0.85$, $p=0.41$).

The results so far can be summarized as follows. Firstly, having higher-level knowledge in the custom instructions is likely to improve the results, but the effect appears limited and inconsistent. The mixed-effects results show that the magnitude of coefficients does not strictly follow the level of expertise. Secondly, the model version has more consistent and stronger effects: GPT-4o outperformed GPT-4 under the same customization. In conclusion, while the advanced model version clearly led to better results, the effect of custom instructions was less clear and inconsistent.

3.2 Analysis of individual problems

3.2.1 Question 1: Palatalization before a front vowel

1) Phonological environment

The correct phonological environment for [ʃ] in Question 1 is before a front vowel (both high and mid) or [i, e]. Table 5 summarizes each model's answers. There were sometimes errors in a model's explanation but the final answer was correct[15]. In such cases, Table 5 shows the final answers only, for brevity.

15 For example, in Prof1/GPT-4, it says that [ʃ] appears before the vowels [i], [e], [a], and [i], but in its final answer it concludes that [ʃ] appears before front unrounded vowels ([i], [e]). The graders deducted points for wrong explanations in the answers.

Table 5. Phonetic environments described by each model and customization setting for Question 1. Errors are underlined. Shaded cells indicate the correct answers.

Custom	Model	Phonemic status	Phonetic environment	
			[ʃ]	[s]
None	GPT-4	Allophone	[i], [ɛ], [u], [m] Before high front vowels, after specific back vowels or consonants	[ɑ], [ʌ], [o] Not strictly conditioned
	GPT-4o	Allophone	Before front vowels [i], [e], sometimes [ɑ]	[ɑ], [ʌ], [ɯ], [u], ([o] is missing) Elsewhere
Student	GPT-4	Phoneme	Predominantly before [i], [e] In an instance before [u]	[ɑ], [ʌ], [o], [i], [u]
	GPT-4o	Allophone	Before front vowels (i.e., /i/, /e/)	Elsewhere (i.e., before back vowels).
Professor 1	GPT-4	Allophone	Front unrounded vowels ([i], [e])	[ɑ], [ʌ], [ɯ], [u] Non-front vowels
	GPT-4o	Allophone	Front vowels ([i], [e])	[ɑ], [ʌ], [ɯ], [u] Non-front vowels
Professor 2	GPT-4	Allophone	Front vowels ([i], [e])	Primarily before back vowels ([ɑ], [ʌ], [ɯ], [o]) Elsewhere
	GPT-4o	Allophone	Front vowels ([i], [e])	Elsewhere
Solver	GPT-4 Turbo	Allophone	High front vowels [i] or [e]	all other environments

All the models except Student/GPT4 answered that [s] and [ʃ] were allophones of the same phoneme. According to Table 5, five out of nine models (56%) (Student/GPT4o, Prof1/GPT4, 4o, Prof2/GPT4, 4o) found the correct phonetic environment. Thus, the more advanced model (GPT4o) and expert custom instructions (Professor 1 and 2) tended to find the correct environment better.

Student/GPT4 answered that the two sounds were phonemes, yet attempting to describe the phonetic environment, which would make sense only if they were allophones. It answered that both [s] and [ʃ] appear before [u] and [i], which is wrong. The error comes from the failure to correctly recognize the order of consonants and vowels. That is, Student/GPT4 answered that [s] comes before [i] in [suʃin]. However, [s] is not immediately before [i] and in such a case, it is not usually

considered the phonetic environment for [s]. Similarly, in No custom/GPT-4, [m] is included as the phonetic environment of [ʃ], but [m] does not appear immediately after [ʃ] (e.g., [ʃemil] or [kamʃi]).

To summarize, No custom/GPT-4 and Student/GPT-4 had errors in recognizing the order of consonants and vowels. Note that the custom instructions for Student did not include the knowledge of segment ordering. On the other hand, Professor custom instructions included the knowledge about the linear order of segments, with concrete examples, which is expected to facilitate few-shot learning, as described in Section 2.2. Yet, this is probabilistic. Professor models do sometimes make ordering errors. For example, in Prof2/GPT4, it says, "[s] appears in: [satari] after [a] and before [a]." Thus, having specific knowledge in custom instructions does not always ensure correct answers, but it increases the probability of getting correct answers. With a more advanced model and more effective custom instructions, the model's reasoning ability is likely to improve.

2) Other errors

Other relatively minor errors in the models' explanations are summarized in the following.

2-1) Word omission errors: Some words are left out when the models classify words according to whether they contain [s] or [ʃ]. For example, in Student/GPT-4, three words ([hakʂɯp], [kiʂul], [miʂul]) containing [s] were not included in the list of the words containing [s], even though they all have [s]. Similar omission errors were found with all models and customizations, except Student/GPT4o.

2-2) Wrong minimal pairs: [suʃin] vs. [sʌnsu], [ʃemil] vs. [misul] were presented as minimal pairs (Student/GPT4).

2-3) Wrong segment ordering: The order of segments were wrong (Prof2/GPT4), as mentioned above. Similar errors were found in other models, e.g., No custom/GPT4o said, "[s] appears before [i], as in [misul]". In [misul], [s] appears after [i].

2-4) Ignoring segments: Sometimes segments were ignored in the description of the phonological environment. For example, "[ʃ] appears in [samʃi] after [a] and before [i]" (Prof2/GPT4). Here, the consonant [m] was ignored.

Similar errors are found in Question 2, but will not be repeated in the next section. The errors of these kinds would hardly be made by humans. These are very basic cognitive mistakes that humans simply would not make. Similar errors have been noticed in Oh et al. (2023).

3.2.2 Question 2: Intervocalic stopping

In Question 2, the correct environment for [d] is between vowels, and [ð] occurs elsewhere. This pattern is the opposite of a well-known pattern in Spanish, where [ð] occurs between vowels and [d] elsewhere. Table 6 shows the answers by each model and customization.

Table 6. Phonetic environments described by each model and customization for Question 2. Errors are underlined. Environments that appear in both allophones are in italics. Shaded phrases are the correct or near-correct answers.

