# P-CoT: A Pedagogically-motivated Participatory Chain-of-Thought Prompting for Phonological Reasoning in LLMs

**Dongjun Jang** and **Youngchae Ahn** and **Hyopil Shin**
Department of Linguistics
Seoul National University
{qwer4107, estelle1026, hpshin}@snu.ac.kr

## Abstract

This study explores the potential of phonological reasoning within text-based large language models (LLMs). Utilizing the PhonologyBench benchmark, we assess tasks like rhyme word generation, g2p conversion, and syllable counting. Our evaluations across 12 LLMs reveal that while few-shot learning offers inconsistent gains, the introduction of a novel Pedagogically-motivated Participatory Chain-of-Thought (P-CoT) prompt, which is anchored in educational theories like scaffolding and discovery learning, consistently enhances performance. This method leverages structured guidance to activate latent phonological abilities, achieving up to 52% improvement and even surpassing human baselines in certain tasks. Future work could aim to optimize P-CoT prompts for specific models or explore their application across different linguistic domains.

## 1 Introduction

The performance of transformer decoder-based large language models (LLMs), such as ChatGPT (OpenAI, 2023), has advanced rapidly in recent years (Brown et al., 2020; Taylor et al., 2022; Chowdhery et al., 2022; Shanahan, 2022; Touvron et al., 2023; Bubeck et al., 2023; Hu et al., 2024; Guo et al., 2025). These models excel at a wide range of tasks through in-context learning (Dong et al., 2022) and have been extensively evaluated for their capabilities in syntax, semantics, morphology, and discourse (Liang et al., 2022; Atox and Clark, 2024; Titus, 2024). In contrast, phonology has received comparatively little attention in this area, primarily because it has traditionally been studied using multi-modal (Zhang et al., 2023a; Yu et al., 2024) or speech-based models (Baevski et al., 2020; Rubenstein et al., 2023; Radford et al., 2023). This is despite its significant role in generating natural prosody (Zhang et al., 2023b) and accurately modeling dialectal variations (Dutoit, 1997).

Recent work such as Suvarna et al. (2024) shows that while text-based LLMs encode latent phonological patterns from language data, they often struggle with explicit tasks, resulting in a notable gap compared to human performance. Therefore, enhancing phonological reasoning in text-based LLMs emerges as a timely and important research direction.

Our study investigates the phonological competence of modern text-based language models using PhonologyBench[1] (Suvarna et al., 2024), a benchmark that evaluates three key tasks: rhyme word generation, grapheme-to-phoneme (g2p) conversion, and syllable counting. We first evaluated 12 state-of-the-art models and examined the impact of few-shot learning (Brown et al., 2020) on their performance. Our findings reveal that while few-shot prompting sometimes results in improvements over baseline, these enhancements are inconsistent and model-dependent, with performance occasionally falling below baseline levels. This underscores the need for more robust techniques to enhance phonological reasoning in text-based LLMs.

To address these challenges, we propose a novel approach that integrates role-playing pedagogy within a Chain-of-Thought (CoT) prompting framework (Wei et al., 2022). We hypothesize that although LLMs inherently encode meta-knowledge in phonology, they lack effective strategies to leverage this information. By integrating two distinct pedagogical techniques—scaffolding (Wood et al., 1976) and discovery learning (Bruner, 1961)—and building on prior studies that have effectively combined these methods (Jatisunda et al., 2020; Alfieri et al., 2011; Schofield et al., 2021), our Pedagogically-motivated Participatory Chain-of-Thought (P-CoT) prompt consistently improves performance across all 12 models on the phonological tasks. These results persist even though we use

[1] https://github.com/asuvarna31/llm_phonology



*Findings of the Association for Computational Linguistics: ACL 2025*, pages 21958–21979
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

| Task | Example Input | Example Output |
|---|---|---|
| **Rhyme Word Generation** | Give 5 words that rhyme with '*education*'. | *circulation, occupation, reputation, population, reservation* |
| **G2P Conversion** | Convert the given grapheme *'basement'* into phoneme according to American English in IPA. | /beɪsmənt/ |
| **Syllable Counting** | Count the number of syllables in the sentence: "*To top it all off, I miss my stunner.*" | *10* |

Table 1: Illustrative Example Tasks from PhonologyBench (Suvarna et al., 2024): Rhyme Word Generation, G2P Conversion, and Syllable Counting

the same examples from standard few-shot learning, indicating that our approach, rather than the data itself drives the enhanced performance. Notably, our method achieves performance gains of up to 52% and even surpasses the human baseline in the specific task. Furthermore, we conducted detailed, task-specific analyses and found that the observed improvements represent true learning advances. These findings underscore the effectiveness of our P-CoT prompts, demonstrating their superiority over traditional few-shot methods in unlocking the latent phonological capabilities of LLMs.

## 2 Related Work

Recent work has begun to explore the phonological skills of text-based LLMs. Suvarna et al. (2024) introduced PhonologyBench—a benchmark for phonological tasks such as g2p conversion, syllable counting, and rhyme word generation (Table 1). Qharabagh et al. (2024) enhanced g2p conversion performance with a two-stage prompting strategy. Additionally, Xue et al. (2021) proposed DeepRapper, a Transformer-based model that generates rap lyrics, while Loakman et al. (2023) introduced the TwistList dataset and benchmarks for tongue twister generation. Together, these studies reveal that research in phonological capabilities is largely task-specific, rather than focusing on a comprehensive evaluation of overall phonological performance.

Alongside these advances, the study of personas in LLMs has progressed significantly, with Role-Playing Language Agents (RPLAs) enhancing human-like performance. Chen et al. (2024) offer a comprehensive categorization of personas, while Tseng et al. (2024) present a taxonomy of LLM role-playing and personalization. Wang et al. (2023) propose Solo Performance Prompting (SPP), which leverages multiple personas to boost model performance. In parallel, researchers have also turned to educational methods to enhance prompt design. For instance, Chang (2023) adopt the Socratic method to guide LLM reasoning through structured questioning, while Jiang et al. (2024) introduce PedCoT, which uses Bloom's taxonomy to guide self-correction through a two-stage prompting process involving multiple model calls.

Building on these insights, our approach integrates both persona-based role-play and educational principles to design effective prompts. Specifically, we draw from discovery learning (Bruner, 1961; Anthony, 1973), where learners explore core principles through inquiry-based methods (Van Joolingen et al., 2005). Although discovery learning fosters self-directed exploration, studies (Moreno, 2004; Tuovinen and Sweller, 1999; Hardiman et al., 1986; Brown, 1994; Kirschner et al., 2006) indicate that structured guidance is crucial to prevent confusion and misconceptions. Given that current LLMs exhibit only rudimentary phonological competence relative to human baselines (Suvarna et al., 2024), we integrate scaffolding and worked examples into our prompts to provide effective, guided instruction.

Accordingly, we combine discovery learning with the scaffolding technique introduced by Wood et al. (1976) to design our prompts. They define scaffolding as a process that enables a novice to tackle tasks beyond their unaided efforts. Later, Bruner (1985) and Cazden (1979) relate this notion to Vygotsky's Zone of Proximal Development (ZPD)—the space between what learners can do alone and what they can accomplish with guidance.