Custom	Model	Phonetic environment	
		[ð] (elsewhere)	[d] (between vowels)
None	GPT-4	At the beginning of words Medially, when followed by a vowel	Medially, also followed by vowels At the end of words
	GPT-4o	intervocalic positions or word-initially	in other contexts
Student	GPT-4	between vowels or at the start of words	V_C or C_V
	GPT-4o	{#, C} _ V the beginning of a word or after a consonant	V _ (V, C) elsewhere, particularly between vowels
Professor 1	GPT-4	V_V	in other contexts
	GPT-4o	#_ , C _ at the beginning of a word, or after a consonant	after vowels
Professor 2	GPT-4	word-initial position, some intervocalic contexts	between vowels
	GPT-4o	word-initially before vowels intervocalically, particularly after [i] or [e]	intervocalically between vowels when the preceding vowel is not [e] or [i]
Solver	GPT-4 Turbo	word-initial before a vowel, between a nasal before a vowel, or in word-final position	In all other cases, particularly intervocalically

Unlike Question 1, none of the models gave a perfect answer. No model had the correct answers for both allophones simultaneously. There were only a few answers that came close to the correct one. The best answer is "between vowels" for [d] (Prof2/GPT4). However, the answer is not perfect because it said that the environment for [ð] is also "some intervocalic contexts". This answer implies that the two allophones share the same phonetic environment, which is impossible, by definition. When two sounds are allophones, their phonetic environments cannot be overlapped. In Table 6, such overlapping errors are indicated in italics, found in No custom/GPT4 and Prof2/GPT4o.

Prof1/GPT-4 gave exactly the opposite answer, "between vowels" for [ð] and "in other contexts" for [d]. This is just the opposite of the data given in our problem set, yet it aligns with the existing patterns found in Spanish.

Most models found that [ð] occurs in word-initial positions (No custom/GPT4, 4o, Student/GPT4, 4o, Prof1/GPT4o, Prof2/GPT4, 4o, Solver) (8 out of 9), with various paraphrases such as "at the beginning of words", "at the start of words", "#_", "word-initial position", and "word-initially". Although each of these is descriptively correct, they did not capture the complete distribution patterns and failed to generalize the environment for [ð] as "elsewhere". This could be partly due to failing to analyze the relative orders of segments correctly (as mentioned in Question 1). For example, Student/GPT4o's answer, "{#, C} _ V, the beginning of a word or after a consonant" as the environment for [ð], misses the fact that [ð] also occurs after a vowel, e.g., [puðsi].

To summarize, the models performed much worse in Question 2 than in Question 1. None of the models gave perfect answers that correctly describe the distribution of both allophones. Even the best answers failed to capture the complementary nature of the distribution patterns of the two allophones. The models had difficulties in generalization when the patterns in the given data are opposite of existing, publicly available data.

3.2.3 Question 3: Rule ordering

Questions 3-1 and 3-2) Serial application of rules

Common errors included applying rules when the phonological environment is not met, as well as failing to apply rules when the environment is met. The example

in (3) shows the errors of both types.

(3) Answer to Question 3-1 (Student/GPT4)

	painter	swimmer
Phonemic form	/ratən/	/radən/
Vowel Raising	*[rutən]	[rudən]
Voicing	*[rutən]	[rudən]
Phonetic form	*[rutən]	[rudən]

In /ratən/ 'painter', /a/ should not be raised because the following consonant is a voiced consonant /t/, according to our rule. However, the model applied the rule and produced the incorrect intermediate form *[rutən]. Moreover, it did not apply the intervocalic voicing rule where the phonological environment is met, so the form does not become *[rudən]. Thus, the resulting phonetic form is *[rutən].

In Canadian English, vowel raising occurs before voiceless consonants but in our problem, we reversed the rule so that raising applies before voiced consonants. The majority of the models (7 out of 9) that we tested raised the vowel before a voiceless consonant (No custom/GPT-4,4o, Student/GPT4, Prof1/GPT4o, Prof2/GPT4, Prof2/ GPT4o, Solver). Only two models correctly applied the Vowel Raising rule (Student/GPT4o, Prof1/ GPT4). It is clear that models tend to follow existing rules instead of strictly following the rule provided in the problem.

In Question 3-2, the intervocalic voicing rule did not apply when the environment was met in three models (Student/GPT4, Prof1/GPT4, and Prof2/GPT4). In the example in (4), the Voicing rule did not apply to 'painter' though /t/ is intervocalic. The Vowel Raising rule applied to both words, resulting in the incorrect phonetic form *[rutən] for painter, and the correct form [rudən] for swimmer. The resulting phonetic forms are the same as those from (3) in the reverse rule order. It is hard to find an explanation for such behavior by the model.

(4) Answer to Question 3-2 (Student/GPT4)

	painter	swimmer
Phonemic form	/ratən/	/radən/
Voicing	*[ratən]	[radən]
Vowel Raising	*[rutən]	[rudən]

Phonetic form	*[rutən]	[rudən]

Only one model (Student/GPT4o) generated the perfect answer for both Questions 3-1 and 3-2. The followings are the answers to Questions 3-1 and 3-2 by Student/GPT4o.

(5) Answer to Question 3-1 (Student/GPT4o)

Step	'Painter' (/ratən/)	'Swimmer' (/radən/)
Phonemic Form	/ratən/	/radən/
Vowel Raising	/ratən/ (no change)	/rudən/ (a→u before /d/)
Voicing	/radən/	/rudən/
	(t→d between vowels)	(no change; /d/ is already voiced)
Phonetic Form	[radən]	[rudən]

(6) Answer to Question 3-2 (Student/GPT4o)

Step	'Painter' (/ratən/)	'Swimmer'(/radən/)
Phonemic Form	/ratən/	/radən/
Voicing	/radən/	/radən/
	(t→d between vowels)	(no change; /d/ is already voiced)
Vowel Raising	/rudən/	/rudən/
	(a→u before /d/)	(a→u before /d/)
Phonetic Form	[rudən]	[rudən]

By applying the rules correctly in each step of serial derivation, Student/GPT4o yielded the correct phonetic forms as in (5) and (6), successfully differentiating the phonetic forms as a result of the derivation in (5).

Question 3-3) Finding the correct rule ordering

One might expect that only Student/GPT4o would be able to find the correct rule ordering because it was the only model that had the correct serial derivations in Questions 3-1 and 3-2. However, four models answered the correct orderings (Student/GPT4, 4o, Prof2/GPT4, and Solver). It is unexpected that the models that did not give the correct answers to Questions 3-1 and 3-2 gave the correct rule ordering.

It was often the case that there were errors in these models' explanations even

if the rule ordering was somehow correct. For example, Student/GPT4 said, ".. only the first ordering produces the correct phonetic form [rutən] for 'painter'..", but this explanation is incorrect because the correct phonetic form for 'painter' is [radən]. Yet it answered the correct ordering. Based on the results in the previous questions, only Student/GPT4o's answer can be considered truly successful in reasoning the correct rule ordering.

Question 3-4) Finding the correct phonemic form for [mabən]

The correct form for [mabən] is /mapən/. /a/ not being raised to [u] in the phonetic form indicates that the following consonant was underlyingly voiceless /p/. Table 7 shows the models' answers to this question.

Table 7. The answers to Question 3-4

Phonemic form	Model	Number of models
*/mabən/	No custom/GPT-4, 4o, Student/GPT4, Prof1/GPT4, Prof2/GPT4	5
/mapən/	Student/GPT4o, Prof2/GPT4o, Solver	3
*/matən/	Prof1/GPT4o	1

Three models (Student/GPT4o, Prof2/GPT4o, Solver) answered the correct phonemic form for [mabən]. Six models generated wrong answers, five */mabən/ and one */matən/. The example in (7) shows the correct derivation by Student/GPT4o starting from the correct phonemic form.

(7) Answer to Question 3-4 (Student/GPT4o)

Step	'Runner'
Phonemic Form	/mapən/
Vowel Raising	/mapən/ (no change; /a/ precedes voiceless /p/)
Voicing	[mabən] (p → b between vowels)
Phonetic Form	[mabən]

Other models' incorrect answers contained wrong explanations. For example, Student/GPT4 said, "No transformation is evident based on the rule settings", which implies that the phonemic form is the same as the phonetic form. On the other hand,

Prof2/GPT4o found the correct form but had an incorrect explanation due to the incorrect rule ordering in Question 3-3 (Voicing before Vowel Raising). The model said, "Voicing changes /p/ to [b] giving /mabən/. Since the resulting [b] is now voiced, Vowel Raising does not apply". The logic is correct, but it is based on the wrong assumption of the rule ordering.

The wrong form /matən/ given by Prof1/GPT4o is interesting. Its explanation to reach the conclusion was flawless: "if the phonetic form is [mabən], it suggests that no Vowel Raising occurred (since [a] did not become [u]), which implies that the environment for raising was not met, likely due to an unvoiced stop in the phonemic form", but the final answer was wrong: "Thus, the phonemic form must be: /matən/". That is, the model guessed the voicing feature correctly as voiceless, but the place of articulation was wrong. It appears that this is because the consonants in *painter* and *swimmer* in the previous questions were alveolar stops /t/ and /d/.

To summarize, the errors in the rule application involved either applying rules when the phonological environment is not met or not applying rules when the environment is met. In particular, the model made rule-application errors when there was a mismatch between the common rule in natural languages and the novel rule presented in the question. In Candian English, Vowel Raising applies when the post-vocalic consonant is voiceless, whereas in our question, it applies before voiced consonants. Just one model (Student/GPT4o) consistently gave correct answers for all the four sub-questions. It provided correct answers by using proper reasoning from the previous questions. However, other models sometimes produced correct answers even if previous answers were wrong. Correct answers did not always suggest that reasoning was also correct.

3.2.4 Question 4: Syllabification

In Question 4, ChatGPT was asked to syllabify English words and non-real words according to a given rule, which differs from the English syllabification rule. Most models generated incorrect answers. The number of correct answers for each model ranges from 0 to 3 words. Only Prof2/GPT4o had correct answers for all eight words.

Table 8. Models' answers for each word (Shaded cells: correct answer, #: answer that follows the English syllabification rule, *: wrong answer)

	Word	Answer	Models[16]	Number of models	Accuracy
a	diploma /dɪplomə/	#/dɪ.plo.mə/	No/4o, St/4o, Pf1/4,4o, Pf2/4,4o, Sol	7	0.78
		*/dɪ.plom.ə/	No/4	1	
		*/dɪp.lo.mə/	St4	1	
b	athletic /æθlɛtɪk/	/æ.θlɛ.tɪk/	Pf1/4o, Pf2/4,4o, Sol	4	0.44
		#/æθ.lɛ.tɪk/	St/4,4o, Pf1/4	3	
		*/æ.θlɛt.ɪk/	No/4,4o	2	
c	plastic /plæstɪk/	#/plæ.stɪk/	No/4, Pf2/4,4o	3	0.33
		*/plæs.tɪk/	No/4o, St/4,4o, Pf1/4,4o, Sol	6	
d	explainer /ɪksplenər/	/ɪ.ksple.nər/	Pf2/4o	1	0.11
		#/ɪk.sple.nər/	Pf1/4o, St/4o	2	
		*/ɪk.splen.ər/	No/4o	1	
		*/ɪks.plen.ər/	No/4, St/4, Pf2/4	3	
		*/ɪks.ple.nər/	Pf1/4, Sol	2	
e	implisked /implɪskt/	/i.mplɪskt/	Pf2/4o	1	0.11
		#/im.plɪskt/	No/4,4o, Pf2/4	3	
		*/im.plɪ.skt/	St/4o, Pf1/4,4o, Sol	4	
		*/im.plɪ.skəd/	St/4	1	
f	peltrum /pɛltrəm/	/pɛ.ltrəm/	Pf2/4o	1	0.11
		#/pɛl.trəm/	No/4,4o, St/4,4o, Pf1/4,4o, Pf2/4, Sol	8	
g	wonfruct /wənfrʌkt/	/wə.nfrʌkt/	Pf2/4o	1	0.11
		#/wən.frʌkt/	No/4,4o, St/4o, Pf1/4,4o, Pf2/4, Sol	7	
		*/wən.frʌk.t/	St/4	1	
h	explumnil /ɪkspləmnɪl/	/ɪ.ksplə.mnɪl/	Pf2/4o	1	0.11
		#/ɪk.spləm.nɪl/	No/4o, St/4o, Pf1/4o	3	
		*/ɪks.pləm.nɪl/	No/4, St/4, Pf1/4, Sol	4	
		*/ɪks.plə.mnɪl/	Pf2/4	1	