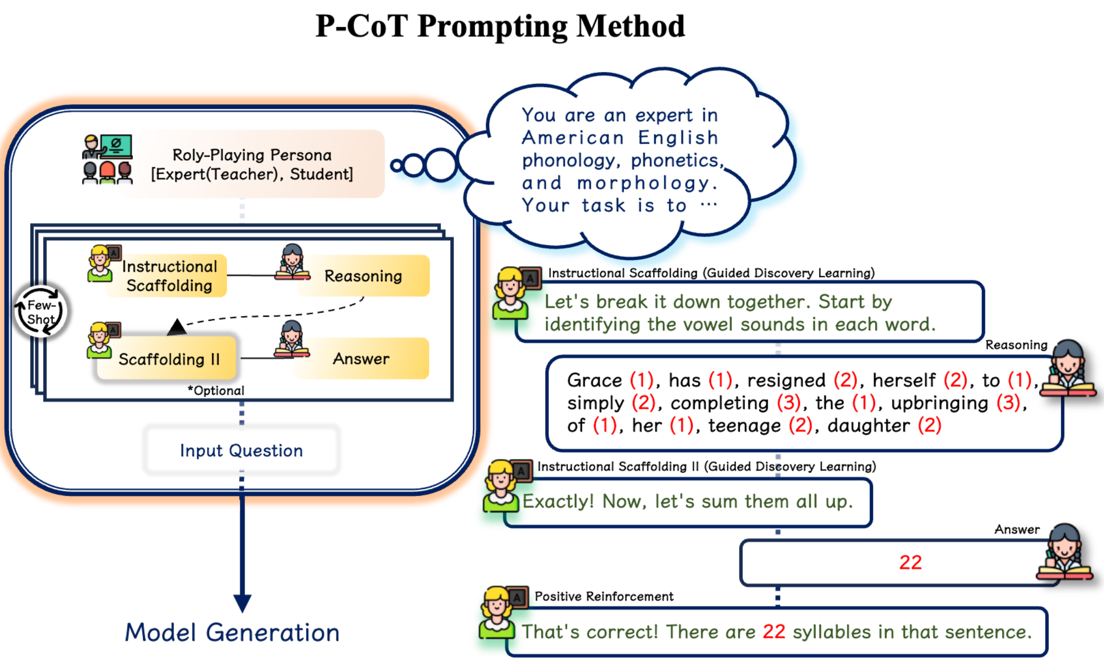



Figure 1: P-CoT Prompting Method

In our approach, scaffolding is used to help the model traverse this zone, offering temporary support that enables it to engage with phonological tasks beyond its current level of competence.

This integration of discovery learning and scaffolding has been shown to be effective in educational contexts. For instance, Ewing McMahon (2000) and Collins et al. (1991) provide real-world examples of scaffolding in educational settings. Collins et al. (1991) particularly outlines a six-phase instructional model centered on modeling, coaching, and scaffolding. Similarly, Jatisunda et al. (2020) found that discovery learning with scaffolding improves creative thinking and self-efficacy, outperforming both unguided discovery and conventional methods. Works such as Alfieri et al. (2011) and Schofield et al. (2021) further support that guided discovery significantly enhances learning outcomes.

## 3 Assessment of Phonological Understanding in LLMs

In this section, we aim to explore the extent to which LLMs inherently possess phonological knowledge by evaluating their performance in a zero-shot context. Additionally, we plan to examine the effectiveness of few-shot learning in further enhancing the models' phonological reasoning. In our exploration of the phonological capabilities of text-based LLMs, we focus on evaluating three key tasks: g2p conversion, rhyme word generation, and syllable counting. We utilize the PhonologyBench framework Suvarna et al. (2024) as a comprehensive benchmark for assessing these tasks.

| Feature | Count |
|---|---|
| G2P Conversion (High) | 1042 |
| G2P Conversion (Low) | 2084 |
| Rhyme Word Generation (Common) | 199 |
| Rhyme Word Generation (Rare) | 110 |
| Syllable Counting | 993 |

Table 2: Distribution of PhonologyBench Data

### 3.1 Phonological Benchmark Definition

We adopted the PhonologyBench framework (Suvarna et al., 2024), which forms a critical component of our study, to systematically evaluate the phonological knowledge of LLMs. Specifically, it includes datasets with 3,126 words for g2p conversion, categorized into 2,084 high-frequency and 1,042 low-frequency words. Additionally, it offers a dataset of 309 words, split into 199 common words and 110 rare ones for rhyme word generation tasks. For each word, a gold reference set was constructed by matching, on average, 1,200 rhyming candidates, ensuring a robust and comprehensive evaluation criterion. For syllable counting,

| Model(Open/Closed) | Baseline | 3-Shot | 5-Shot | **P-CoT1** | **P-CoT3** | **P-CoT5** |
|---|---|---|---|---|---|---|
| Llama-3.3-70B-Instruct | 66.3/34.4 | 65.3/38.4 | 64.7/36.9 | **77.2**/47.1 | 76.5/46.4 | 76.4/**47.5** |
| Llama-3.1-8B-Instruct | 41.7/17.5 | 41.3/21.6 | 39.7/27.1 | 54.7/30.2 | 55.5/33.3 | **62.9/37.1** |
| Mistral-7B-Instruct-v0.2 | 26.8/12.0 | 28.5/9.3 | 27.9/9.3 | 31.9/15.2 | 71.6/30.2 | **78.8/36.2** |
| Ministral-8B-Instruct-2410 | 16.2/8.7 | 26.6/6.9 | 28.6/8.9 | 34.4/20.8 | 50.2/22.7 | **63.9/31.9** |
| Qwen2.5-72B-Instruct | 51.8/26.7 | 54.8/25.8 | 53.6/22.4 | 66.6/35.8 | 68.1/**36.2** | **69.9**/35.4 |
| Qwen2.5-7B-Instruct | 23.7/9.5 | 21.2/7.3 | 22.1/6.7 | 43.5/15.6 | 45.8/29.2 | **48.5/30.2** |
| gemma-2-9b-it | 55.6/40.5 | 57.1/38.7 | 57.6/39.3 | **73.0**/52.5 | 71.8/47.7 | 71.2/**62.3** |
| gemma-2-27b-it | 62.4/39.6 | 64.5/34.9 | 63.4/34.7 | **78.4/52.7** | 77.3/51.3 | 77.7/51.2 |
| gpt-3.5-turbo | 70.7/47.3 | 72.9/45.1 | 73.1/45.5 | 79.2/**61.9** | 80.5/59.2 | **81.3**/51.2 |
| gpt-4o | 73.0/50.7 | 73.8/44.9 | 74.7/49.5 | - | **86.0/60** | 84.3/57.5 |
| Claude 3.5 Sonnet | 76.4/38.2 | 76.9/36.7 | 77.3/37.1 | 82.5/50.8 | 83.1/**53.3** | **83.2**/51.2 |
| Claude 3.5 Haiku | 68.5/41.5 | 70.2/41.1 | 70.5/41.5 | 80.2/34.8 | 81.7/58.7 | **82.4/58.8** |
| Human Baseline | 86.4/60.4 | | | | | |

Table 3: Performance Comparison on Rhyme Word Generation Tasks (Common/Rare Word Sets)

the benchmark includes 993 sentences (Table 2).

### 3.2 Models

We aim to evaluate the overall performance of LLMs on typical phonological tasks by assessing the performance of up to 12 different LLM models. Among them, 8 are open-source models and 4 are closed models.

**Open Models** We select a diverse set of open-source models, including both foundational and advanced versions of the latest LLMs, with parameter counts ranging from 7 billion to 72 billion. This lineup features Meta's LLaMA series models like Llama-3.3-70B and Llama-3.1-8B, as well as Mistral's Mistral-7B and Ministral-8B (Dubey et al., 2024; Jiang et al., 2023; AI, 2024). Additionally, we include various iterations from the Qwen and gemma series (Yang et al., 2024; Team et al., 2024).

**Closed Models** In addition to the open-source models, we also examine a selection of top-tier, proprietary models, which include OpenAI's GPT series (featuring GPT-3.5-turbo and GPT-4o (OpenAI, 2022; Anand et al., 2023)) and Anthropic's Claude series (Claude 3.5 Sonnet and Claude 3.5 Haiku (Anthropic, 2024b,a)).

### 3.3 Metrics for Model Evaluation

To assess the performance of the models on g2p conversion and syllable counting tasks, we use the Exact Match criterion as our measure of accuracy. For the rhyme word generation task, we evaluate performance using a Success Rate metric. This metric calculates the proportion of rhyming words generated by the model that match those in a predefined set of expected outcomes. Specifically, we determine the success rate for each word by finding the ratio of correctly generated candidates present in the ground-truth set. The overall Success Rate (SR) is then derived as the average of these individual success rates across all words. By applying these metrics, we remain consistent with the evaluation methods used in PhonologyBench (Suvarna et al., 2024).