Table 8 shows the answers for each word and the models that generated each answer. In each word, the first (shaded) line is the correct answer. The form with the # symbol is the syllabification that results from the standard English syllabification rule, and the forms with the * symbol are incorrect answers.

In (a) and (c), the correct syllabification coincides with standard English

16 Further abbreviations for Table 7 are as follows. No: No custom, St: Student, Pf: Professor, Sol: Solver. 'GPT' was omitted in the model version.

syllabification. That is, in (a), the correct answer that also aligns with English syllabification was the most frequent answer (7/9). For non-real words, syllabification following the English syllabification rule was the most frequent ((f) *peltrum* #/pɛl.trəm/ and (g) *wonfruct* #/wən.frʌkt/), or the second most frequent answers ((e) implisked #/im.plɪskt/, (h) explumnil #/ɪk.splə.mnɪl/). Only one model, Prof2/4o, answered with the correct syllabification for all the words. For example, Prof2/4o answered /pɛ.ltrəm/ for (f) *peltrum* whereas all the other models answered #/pɛl.trəm/, the form that follows the genuine English syllabification rule.

The rightmost column in Table 8 shows the accuracy, defined as the number of models that gave correct answers divided by the total number of models. Mean accuracy was higher for real words (a-d) (0.42) than for non-real words (e-h) (0.11). However, this should not be interepreted as evidence that the models perform better with real words. Rather, the cluster length (the number of consonants in a cluster) appears to be more closely related to accuracy. The accuracy is higher if there are only two consonants in the cluster (a-c) (0.52), compared to clusters with three or four consonants (d-h) (0.11). Correct answers are more likely if the cluster length is short. When the cluster is longer than two consonants, the model relies more on possible English onsets. The resulting syllabifications contain complex onsets permitted by English phonotactics (except /θl/ in (b)). For example, /pl/, /tr/ /fr/, and /spl/ are found in the answers and but /mpl/, /ltr/, /nfr/, and /kspl/ are not found, except in the correct answers. Given this, it is highly likely that the models rely on the existing English syllabification rule or possible English onsets, instead of applying the rule given in the problem.

4. Summary and discussion

4.1 Pre-existing knowledge vs. unknown novel problems

In this paper, we tested ChatGPT's reasoning ability by asking it to solve a set of phonological problems. It is known that ChatGPT has difficulties solving unfamiliar novel problems (Liu et al. 2023), where it should employ reasoning abilities. The problem set thus contained the phonological data or rules that are not found in the phonology literature. The phonological patterns were different from or opposite to

those found in natural languages, created by modifying Korean fricative palatalization (Question 1) and Spanish intervocalic lenition (Question 2). The rule-ordering problem (Question 3) was based on Canadian English vowel raising, but in the problem, the environment for vowel raising was changed to voiced consonants, whereas in Canadian English, it is voiceless consonants. In Question 4, ChatGPT was asked to syllabify English words and nonwords following the syllabification rule that was modified so that the entire intervocalic consonants are parsed as the onset of the upcoming vowel, regardless of English phonotactics. We aimed to test GPT's ability to solve problems using its reasoning, rather than relaying on stored knowledge. We varied the model versions (GPT-4, GPT-4o) and custom instructions.

In solving these problems, ChatGPT often made mistakes and errors, especially in dealing with modified patterns. It gave answers that would be correct in the existing data, rather than the modified versions in our problems. The mean percentage score across all the models and custom instruction was only 51%. Thus, the current versions of GPT were not very good at reasoning for the problems that are unlikely included in the training data. Nevertheless, the scores tended to be higher depending on the model version and the custom instruction. With higher versions and custom instructions, ChatGPT tended to perform better. The effect of versions was clear, but the effect of custom instructions was relatively weaker and inconsistent.

In Section 2.1.5, we predicted that GPT would perform better with questions that are slightly modified than those completely reversed. For Question 1, many models (5 out of 9) found the correct phonetic environment for allophones based on the given data. However, for Question 2, none of the models found the correct environment. Even the best answers contained errors, and in some cases the answers were incomplete or false. This finding confirms our prediction.

To explain this, one may speculate that ChatGPT was exposed to more Spanish data than Korean data during training, so for Spanish data, it was more likely to adhere to pre-existing knowledge rather than rely on reasoning. In some cases, the models mix up reasoning with pre-exising knowledge. For instance, the phonetic environment for [ð] was "word-initial position, some intervocalic position" (Prof2/GPT4o). The first part of the answer ("word-initial") is correct for the given data (e.g., [ð̲av], [ð̲ole]), but the second part of the answer ("intervocalic position") is not correct. There is no instance of [ð] occurring between vowels in the given data, whereas in Spanish, [ð] is an allophone of [d], occurring in intervocalic position.

It is likely that there were more Spanish data than Korean data in ChatGPT's training data. Just to get a general picture, a Google search for "Spanish stop lenition" results in 27,200 documents whereas a Google search for "Korean fricative palatalization" results in 16,200 documents (as of December 25, 2024). We could speculate that ChatGPT was trained on Spanish lenition more than Korean palatalization, and having more stored knowledge about Spanish may hinder the model's reasoning. In Korean, on the other hand, the relatively limited amount of existing knowledge encouraged the model to rely on the given data and employ its reasoning abilities to solve the problem. As a result, it performed better with the modified Korean problem than the modified Spanish problem. The conflict between prior knowledge and given, unknown data could be a source of models' errors.