### 3.4 Configuration

To ensure consistent and reproducible results across our experiments, we fix the model generation probability by setting a deterministic seed. This allows us to maintain consistent outputs across different trials and accurately evaluate the true performance improvements attributed to our methods. We conduct inference using several A100 GPUs, each equipped with 80GB of memory, without applying techniques like quantization (Frantar et al., 2022; Liu et al., 2023) or FlashAttention (Dao, 2023). Also, we adhere closely to the zero-shot prompts provided by PhonologyBench to accurately assess the models.

## 4 Results and Analysis

### 4.1 Analysis of Few-Shot Learning Effectiveness

We analyzed the effectiveness of few-shot learning in phonological reasoning tasks. We compared this approach to traditional baseline models across various LLMs and phonological tasks to assess

| Model (Open/Closed) | Baseline | 3-Shot | 5-Shot | **P-CoT1** | **P-CoT3** | **P-CoT5** |
|---|---|---|---|---|---|---|
| Llama-3.3-70B-Instruct | 27.0/40.5 | 29.9/42.9 | 29.6/43.8 | 32.2/45.6 | 32.8/48.8 | **33.2/49.7** |
| Llama-3.1-8B-Instruct | 17.2/24.1 | 21.3/31.5 | 21.4/31.9 | 21.8/**31.9** | **22.4**/29.1 | 22.0/29.8 |
| Mistral-7B-Instruct-v0.2 | 9.5/15.0 | 13.5/19.2 | 13.1/19.3 | 14.3/22.2 | **16**/21.9 | 15.6/**22.4** |
| Ministral-8B-Instruct-2410 | 19.9/31.4 | 20.3/30.8 | 20.6/32.1 | 22.3/32.5 | **23.7/35.3** | 23.1/34.6 |
| Qwen2.5-72B-Instruct | 25.3/38.0 | 30.0/42.1 | 28.7/43.2 | 28.2/40.2 | **30.1/43.4** | 28.8/41.4 |
| Qwen2.5-7B-Instruct | 12.5/18.6 | 13.3/19.9 | **13.6/22.3** | 10.0/17.4 | 10.8/17.9 | 10.2/16.6 |
| gemma-2-9b-it | 24.1/36.2 | 25.9/40.0 | 26.4/39.8 | 27.7/40.5 | **28.6/44.2** | 27.4/42.3 |
| gemma-2-27b-it | 29.6/43.8 | 30.4/46.9 | 31.1/46.6 | 31.1/46.7 | **32.1/48.7** | 31.4/47.9 |
| gpt-3.5-turbo | 28.9/43.2 | 42.6/60.3 | 38.7/56.8 | 45.4/61.9 | **50.1/68.3** | 49.3/67.5 |
| gpt-4o | 32.0/49.8 | 46.9/69.5 | 44.1/57.1 | 49.7/69.2 | 51.7/**69.8** | **52.1**/69.6 |
| Claude 3.5 Sonnet | 35.5/51.6 | 57.2/79.3 | 57.0/79.8 | 59.9/80.4 | 61.0/81.3 | **61.6/82.2** |
| Claude 3.5 Haiku | 33.5/49.3 | 55.9/**76.4** | 54.4/74.5 | 54.0/74.5 | **56.9**/75.1 | 56.5/75.4 |

Table 4: Performance Comparison on G2P Conversion Tasks (low/high Word Sets)

performance differences and identify any notable improvements.

**Rhyme Word Generation:** In the rhyme word generation task, few-shot learning demonstrated mixed results, as shown in Table 3. While some models, like GPT-3.5-turbo, showed improvements over the baseline with a 2-3 percentage point increase in accuracy for common rhymes, others, such as the Llama, Qwen, and gemma series, experienced only marginal gains or even declines. This inconsistency underscores the varying ability of models to leverage contextual examples efficiently, indicating that the effectiveness of few-shot learning is heavily dependent on the model's intrinsic capacity to process and learn from examples.

**G2P Conversion:** In the g2p conversion task, as shown in Table 4, few-shot learning led to variable results. Models like Claude 3.5 Sonnet experienced a significant boost, increasing from a baseline of 35.5% to 57.2% accuracy on the low-frequency set. Conversely, other models, such as gemma-2-27b-it, showed minimal progress, remaining near their baseline scores.

**Syllable Counting:** As presented in Table 5, few-shot learning provided modest gains in the syllable counting task. Notably, models like Claude 3.5 Haiku achieved a significant improvement, rising from a baseline accuracy of 21.1% to 57.4%. However, others, such as Mistral-7B and Ministral-8B, showed little advancement.

### 4.2 P-CoT: A Pedagogically-motivated Participatory Chain of Thought Prompting

[2] Building on our initial observations that revealed limited success with traditional few-shot learning in enhancing phonological reasoning, we propose a novel approach: the Pedagogically-motivated Participatory Chain-of-Thought (P-CoT). This framework is specifically designed to expand the latent phonological capabilities inherent in text-based LLMs (Figure 1). Our method integrates discovery learning (Bruner, 1961; Anthony, 1973), and scaffolding (Wood et al., 1976) to provide structured guidance and contextual support. Additionally, inspired by Vygotsky's ZPD (Vygotsky, 1978) and the instructional strategies of Collins et al. (1991) and Jatisunda et al. (2020), we offer temporary reciprocal guidance through targeted hints, clear definitions, and specific tasks. Moreover, by emphasizing the active engagement of the learner, our prompt design ensures that attentive and motivated models effectively navigate the discovery process and avoid misconceptions, ultimately improving performance on phonological tasks.

Our approach actively engages LLMs with structured prompts that simulate guided exploration, leveraging their capacity to infer and generalize from examples, as previous educational research has recommended (Tuovinen and Sweller, 1999; Moreno, 2004). This also facilitates their internalization of complex phonological rules and is expected to surpass the capabilities of traditional few-shot learning (Brown et al., 2020), enabling a more intricate application of phonological knowledge.

[2] https://github.com/Junmaij/P-CoT

| Model (Open/Closed) | Baseline | 3-Shot | 5-Shot | **P-CoT1** | **P-CoT3** | **P-CoT5** |
|---|---|---|---|---|---|---|
| Llama-3.3-70B-Instruct | 18.6 | 13.5 | 11.6 | **41.6** | 28.7 | 28.1 |
| Llama-3.1-8B-Instruct | 13.0 | 14.2 | 13.3 | **41.3** | 37.4 | 39.6 |
| Mistral-7B-Instruct-v0.2 | 2.9 | 3.7 | 4.4 | 13.0 | **26.5** | 14.9 |
| Ministral-8B-Instruct-2410 | 15.6 | 3.9 | 5.8 | 10.8 | **26.5** | 14.9 |
| Qwen2.5-72B-Instruct | 12.9 | 14.6 | 15.0 | 14.8 | **15.6** | 10.7 |
| Qwen2.5-7B-Instruct | 1.0 | 2.1 | 4.8 | 2.5 | **10.3** | 4.3 |
| gemma-2-9b-it | 7.3 | 9.3 | 8.7 | **19.4** | 8.7 | 12.6 |
| gemma-2-27b-it | 13.6 | 23.4 | **23.5** | 17.1 | 20.3 | 20.6 |
| gpt-3.5-turbo | 16.0 | 19.1 | 19.1 | **48.8** | 46.5 | 42.8 |
| gpt-4o | 20.8 | 20.9 | 19.5 | 20.4 | **22.2** | 19.8 |
| Claude 3.5 Sonnet | 20.2 | 24.8 | **32.9** | 23.4 | 28.8 | 32.6 |
| Claude 3.5 Haiku | 21.1 | 20.7 | 26.9 | 48.2 | **57.4** | 54.5 |

Table 5: Performance Comparison on Syllable Counting Tasks

#### 4.2.1 Prompt Design

In our exploration of the P-CoT methodology, we utilized the same sample sets used in our earlier few-shot learning experiments to ensure consistency in evaluating the models' performance enhancements. This approach allowed us to directly compare the efficacy of our P-CoT approach against the baselines established in previous tests.

Specifically, we defined three distinct configurations based on the number of interactions or dialogues included in the P-CoT prompt:

**P-CoT Prompt**

- **P-CoT1:** Presents a single teacher-student interaction, where the model is guided to derive a phonological rule from a single example.
- **P-CoT3:** Extends the same instructional format to three examples, while maintaining the dialogic structure.
- **P-CoT5:** Further increases the number of interactions to five, reinforcing the target concept through guided discovery learning.

The general structure of our P-CoT prompts is designed to simulate an instructional setting grounded in educational theory. Specifically, the prompts implement a form of guided discovery learning, which combines the principles of discovery learning with scaffolding techniques that support the learner's reasoning process. Each prompt begins by assigning teacher and student personas to the model and the user, respectively, thereby establishing an interactive and participatory learning environment. This role-based setup reflects a classroom-like context.

To initiate this interaction, the student persona is either explicitly signaled by phrases like "I'm ready to learn about ~," or implicitly conveyed through contextual cues and role-setting instructions. This design is inspired by Jatisunda et al. (2020) and Wood et al. (1976), who emphasize that guided discovery learning requires learners to actively pursue knowledge. Following this setup, the teacher persona introduces the conceptual basis of the task, providing necessary definitions to prevent misconceptions—serving as the first scaffolding step (Jatisunda et al., 2020). Then, consistent with Schofield et al. (2021), our prompt design does not aim to convey abstract concepts directly. Instead, the teacher persona guides the student role toward understanding by prompting it to solve concrete problems. To further support this learning process, we apply scaffolding by decomposing the problem into subtasks—a core principle emphasized by Collins et al. (1991) and Ewing McMahon (2000). For instance, in the syllable counting task, the teacher presents a supporting subtask, prompting the student to identify the vowel sounds in each word before proceeding to count the total syllables in the given sentence.

As discussed above, we embed a teacher-student dialogue that simulates an instructional exchange into the prompt, aiming to enhance the model's reasoning ability by situating it within a step-by-step learning process. By applying these structured prompts designed around educational theories of discovery learning and scaffolding, we aim to comprehensively assess the impact of P-CoT on improving phonological reasoning over traditional few-shot learning methods. The detailed design of these prompts is provided in the Appendix A.

#### 4.2.2 Result

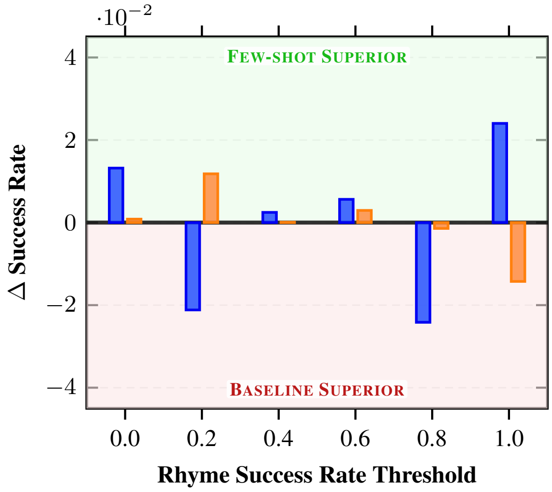
The implementation of P-CoT prompts consistently improved performance over the baseline across all examined models and tasks. By simulating a guided discovery learning environment, this approach capitalized on structured exploration and the deliberate scaffolding of knowledge, thereby facilitating deeper cognitive engagement and enhancing phonological competence.

**Rhyme Word Generation:** In the domain of rhyme word generation, the P-CoT strategy significantly enhanced model performance, as shown in Table 3. The P-CoT approach leveraged structured prompts to effectively improve the model's ability to generate rhymes, achieving notable advancements in both common and rare word categories. For instance, models such as Mistral-7B-Instruct-v0.2 and Ministral-8B-Instruct-2410 demonstrated performance increases of approximately 47.7 to 52 percentage points over the baseline when tasked with common rhymes. Even within closed models like GPT-4, structured pedagogical approaches led to performance levels approaching that of human baseline.

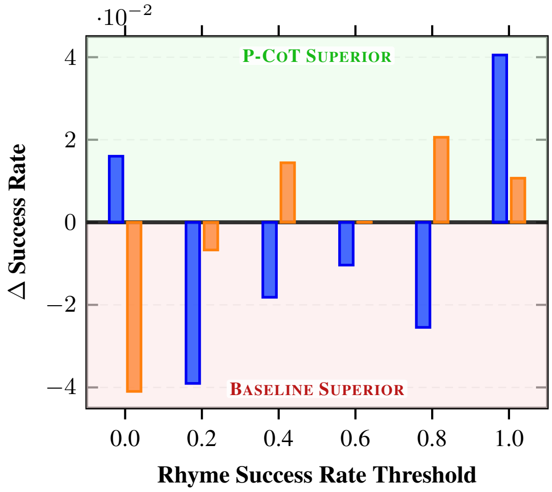
**G2P Conversion:** In the g2p conversion, the P-CoT approach significantly boosted model performance, as evidenced in Table 4. By utilizing structured prompts that guide learning, we observed substantial improvements in accuracy across both high-frequency and low-frequency word categories. Notably, models like Claude 3.5 Sonnet experienced a marked increase in successful conversions, with accuracy improvements reaching up to 30 percentage points above baseline levels.

**Syllable Counting:** In the syllable counting task, the P-CoT method showed notable progress, significantly enhancing the performance of multiple models, as detailed in Table 5. For instance, GPT-3.5-turbo saw its accuracy leap from a baseline of 16.0% to 48.8% under the P-CoT1 configuration. Similarly, Claude 3.5 Haiku improved from 21.1% to 57.4%, illustrating the potency of structured learning prompts.

## 5 Evaluating the Pedagogical Impact on Phonological Comprehension: A Deep Dive into P-CoT Efficacy




(a) Few-shot over Baseline (Blue: Common Words, Orange: Rare Words)




(b) P-CoT over Baseline (Blue: Common Words, Orange: Rare Words)

Figure 2: Success Rate Improvement of Rhyme Word Generation Analysis

In our exploration of enhancing phonological reasoning within text-based LLMs through the P-CoT methodology, our results underscore the significant potential of educationally-inspired frameworks to enhance model performance across all three evaluated phonological tasks: rhyme word generation, g2p conversion, and syllable counting.

### 5.1 Rhyme Word Generation

In Figure 2, the performance of two instructional strategies—few-shot and P-CoT—against the baseline is compared for rhyme word generation. Both results illustrate the success rate differences across six different conditions (success rate thresholds from '0.0' to '1.0'), with the y-axis explicitly representing the performance difference between the strategies and the baseline.

For common words, few-shot displays instability, even falling below baseline at higher success rate thresholds like *0.8*. In contrast, P-CoT steadily improves over the baseline, notably at *1.0* thresholds, demonstrating its effectiveness through structured prompts that consistently enhance model performance.