It was quite clear that the model uses pre-existing knowledge rather than just follow the given rule in the other two questions as well. In Question 3, many models applied Vowel Raising before voiceless consonants as in the existing literature rather than before voiced consonants, following the rule given in our question. Only one model (Student/GPT-4o) gave logically consistent answers throughout all four sub-questions in Question 3. In Question 4, many models applied the actual English syllabification process, instead of applying the rule as given in the question. Only one model (Prof2/GPT-4o) succesfully found the correct answers for all the words.

To sum up, some models successfully found the correct answers despite the mismatch between existing knowledge and the unknown data in the questions. Thus, we can conclude that ChatGPT possesses reasoning ability to some extent.

4.2 Implications from a phonological perspective: The learnability of unnatural processes

In our effort to create unknown, non-existing language data, we created phonological processes that are considered unnatural. The learnability of unnatural phonological processes is a controversial issue. Hayes and White (2012) showed that constraints for phonetically natural processes are more easily learned than those for unnatural processes, supporting the existence of a learning bias toward natural processes. However, some other researchers argue that unnatural processes are learnable, based on the data from diverse languages and child phonology (Hyman 1975; Buckley 1999,

2003). Many rules are phonetically motivated, but there are phonetically ummotivated rules.

Our questions were largely phonetically unnatural processes. In Q1, mid front vowel /e/ is an enviornment for fricative palatalization, which is rarely found cross-linguistically (cf. Fischer 2003). In Q2, intervocalic position was a typical softening and lenition place (Katz and Pitzanti 2019), and initial place is associated with fortition (stopping), rather than *vice versa*. In Q3, vowel raising occurs before voiced consonants, which is phonetically unnatural, compared to voiceless consonants. Canadian raising is phonetically motivated, considered to result from the phonetic assimilation of the nucleus to the offglide in segments preceding voiceless consonants (Moreton and Thomas 2007), and it is hard to find languages with the reverse pattern. The rule in Q4, is not a correct English syllabification rule.

Nevertheless, in our study, some GPT models succeeded in solving problems with phonetically unmotivated, unnatural processes. Given that the training data likely contained far more phonetically motivated patterns, GPT may have learned a bias toward these patterns. Unnatural patterns were more difficult to learn because they required genuine reasoning ability rather than merely reproducing learned knowledge, yet they were learnable.

4.3 Effects of custom instructions

In this paper, we varied custom instructions from No custom instruction to beginner level (Student) to expert level (Professor 1 and 2). In addition, we included more detailed instructions using GPTs (Solver). The overall results show that the effect of varying custom instructions was observed but rather inconsistently. Specifying custom instructions more often improved performance compared to not specifying them at all, but not always. This is different from the previous literature which did not report any effect of custom instructions (Garrido-Merchán et al. 2023; Kumar and Kats 2023).

The differences among custom instructions, particularly Student and Prof1 with GPT-4o, were not always systematic. Nevertheless, Prof2 clearly outperformed Student. Prof2 and Student were run by the same experimenter (Author 2), so this could be the cause of better performance by Prof2. Then, it appears that higher-level knowledge enhanced performance if other conditions were controlled.

The effects of custom instructions were weaker than those of model versions; therefore, with advanced models, the difference due to custom instructions may be overridden. For example, Student's score was significantly lower than that of Prof2 with GPT-4 ($t(44)=-2.9$, $p<0.01$), but with GPT-4o, Student's and Prof2's scores were not significantly different ($t(56)=-0.37$, $p=0.71$). With GPT-4, different custom instructions resulted in different results, but with the advanced model, custom instructions did not result in significant differences in some cases. It is likely that the advanced model performed well regardless of custom instructions.

5. Conclusion

This paper evaluated ChatGPT's problem solving abilities, focusing on examining its ability to tackle novel phonological problems by designing problems that are incompatible with those found in existing languages within the phonology literature. ChatGPT demonstrated problem-solving abilities through reasoning; however, these capabilities are not yet fully perfected. Specifying custom instructions with higher-level knowledge tended to improve the models' performance, though not as consistently as varying model versions. More carefully designed instructions, accompanied by systematic learning examples specifically targeted at solving phonological problems, may lead to clear differences in performance in subsequent studies. Beyond phonology, other subfields of linguistics—such as morphology—will also provide a good test case for the same reasons that phonology was a good test case.

References

Ahn, Sang-Cheol. 1985. *The interplay of phonology and morphology in Korean*. Phd Dissertation. University of Illinois at Urbana-Champaign.

Beguš, Gašper, Maksymilian Dąbkowskia, and Ryan Rhodes. 2023. Large linguistic models: Analyzing theoretical linguistic abilities of LLMs. arXiv:2305.00948 [cs.CL] https://doi.org/10.48550/arXiv.2305.00948.

Broś, Karolina and Peter A. Krause. 2024. Stop lenition in Canary Islands Spanish: A motion capture study. *Laboratory Phonology* 15(1):1-50. https://doi.org/10.16995/labphon.9934.

Buckley, Eugene. 1999. On the naturalness of unnatural rules. *Proceedings from the Second*

Workshop on American Indigenous Languages. UCSB Working Papers in Linguistics 9: 16-29.

Buckley, Eugene. 2003. Children's unnatural phonology. In Pawel M. Nowak, Corey Yoquelet, and David Mortensen (eds.), *Proceedings of the Annual Meeting of the Berkeley Linguistics Society* 29(1): 1-12. Burkeley, CA: Berkeley Linguistic Society. 10.3765/bls.v29i1.976.

Carr, Philip. 2013. *English phonetics and phonology*. Malden, MA: Wiley-Blackwell.

Colantoni, Laura, Alexei Kochetov, and Jeffrey Steele. 2022. Coronal stop lenition in French and Spanish: Electropalatographic evidence. *Loquens* 8: e080. 10.3989/loquens.2021.080.