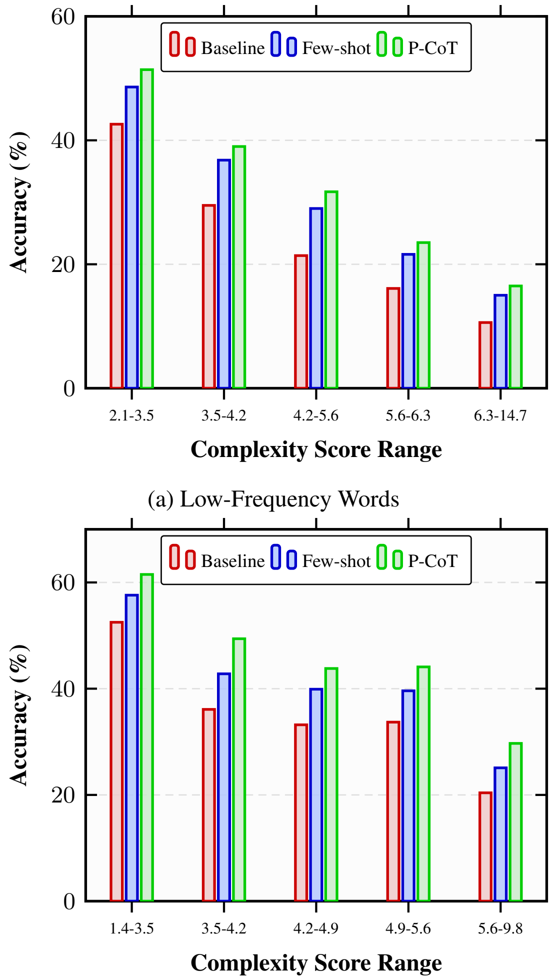
For rare words, few-shot shows limited gains and often underperforms relative to the baseline, particularly at *1.0*, while P-CoT maintains positive improvements at *1.0* success rate. This consistency highlights P-CoT's reliable handling of complex word sets, emphasizing its robustness over few-shot. Thus, P-CoT stands out as the more effective strategy for achieving high reliability and performance in generating rhymes, making it a preferable approach for both common and rare word categories.

## 5.2 G2P Conversion

To better understand the result in Table 4, we analyze model performance against a complexity score $S$ derived from the formula $S = 0.4L + 0.3V + 0.3C$, where $L$ indicates word length, $V$ counts vowels, and $C$ assesses consonant quantity. Figures 3a and 3b illustrate a clear trend: as word complexity rises, accuracy tends to decline across all evaluated methods, including baseline, few-shot, and P-CoT. However, P-CoT consistently outshines the other methods, even as complexity increases, demonstrating its robust capability to handle intricate g2p conversion.

For words with lower complexity, all approaches achieve relatively high accuracy. It is at higher levels of complexity that the performance gap becomes more pronounced, with P-CoT showing considerable advances. This effectiveness is further confirmed by Table 6, where we see statistically significant improvements of P-CoT over baseline models in both types of words. Consequently, P-CoT clearly emerges as a superior technique for enhancing phonological reasoning in complex scenarios.

Also, our result is supported by statistically significant p-values from the Mann-Whitney U Test (Nachar et al., 2008): notably, P-CoT shows significant enhancements over baseline performance for high-frequency (p = $5.72 \times 10^{-6}$) and low-frequency words (p = $6.44 \times 10^{-12}$). It also surpasses few-shot techniques for high-frequency words (p = $1.84 \times 10^{-2}$), though the improvement for low-frequency words is not significant (p = $2.37 \times 10^{-1}$). These results highlight P-CoT as an exceptionally effective approach for advancing g2p conversion task performance, effectively adapting to varying word complexities and frequencies.




(a) Low-Frequency Words

(b) High-Frequency Words

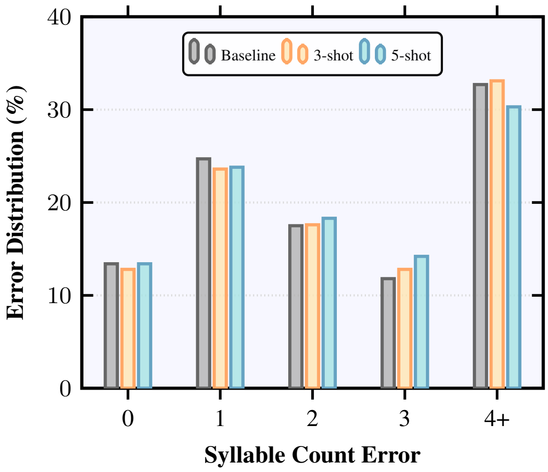
Figure 3: Accuracy by Word Complexity of G2P Conversion (Red: Baseline, Blue: Few-shot, Green: P-CoT)

## 5.3 Syllable Counting

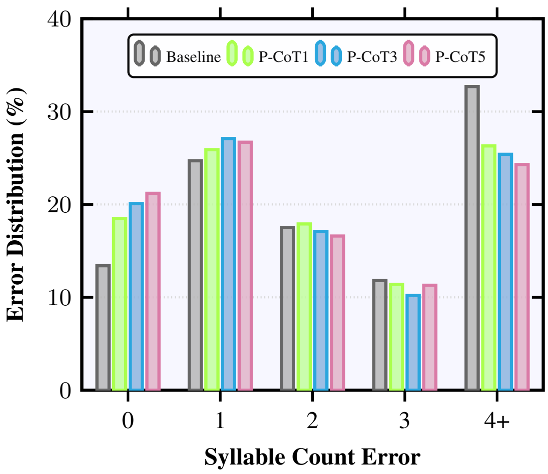
The improvement in Table 5 is further illustrated by the error analysis in Figure 4. Our error analysis compares few-shot learning over the baseline across error categories, represented on the x-axis by the number of syllabic errors $[0, 1, 2, 3, 4+]$. Despite using few-shot learning, the distribution remains relatively similar to the baseline, reflecting minimal impact on reducing errors, particularly in the 4+ error category, showing only a slight decrease with few-shot learning.

| Comparison | **High** | **Low** |
|---|---|---|
| Base vs Few | $1.47 \times 10^{-2}$ | $1.39 \times 10^{-8}$ |
| Base vs P-CoT | $5.72 \times 10^{-6}$ | $6.44 \times 10^{-12}$ |
| Few vs P-CoT | $1.84 \times 10^{-2}$ | $2.37 \times 10^{-1}$ |

Table 6: Mann-Whitney U Test p-values for G2P Conversion task (High/Low)




(a) Few-shot over Baseline




(b) P-CoT over Baseline

Figure 4: Syllable Counting Error Analysis over Baseline (Gray: Baseline, Orange/Teal: Few-shot, Lime/Cyan/Magenta: P-CoT variants)

Conversely, Figure 4b shows the distributions under P-CoT configurations. Notably, P-CoT prompts lead to a dramatic improvement in reducing instances with 4+ errors, evident from the baseline score of 32.67% dropping to 24.33% with P-CoT5. The 4+ error category's significant reduction underscores P-CoT's effectiveness, highlighting its superior ability to guide models towards more accurate syllabic prediction through structured exploration and learning.

These observations strongly support that the P-CoT framework enables LLMs to better interpret syllabic structure, reducing error frequency and distribution markedly. Specifically, the reduction of high-error categories demonstrates that P-CoT facilitates a deeper understanding of syllable structure, allowing models to achieve closer alignment with the correct syllabic count. This pattern affirms that educationally-inspired prompting significantly enhances LLMs' phonological reasoning capabilities, especially for more complex tasks.

## 6 Conclusion

Our study provides a foundational analysis of phonological reasoning within text-based LLMs, a domain that has been largely overlooked in prior research. Contrary to earlier assumptions, our findings reveal a substantial potential for phonological competence within these models. By introducing the Pedagogically-motivated Participatory Chain of Thought (P-CoT) prompting method, grounded in guided discovery learning and scaffolding, we demonstrate significant performance enhancements across rhyme word generation, g2p conversion, and syllable counting tasks, highly surpassing traditional few-shot methods. Our work paves the way for future investigations to refine LLM capabilities in phonology, and encourages the exploration of similar pedagogical approaches in other linguistic domains. Our framework offers a persuasive model for enhancing LLM performance, demonstrating the ability to tap into hidden capabilities using well-structured, education-inspired prompts.