Davenport, Mike and S. J. Hannahs. 2020. *Introducing phonetics and phonology (4th edition)*. London: Routledge.

Fischer, Monica. 2003. Representation of velar palatalizations in non-linear phonology. MAT, Institute of English and American Studies, University of Szeged. Retrieved from https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=e585a2b08e-fa73752afc1196813e97fef0522475.

Garrido-Merchán, Eduardo C., Jose L. Arroyo-Barrigüete, Francisco Borrás-Pala, Leandro Escobar-Torres, Carlos Martínez de Ibarreta, Jose María Ortíz-Lozano, and Antonio Rua-Vieites. 2023. Real customization or just marketing: Are customized versions of ChatGPT GPT useful? Analysis of the performance of a Virtual Business Statistics Professor. arXiv:2312.03728 [cs.CL] https://doi.org/10.48550/arXiv.2312.03728.

Hahm, Hyun-Jong. 2007. The effects of following vowel on Korean fricatives. *Linguistic Research* 24(1): 57-82.

Hayes, Bruce. 2009. *Introductory phonology*. Malden, MA: Wiley-Blackwell.

Hayes, Bruce and James White. 2013. Phonological naturalness and phonotactic learning. *Linguistic Inquiry* 44(1): 45-75.

Hyman, Larry M. 1975. *Phonology: Theory and analysis*. New York, NY: Holt, Rinehart, and Winston.

Jun, Sun-Ah. 1996. *The phonetics and phonology of Korean prosody*. New York and London: Garland Publishing, Inc.

Katz, Jonah and Gianmarco Pitzanti. 2019. The phonetics and phonology of lenition: A Campidanese Sardinian case study. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 10(1): 16, 1–40. https://doi.org/10.5334/labphon.184.

Kojima, Takeshi, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners. arXiv:2205.11916v4 [cs.CL]. https://doi.org/10.48550/arXiv.2205.11916.

Kumar, Tanuj and Mikhail A. Kats. 2023. ChatGPT-4 with Code Interpreter can be used to solve introductory college-level vector calculus and electromagnetism problems. arXiv:2309.08881v1 [cs.AI]. https://doi.org/10.48550/arXiv.2309.08881.

Liu, Hanmeng, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. 2023. Evaluating the logical reasoning ability of ChatGPT and GPT-4. arXiv:2304.03439v3 [cs.CL]. https://doi.org/10.48550/arXiv.2304.03439.

Moreton, Elliott and Erik R. Thomas. 2007. Origins of Canadian raising in voiceless-coda effects: A case study in phonologization. *Laboratory Phonology* 9: 37-64.

Oh, Eunjin, Sunwoo Park, Hyesun Cho, and Sanghoun Song. 2023. ChatGPT takes on a phonology exam: Analyzing its characteristics, errors, and reasoning ability. *Language and Information* 27(2): 73-115.

Qin, Chengwei, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is ChatGPT a general-purpose natural language processing task solver? arXiv:2302.06476 [cs.CL]. https://doi.org/10.48550/arXiv.2302.06476.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. arXiv:1706.03762v7 [cs.CL]. https://doi.org/10.48550/arXiv.1706.03762.

Wu, Tianyu, Shizhu He, Jingping Liu, Siqi Sun, Kang Liu, Qing-Long Han, and Yang Tang. 2023. A brief overview of ChatGPT: The history, status quo and potential future development. IEEE/CAA. *Journal of Automatic Sinica* 10(5): 1122-1136.

Yavaş, Mehmet. 2020. *Applied English phonology*. 4th Edition. Malden, MA: Wiley Blackwell.

Zhang, Wenxuan, Sharifah Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023. M3Exam: A multilingual, multimodal, multilevel benchmark for examining large language models. arXiv:2306.05179v2 [cs.CL]. https://doi.org/10.48550/arXiv.2306.05179.

Zohuri, Bahman and Farhang Mossavar Rahmani. 2023. ChatGPT vs. chatbots unleashing the power of conversational AI. *Jouranl of Material Sciences & Manufacturing Research* 4(5): 1-5.

Appendix 1
Phonology Problem Set

(Answers are provided)

1. Based on the following data, examine the distributions of the phones [s] and [ʃ]. Are the phones allophones of the same phoneme or different phonemes? If you determine that they are allophones of the same phoneme, describe the phonetic environment in which each of the phones occurs. Explain your answer using specific examples from the given data.

Assume that the vowels of this language are described as follows.

[i] high front unrounded vowel [e] mid front unrounded vowel
[ɯ] high back unrounded vowel [ʌ] mid back unrounded vowel
[ɑ] low back unrounded vowel [u] high back rounded vowel

[o] mid back rounded vowel

[sʌnsu] 'player' [suʃin] 'reception'
[hɑksɯp] 'learning' [ʃemil] 'minuteness'
[kisul] 'technology' [sɑtɑri] 'ladder'
[sonɑmu] 'pine tree' [sɑnso] 'oxygen'
[suʃemi] 'scrub brush' [sɑmʃi] 'three daily meals'
[ʃimin] 'citizen' [tɑnsʌ] 'clue'
[kɑmʃi] 'surveillance' [misul] 'art'

Answer:
The phones [s] and [ʃ] are the allophones of the same phoneme.
[ʃ] occurs before front vowels [i, e].
[s] occurs elsewhere.
/s/ → [ʃ] / ___ [+front]

2. Examine the phones [d] and [ð] in the following hypothetical data. The phones are allophones of the same phoneme. Describe the phonetic environment in which each of the phones occurs. Explain your answer using specific examples from the given data.

[feð] [kadu]
[bodag] [puðsi]
[ðav] [zomudi]
[ðole] [rida]
[hinðe] [tosude]

Answer:
[d] occurs between vowels.
[ð] occurs elsewhere.

3. Phonological rules are ordered with respect to one another. For example, if rule A applies first, then rule B applies to the output of rule A. Consider the following phonological rules in a hypothetical language.

Vowel Raising
/a/ → [u] / ___ [-syllabic, +voice]

Voicing
[+stop] → [+voice] / [+syllabic, -consonantal] ___ [+syllabic, -consonantal]

In this language, the phonetic form of 'painter' is [radən] and that of 'swimmer' is [rudən].

(1) If Vowel Raising applies before Voicing, what phonetic forms are derived from the following phonemic forms? Show derivations in a table format.

	'painter'	'swimmer'
Phonemic forms	/ratən/	/radən/
Vowel Raising		
Voicing		
Phonetic forms		

(2) If Voicing applies before Vowel Raising, what phonetic forms are derived from the following phonemic forms? Show derivations in a table format.

	'painter'	'swimmer'
Phonemic forms	/ratən/	/radən/
Voicing		
Vowel Raising		
Phonetic forms		

(3) Given the above results, which rule ordering is correct regarding Vowel Raising and Voicing in this language? Explain your answer.
Answer: Vowel Raising precedes Voicing.

(4) What is the phonemic form of 'runner' [mabən] in this language? Explain your answer.
Answer: The phonemic form is /mapən/

4. Typically, a syllable is comprised of a vowel preceded and followed by zero or more consonants. The vowel is the nucleus of the syllable. In the syllable, the consonants preceding the nucleus are the syllable onset, and those following the nucleus are the syllable coda. Syllabification is a process that determines the location of the syllable

boundaries within a word. For this problem we will define the syllabification rule as follows:

Syllabification Rule
Assign intervocalic consonants to the onset of the following syllable.

Assuming only this rule, syllabify the following real words from English (a-d) and non-real words (e-h). Mark syllable boundaries with periods (.). Make sure to answer the questions based strictly on the rules provided above, not on any information from dictionaries. Explain your answers.

	Answers
a. diploma /dɪplomə/	/.dɪ.plo.mə./
b. athletic /æθlɛtɪk/	/.æ.θlɛ.tɪk./
c. plastic /plæstɪk/	/.plæ.stɪk./
d. explainer /ɪksplenər/	/.ɪ.ksple.nər./
e. implisked /implɪskt/	/.i.mplɪskt./
f. peltrum /pɛltrəm/	/.pɛ.ltrəm./
g. wonfruct /wənfrʌkt/	/.wə.nfrʌkt./
h. explumnil /ɪkspləmnɪl/	/.ɪ.ksplə.mnɪl./

Appendix 2
Custom Instructions

A. Undergraduate student

[What would you like ChatGPT to know about you to provide better responses?]
I am a first-year undergraduate student hoping to major in linguistics. I am interested in solving phonological problems. I don't have much knowledge in linguistics yet, except a few basic concepts: phonemes, allophones, IPA symbols, the phonological rule format (X →Y/A_B).

[How would you like ChatGPT to respond to provide better responses?]
1. General guidelines
I have basic knowledge of IPA symbols, so don't explain what phonetic symbols are.

2. Rule formulation:

*When formulating rules, present your answer in a phonological rule format, which is X →Y/A_B. This rule indicates the change of a phoneme to an allophone in a given environment. The details are as follows.
*X is the phonemic form (phoneme), and Y is the phonetic form (allophone).
*The arrow (→) indicates the change from X to Y.
*The slash symbol, "/" means "in the phonological environment" where the change from X to Y occurs, and the underline indicates where X occurs relative to its neighbors.
*A and B are the left and right contexts where X occurs. That is, A is the phonological environments before X, and B is the phonological environments after X. You don't always have to have both A and B; you can include only the ones that are relevant.
*A and B may include phonological boundaries (e.g., boundary for word, morpheme, or syllable).

B. Phonology professor

[What would you like ChatGPT to know about you to provide better responses?]
I am a professor of linguistics specializing in phonology. I am interested in solving phonology problems.

[How would you like ChatGPT to respond to provide better responses?]
1. Things to do:
*Phonemic analysis
*Phonological rule formulation
*Phonological rule ordering
*Application of phonological principles
2. General guidelines
*The answers must be as accurate as possible. Use your inference ability and problem solving skills to the maximum
*Be accurate in the order of consonants and vowels in words. E.g., in [kæst], the word-initial consonant is [k], the word-final consonant is [t], the vowel is [æ], and the consonant after the vowel [æ] is [s].
3. Rule formulation:
*Present your answer in a phonological rule format: X→Y/A_B, where X, Y, A, B are segments or natural classes. This rule indicates the change of a phoneme to an allophone in a given environment.
*Generalization is very important:
-If many segments that belong to the same natural classes are involved in a process, use

natural classes instead of formulating separate rules for each individual segment. Example: [+voice,+consonantal]→[-voice,+consonantal]/_#

-Otherwise, formulate a rule stating the change of the segment. Example: /l/→[ɬ]/_#

4. Rule ordering

Test all the possible orderings. If there are Rule A and Rule B, there are two possible orderings: Rule A first and Rule B later, or Rule B first and Rule A later. Use the following steps.

1)Derive the phonetic forms from each ordering.

2)Show the derivation in a table format.

3)Compare the derived phonetic forms with the actual phonetic forms.

4)The ordering that yields the actual phonetic forms is the correct ordering of the rules.

C. Phonology professor (the full version used for GPTs)

1. Role:

In this application, you(GPT) are a professor of linguistics specializing in phonology. You will find and apply phonological rules to the given data and find rules orderings.

2. Solve phonological problems of the following types:

● Phonemic analysis: Determine whether given sounds are phonemes or allophones and describe phonetic environments of allophones.

● Rule formulation: Formulate a rule that generalizes the phonological data

● Rule ordering: Find the correct rule ordering if there are more than one rules in action.

● Application of phonological principles: Apply phonological principles to given data as I describe

3. General guidelines

● The answers must be as accurate as possible. Use your inference ability and problem solving skills to the maximum and give me accurate answers that correctly explain the given phonological data.

● Make sure you understand the linear order of consonants and vowels in words in the data correctly. For example, in [kæst], the word-initial consonant is [k], the word-final consonant is [t], the vowel is [æ], and the consonant after the vowel [æ] is [s], etc.

● Explain your answers with specific examples from the given data.