## 7 Limitations

While offering valuable insights into enhancing phonological reasoning within text-based large language models, our study has some limitations. First, the proposed P-CoT prompting framework may not be uniformly optimal across all LLM architectures and sizes; different models could potentially benefit from further tailored prompt designs. Second, the evaluation primarily relies on PhonologyBench, which, while comprehensive, may not capture all nuances of phonological tasks across diverse languages and dialects. Third, our research focused on three specific phonological tasks, and the findings might not generalize across other, untested phonological aspects or linguistic phenomena. Furthermore, the number of example interactions in our prompts do not show a consistently linear relationship with performance improvements. This indicates that simply increasing the number of examples does not always lead to better outcomes, suggesting a potential interplay with factors like example quality, model capacity to process additional context, and inherent task complexity. Lastly, the effectiveness of our method is contingent on the quality and representativeness of the few-shot examples used, which could potentially introduce bias if unaccounted for. Future work should explore cross-linguistic evaluations and adaptive prompt strategies to further refine and validate our approach in broader, more varied contexts.

## 8 Ethical Consideration

Our research adheres to ethical standards, prioritizing transparency, inclusivity, and the responsible deployment of language models. We acknowledge the potential biases inherent in pretrained models and have consciously selected datasets to minimize skew, ensuring diverse representation within the scope of our study. The data and benchmark tasks utilized are publicly available and sourced from reputable repositories to uphold research integrity.

Furthermore, the advancements proposed by our P-CoT prompting are intended to enhance linguistic model capabilities without exacerbating existing issues such as bias or misinformation propagation. While our approach offers significant performance improvements, it is crucial for practitioners to critically evaluate the contexts in which these models are deployed, emphasizing responsible AI usage. This study encourages continuous dialogue and collaboration within the research community to foster ethically sound AI development and application.

## References


Mistral AI. 2024. Ministral-8b-instruct-2410.

Louis Alfieri, Patricia J Brooks, Naomi J Aldrich, and Harriet R Tenenbaum. 2011. Does discovery-based instruction enhance learning? *Journal of educational psychology*, 103(1):1.

Yuvanesh Anand, Zach Nussbaum, Brandon Duderstadt, Benjamin Schmidt, and Andriy Mulyar. 2023. Gpt4all: Training an assistant-style chatbot with large scale data distillation from gpt-3.5-turbo. https://github.com/nomic-ai/gpt4all.

WS Anthony. 1973. Learning to discover rules by discovery. *Journal of Educational Psychology*, 64(3):325.

Anthropic. 2024a. Claude 3.5 haiku.

AI Anthropic. 2024b. Claude 3.5 sonnet model card addendum. *Claude-3.5 Model Card*, 3(6).

Nathan Atox and Mason Clark. 2024. Evaluating large language models through the lens of linguistic proficiency and world knowledge: A comparative study. *Authorea Preprints*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

AL Brown. 1994. Guided discovery in a community of learners. *Classroom lessons: Integrating cognitive theory and classroom practice/press/Bradford Books*.

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

Jerome Bruner. 1985. Vygotsky: A historical and conceptual perspective. *Culture, communication, and cognition: Vygotskian perspectives*, 21:34.

Jerome S Bruner. 1961. The act of discovery. *Harvard educational review*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *Preprint*, arXiv:2303.12712.

Courtney B. Cazden. 1979. Peekaboo as an instructional model: Discourse development at home and at school. papers and reports on child language development, no. 17.

Edward Y. Chang. 2023. Prompting large language models with the socratic method. *Preprint*, arXiv:2303.08769.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, et al. 2024. From persona to personalization: A survey on role-playing language agents. *arXiv preprint arXiv:2404.18231*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Allan Collins, John Seely Brown, and Ann Holum. 1991. Cognitive apprenticeship: Making thinking visible. *American Educator: The Professional Journal of the American Federation of Teachers*, 15.

Tri Dao. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Thierry Dutoit. 1997. High-quality text-to-speech synthesis: An overview. *Journal Of Electrical And Electronics Engineering Australia*, 17(1):25–36.

Bronwyn F Ewing McMahon. 2000. Scaffolding: A suitable teaching characteristic in one-to-one teaching in maths recovery. In *Mathematics Education Beyond 2000. Twenty Third Annual Conference of the Mathematics Education Research group of Australasia Incorporated. 5-9 July, Fremantle, Western Australia.*, pages 417–423.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2022. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Pamela Thibodeau Hardiman, Alexander Pollatsek, and Arnold D Well. 1986. Learning to understand the balance beam. *Cognition and Instruction*, 3(1):63–86.

Yuchen Hu, Chen Chen, Chao-Han Huck Yang, Ruizhe Li, Dong Zhang, Zhehuai Chen, and Eng Siong Chng. 2024. Gentranslate: Large language models are generative multilingual speech and machine translators. *Preprint*, arXiv:2402.06894.

Mohamad Gilar Jatisunda, Vici Suciawati, and Dede Salim Nahdi. 2020. Discovery learning with scaffolding to promote mathematical creative thinking ability and self-efficacy. *Al-Jabar: Jurnal Pendidikan Matematika*, 11(2):351–370.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Zhuoxuan Jiang, Haoyuan Peng, Shanshan Feng, Fan Li, and Dongsheng Li. 2024. Llms can find mathematical reasoning mistakes by pedagogical chain-of-thought. *arXiv preprint arXiv:2405.06705*.

Paul Kirschner, John Sweller, and Richard E Clark. 2006. Why unguided learning does not work: An analysis of the failure of discovery learning, problem-based learning, experiential learning and inquiry-based learning. *Educational Psychologist*, 41(2):75–86.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Zechun Liu, Barlas Oguz, Changsheng Zhao, Ernie Chang, Pierre Stock, Yashar Mehdad, Yangyang Shi, Raghuraman Krishnamoorthi, and Vikas Chandra. 2023. Llm-qat: Data-free quantization aware training for large language models. *arXiv preprint arXiv:2305.17888*.

Tyler Loakman, Chen Tang, and Chenghua Lin. 2023. Twistlist: Resources and baselines for tongue twister generation. *Preprint*, arXiv:2306.03457.

Roxana Moreno. 2004. Decreasing cognitive load for novice students: Effects of explanatory versus corrective feedback in discovery-based multimedia. *Instructional science*, 32(1):99–113.

Nadim Nachar et al. 2008. The mann-whitney u: A test for assessing whether two independent samples come from the same distribution. *Tutorials in quantitative Methods for Psychology*, 4(1):13–20.

OpenAI. 2022. Gpt-3.5 turbo.

R OpenAI. 2023. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2(5).

Mahta Fetrat Qharabagh, Zahra Dehghanian, and Hamid R. Rabiee. 2024. Llm-powered grapheme-to-phoneme conversion: Benchmark and case study. *Preprint*, arXiv:2409.08554.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Paul K Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, et al. 2023. Audiopalm: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*.

Alexandra Schofield, Richard Wicentowski, and Julie Medero. 2021. Learning how to learn nlp: Developing introductory concepts through scaffolded discovery. In *Proceedings of the Fifth Workshop on Teaching NLP*, pages 131–137.

Murray Shanahan. 2022. Talking about large language models. *arXiv preprint arXiv:2212.03551*.

Ashima Suvarna, Harshita Khandelwal, and Nanyun Peng. 2024. Phonologybench: Evaluating phonological skills of large language models. *arXiv preprint arXiv:2404.02456*.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Lisa Miracchi Titus. 2024. Does chatgpt have semantic understanding? a problem with the statistics-of-occurrence strategy. *Cognitive Systems Research*, 83:101174.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Yu-Ching Hsu, Jia-Yin Foo, Chao-Wei Huang, and Yun-Nung Chen. 2024. Two tales of persona in llms: A survey of role-playing and personalization. *arXiv preprint arXiv:2406.01171*.

Juhani E Tuovinen and John Sweller. 1999. A comparison of cognitive load associated with discovery learning and worked examples. *Journal of educational psychology*, 91(2):334.

Wouter R Van Joolingen, Ton de Jong, Ard W Lazonder, Elwin R Savelsbergh, and Sarah Manlove. 2005. Co-lab: research and development of an online learning environment for collaborative scientific discovery learning. *Computers in human behavior*, 21(4):671–688.

Lev S Vygotsky. 1978. *Mind in society: The development of higher psychological processes*, volume 86. Harvard university press.

Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2023. Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. *arXiv preprint arXiv:2307.05300*.

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

David Wood, Jerome S Bruner, and Gail Ross. 1976. The role of tutoring in problem solving. *Journal of child psychology and psychiatry*, 17(2):89–100.

Lanqing Xue, Kaitao Song, Duocai Wu, Xu Tan, Nevin L Zhang, Tao Qin, Wei-Qiang Zhang, and Tie-Yan Liu. 2021. Deeprapper: Neural rap generation with rhyme and rhythm modeling. *arXiv preprint arXiv:2107.01875*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Tengfei Yu, Xuebo Liu, Zhiyi Hou, Liang Ding, Dacheng Tao, and Min Zhang. 2024. Self-powered llm modality expansion for large speech-text models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12401–12417.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023a. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv preprint arXiv:2305.11000*.

Ya-Jie Zhang, Chao Zhang, Wei Song, Zhengchen Zhang, Youzheng Wu, and Xiaodong He. 2023b. Prosody modelling with pre-trained cross-utterance representations for improved speech synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2812–2823.

# A Details on P-CoT Prompting Method

We explained that our approach delivers temporary, reciprocal guidance through targeted hints, clear definitions of key concepts, and the assignment of specific tasks. This design also enables models to reach a deeper level of understanding by emphasizing the learner's active engagement. In developing our detailed prompts, we specifically consulted real-world educational applications of these methods, as documented by Wood et al. (1976), Collins et al. (1991), Ewing McMahon (2000), Jatisunda et al. (2020), and Schofield et al. (2021). In this section, we provide a detailed overview of how we construct prompts based on the educational principles, illustrating each component of our framework.

Collins et al. (1991) propose a reciprocal teaching approach in which teachers and students alternate roles. Building on this idea, we tested two scenarios: one where the model served as the teacher and another where it acted as the student. Here we present the scenario that achieved the best performance for each task. The model functioned optimally as the teacher for syllable counting and rhyme word generation, while it achieved higher accuracy as the student in g2p conversion tasks. Although we report the optimal role for each task, the effectiveness may stem more from exposing the model to an interactive learning context that offers cues for reasoning.

### A.1 Rhyme Word Generation

Here is a detailed overview of our 3-shot P-CoT prompt for the rhyme word generation task. As shown in Table 8, we adopt a **role-based prompt** format, implemented using a system/user message structure common in instruction-tuned LLMs. In this setting, the role-setting message assigns the model a teacher persona by stating:

**Role Assignment**

*"You are an expert in American English phonology, phonetics, and morphology." "Your goal is to assist the student in discovering words that rhyme with a given word using discovery learning methods."*

Following Jatisunda et al. (2020), we also provide a clear definition of the core concept to prevent misconceptions:

**Conceptual Foundation:** *"Rhyming words are words that have the same ending sound. In simpler terms, it is the repetition of similar ending sounds."*

Before the guided exploration begins, the **User** adopts the student role, declaring:

**Student Readiness:** *"I'm ready to learn about rhyming words and how to identify them."*

In line with Jatisunda et al. (2020) and Wood et al. (1976), this statement frames the user as an attentive, motivated learner. The teacher (i.e., the **Assistant**) then reaffirms essential information:

**Scaffolding 1(Concept Understanding):** *"Let's start by understanding what rhyming words are. Remember, they share the same ending sounds."*

Rather than providing the answer outright, the teacher encourages the student to identify the ending sound of a sample word. This scaffolding is repeated three times in the 3-shot P-CoT, allowing the student to actively engage in the discovery process. When the student appears unsure, the teacher further breaks down the task:

**Scaffolding 2 (Task Decomposition):** *"Start by identifying the ending sound of 'information' and try to think of other words with a similar ending."*

This stepwise approach divides the problem into smaller tasks, promoting deeper understanding (Collins et al., 1991; Ewing McMahon, 2000). By embedding these teacher–student interactions directly into the prompt, our 3-shot P-CoT design demonstrates how guided discovery learning can ultimately enhance performance in rhyme word generation.

### A.2 G2P Conversion

This is a detailed overview of our 3-shot P-CoT prompt for g2p conversion, as illustrated in Table 8. We adopt a **role-based prompt** format, but in this scenario, we do not explicitly assign a student persona to the model. Instead, the role-setting message states:

**Role Assignment**

*"You are a leading expert in General American English (GAE) phonology. Your specialty is converting written words (graphemes) into their precise GAE pronunciation (phonemes)."*

This setup provides the model with a phonology-focused persona, implicitly allowing the **Assistant** to fulfill the role of an attentive learner, consistent with Jatisunda et al. (2020) and Wood et al. (1976).

The teacher (i.e., the **User**) initiates the learning process by offering an example, saying:

**Scaffolding 1 (Provide hints):** *"Let's analyze the word 'apparently' using standard GAE pronunciation. Show me its exact phonemic transcription, paying special attention to American rhotic sounds and stress patterns."*

Rather than supplying a direct solution, this prompt aligns with discovery learning by prompting the student to explore the g2p conversion process. Hints about GAE-specific pronunciations (e.g., rhotic sounds) serve as scaffolding, helping the student avoid misconceptions and focus on critical aspects of phoneme conversion.

In response, the student (i.e., the **Assistant**) details its reasoning, for instance:

> **Student Phonological Analysis:** *"In GAE, we have the unstressed initial əp, followed by the stressed syllable ˈɛɾ with the characteristic American rhotic ɾ, then the reduced ənt, and finally li."*

By explicitly describing its thought process, the student demonstrates the inquiry-driven nature of discovery learning (Bruner, 1961), gradually refining its approach through teacher-provided scaffolding (Wood et al., 1976). This interaction repeats three times in the 3-shot P-CoT, each iteration offering guidance on tricky GAE features.

Finally, recognizing that scaffolding should be gradually removed, the last request no longer provides word-specific hints. Instead, the teacher offers general guidelines and then states:

> **Independent Application Task:**
> *"Now, provide the precise General American English phonemic transcription for '{text}'."*

This gradual removal of support requires the student to apply what it has learned independently, aligning with the principle that scaffolding is intended to be temporary. Consequently, the student takes an active role in completing the g2p conversion, illustrating how our 3-shot P-CoT framework effectively integrates discovery learning and scaffolding to enhance phonological reasoning.

### A.3 Syllable Counting

Below is a detailed description of our 3-shot P-CoT prompt for the syllable counting task, illustrating its key components and flow. As with the other tasks, we adopt a **role-based prompt** format. As shown in Table 9, we begin by assigning the model a teacher persona through the role-setting message, which states:

> **Role Assignment**
>
> *"You are an expert in American English phonology and phonetics."*
> *"Your goal is to assist the student in counting the total number of syllables in a given sentence using discovery learning methods."*

Before the guided learning begins, the **User** adopts the student role by declaring:

> **Student Readiness:** *"I'm ready to learn about syllables and how to count them in a sentence."*

Following Jatisunda et al. (2020) and Wood et al. (1976), this statement establishes an attentive, motivated learner. The teacher (i.e., the **Assistant**) then reiterates essential information to prevent misconceptions, saying:

> **Scaffolding 1(Concept Understanding):** *"Let's start by understanding what a syllable is. Remember, a syllable has one vowel sound."*

The teacher first offers an example and encourages the student to attempt a solution, reflecting the discovery learning principle that learners must actively engage in problem-solving. If the student appears uncertain, the teacher further breaks down the task, for instance as suggested in Collins et al. (1991) and Ewing McMahon (2000):

> **Scaffolding 2 (Task Decomposition):**
> *"Start by identifying the vowel sounds in each word."*

This progressive guidance fosters discovery by dividing the overall process into smaller steps. When the student completes each step (e.g., counting the vowel sounds per word), the teacher offers additional support:

> **Scaffolding 3 (Task Decomposition):** *"Exactly! Now, let's sum them all up."*

In the 3-shot P-CoT setup, this cycle repeats three times. Finally, to maintain the principle that scaffolding is temporary, the final request does not break the task into sub-steps. Instead, it offers only a hint about identifying each vowel sound before instructing the teacher-model to:

**Independent Application Task:** *"Count the total number of syllables in the given word (in Target Sentence)."*

Through this reciprocal interplay, guided discovery learning is effectively realized, demonstrating how our P-CoT prompt facilitates deeper understanding in the syllable counting task.