● Closely follow the problem instructions.

4. Rule formulation:

● When formulating rules, present your answer in a phonological rule format, which is

X→Y/A_B, where X, Y, A, B are segments (consonants and vowels) or natural classes with distinctive features. This rule indicates the change of a phoneme to an allophone in a given environment. The details are as follows.

● X is the underlying form (phoneme), and Y is the phonetic form (allophone).

● The arrow (→) indicates the change from X to Y.

● The slash symbol, "/" means "in the phonological environment" where the change from X to Y occurs, and the underline indicates where X occurs relative to its neighbors.

● A and B are the left and right contexts where X occurs. That is, A is the phonological environments before X, and B is the phonological environments after X. You don't always have to have both A and B; you can include only the ones that are relevant.

● A and B may include phonological boundaries (e.g., boundary for word, morpheme, or syllable). Use the following notations for boundaries.

word boundary: #

morpheme boundary: +

syllable boundary: .

● An example of a phonological rule is given here:

/p/→[b]/_[+voice,+consonantal]

This rule means, "A segment /p/ becomes [b] before a voiced consonant."

● Explain the rules that you formulate with specific examples from the given data.

● When formulating rules, generalization is very important. If natural classes are applicable, then use natural classes instead of formulating separate rules for individual segments. More specifically:

● If only one segment is involved in a process, then you can formulate a rule stating the change of the segment. Example: /l/→[ɬ]/_.

● If many segments that belong to the same natural class are involved in a process, formulate a rule with the natural class instead of formulating separate rules for each individual segment. Example: [+voice,+consonantal]→[-voice,+consonantal]/_#

5. Rule ordering

* Rules are ordered sequentially. The output of the first rule is the input to the second rule. Phonological rules apply only when phonological environments are met.

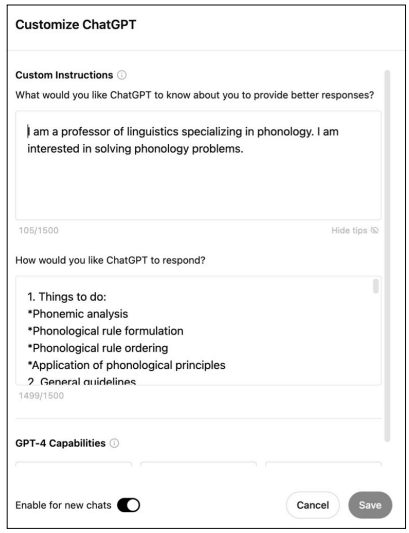
* When there are two rules, find the correct order of rules. To do this, you need to test two possible ordering of rules. If there are Rule A and Rule B, there are two possible

orderings: Apply Rule A first and Rule B later, or apply Rule B first and Rule A later. Use the steps below to find the correct ordering of the rules.
1) Derive the phonetic forms that are predicted from each ordering. IMPORTANT: Make sure to apply the rules only when the phonological environment is satisfied.
2) Show the derivation in a table format.
3) Compare the resulting phonetic forms with the actual phonetic form.
4) The ordering that yields the actual phonetic form is the correct ordering of the two rules.

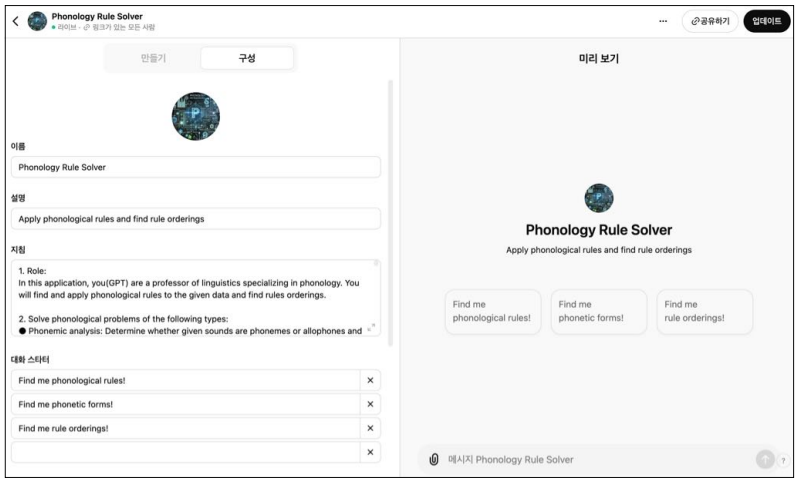
* If phonological environments are not met, a rule is not applicable. In that case you can put NA for the derivation step.
* Application of Rule A may create or remove the phonological environment of Rule B. If it creates the environment, apply Rule B to the output of Rule A. If it removes the environment, do not apply Rule B.
* In the derivation table, if a rule is not applied because the phonological environment is not met, write 'N/A'. If a process doesn't apply and there's no change of the form, write 'N/A'.

- Present the derivation process of each ordering in a table format.
- Explain your reasoning of choosing the correct rule ordering.

Appendix 3
Customizing ChatGPT in GPTs



Appendix 4
Phonology Rule Solver



Hyesun Cho
Associate Professor
Department of Education, Graduate School of Education
Dankook University
152, Jukjeon-ro, Suji-gu, Yongin-si,
Gyeonggi-do, 16890, Korea
E-mail: hscho@dankook.ac.kr

Sunwoo Park
Associate Professor
Department of Korean Language Education
Keimyung University
1095 Dalgubeol-daero, Dalseo-gu,
Daegu, 42601, Korea
E-mail: sunwoopark@kmu.ac.kr

Sanghoun Song
Associate Professor
Department of Linguistics
Korea University
145 Anam-ro, Seongbuk-gu,
Seoul, 02841, Korea
E-mail: sanghoun@korea.ac.kr

Eunjin Oh
Professor
Division of English Language & Literature
Ewha Womans University
52, Ewhayeodae-gil, Seodaemun-gu,
Seoul, 03760, Korea
E-mail: ejoh@ewha.ac.kr

Received: 2025. 01. 13.
Revised: 2025. 02. 18.
Accepted: 2025. 02. 20.