## P-CoT5 Prompting Flow (Syllable Counting)

Dialogue Flow

Discovery Learning

**System:** You are an expert in American English phonology and phonetics. A syllable is a unit of pronunciation having exactly one vowel sound, forming the whole or a part of a word. Your goal is to assist the student in counting the total number of syllables in a given sentence using discovery learning methods.

**User:** "I'm ready to learn about syllables and how to count them in a sentence."

**Assistant:** "Great! Let's start by understanding what a syllable is. Remember, a syllable has one vowel sound. Now let's try an example together."

**Example 1:** "Sentence: 'Grace has resigned herself to simply completing the upbringing of her teenage daughter.' How many syllables do you think it has?"

**Assistant:** "Let's break it down together. Start by identifying the vowel sounds in each word."

**User:** "Grace (1), has (1), resigned (2), herself (2), to (1), simply (2), completing (3), the (1), upbringing (3), of (1), her (1), teenage (2), daughter (2)"

**Assistant:** "Exactly! Now, let's sum them all up."

**User:** "22"

**Assistant:** "That's correct! There are 22 syllables in that sentence."

**Example 2:** "Sentence: 'This story is about a young girl's redemption in a small town.' How many syllables do you think it has?"

**Assistant:** "Let's work through this one as well. Identify the vowel sounds in each word."

**User:** "This (1), story (2), is (1), about (2), a (1), young (1), girl's (1), redemption (3), in (1), a (1), small (1), town (1)"

**Assistant:** "Good job! Now add them together."

**User:** "16"

**Assistant:** "Correct again! It has 16 syllables."

**Examples 3-5:** Similar pattern with sentences yielding 13, 16, and 20 syllables respectively, following identical discovery-learning methodology.

**Final Task:** "Now it's your turn! Count the *total* number of syllables in the given word (in Target Sentence): '{text}.' Do not randomly generate the number; rather, you must analyze the vowel sounds and sum them accurately. Provide *only* the total count."

Figure 5: **P-CoT5 Prompting Flow for Syllable Counting Task.**

| Task | Prompt |
|---|---|
| **P-CoT1: Rhyme Word Generation** | **System:** You are an expert in American English phonology, phonetics, and morphology. Rhyming words share the same ending sound. As an expert, guide the user to discover 5 different words that rhyme with a given word by giving subtle hints and ensuring they find the exact answer.<br>**User:** "As the Prominent Phonology Professor, I know that rhyming words have the same ending sound. Give exactly 5 *different* words that rhyme with '{text}', separated by commas. Do not include any breakdown or additional explanation." |
| **P-CoT1: G2P Conversion** | **System:** You are a leading expert in General American English (GAE) phonology. Your specialty is converting written words (graphemes) into their precise GAE pronunciation (phonemes).<br>**User (example 1):** "Let's analyze the word 'apparently' using standard GAE pronunciation. Show me its exact phonemic transcription, paying special attention to American rhotic sounds and stress patterns."<br>**Assistant:** "The General American English pronunciation of 'apparently' is /əpˈɛɹəntli/. In GAE, we have the unstressed initial əp, followed by the stressed syllable ˈɛɹ with the characteristic American rhotic ɹ, then the reduced ənt, and finally li."<br>**User (final request):** "Now, provide the precise General American English phonemic transcription for '{text}'. Focus on accurate vowels, rhotics (ɹ), and stress. Give me *only* its complete GAE phonemic transcription." |
| **P-CoT1: Syllable Counting** | **System:** You are an expert in American English phonology and phonetics. A syllable is a unit of pronunciation having exactly one vowel sound. Your goal is to assist in counting the total number of syllables in a given sentence using discovery learning.<br>**User:** "I'm ready to learn about syllables and how to count them in a sentence."<br>**Assistant:** "Great! Let's start by understanding what a syllable is. Remember, a syllable has one vowel sound. Now let's try an example together."<br>**User (example 1):** "Sentence: 'Grace has resigned herself to simply completing the upbringing of her teenage daughter.' How many syllables do you think it has?"<br>**Assistant:** "Let's break it down together. Identify the vowel sounds in each word."<br>**User:** "Grace (1), has (1), resigned (2) ... teenage (2), daughter (2)"<br>**Assistant:** "Exactly! Now, let's sum them all up."<br>**User:** "22"<br>**Assistant:** "That's correct! There are 22 syllables in that sentence."<br>**User (final request):** "Now it's your turn! Count the TOTAL number of syllables in the given word (in Target Sentence): '{text}'. Do not randomly guess the number; you must identify vowel sounds and add them up. Provide *only* the total count." |

Table 7: P-CoT1 tasks (Rhyme Word Generation, G2P Conversion, Syllable Counting) with sample prompts.

| Task | Prompt |
|---|---|
| **P-CoT3: Rhyme Word Generation** | **System:** You are an expert in American English phonology, phonetics, and morphology. Rhyming words are words that have the same ending sound. In simpler terms, it is the repetition of similar ending sounds. Your goal is to assist the student in discovering words that rhyme with a given word using discovery learning methods.<br>**User:** “I’m ready to learn about rhyming words and how to identify them.”<br>**Assistant:** “Great! Let’s start by understanding what rhyming words are. Remember, they share the same ending sounds. Now let’s try an example together.”<br>**User (example 1):** “Word: ‘information.’ Can you think of any words that rhyme with ‘information’?”<br>**Assistant:** “Let’s break it down together. Start by identifying the ending sound of ‘information’ and try to think of other words with a similar ending.”<br>**User:** “Here are some words that rhyme with ‘information’: isolation, operation, conversation, corporation, demonstration.”<br>**Assistant:** “That’s correct! Those are great examples of rhyming words with ‘information’.”<br>**User (example 2):** “Word: ‘available.’ Can you think of any words that rhyme with ‘available’?”<br>**Assistant:** “Let’s work through this one as well. Identify the ending sound of ‘available’ and find other words with similar endings.”<br>**User:** “Here are some words that rhyme with ‘available’: distrainable, explainable, restrainable, retainable, retrainable.”<br>**Assistant:** “Correct again! Those are good examples of rhyming words with ‘available’.”<br>**User (example 3):** “Word: ‘transport.’ Can you think of any words that rhyme with ‘transport’?”<br>**Assistant:** “Let’s do the same process and find words with similar ending sounds to ‘transport’.”<br>**User:** “Here are some words that rhyme with ‘transport’: passport, escort, report, resort, retort.”<br>**Assistant:** “Perfect! Those words rhyme with ‘transport’.”<br>**User (final request):** “Now it’s your turn! Identify the ending sound and find words that rhyme with the given word. Word: ‘{text}.’ Give exactly 5 *different* words that rhyme with it, separated by commas. Begin with the phrase ‘Here are some words that rhyme with {text}:’” |
| **P-CoT3: G2P Conversion** | **System:** You are a leading expert in General American English (GAE) phonology. Your specialty is converting written words (graphemes) into their precise GAE pronunciation (phonemes).<br>**User (example 1):** “Let’s analyze the word ‘apparently’ using standard GAE pronunciation. Show me its exact phonemic transcription, paying special attention to American rhotic sounds and stress patterns.”<br>**Assistant:** “The General American English pronunciation of ‘apparently’ is /əpˈɛrəntli/. In GAE, we have the unstressed initial əp, followed by the stressed syllable ˈɛr with the characteristic American rhotic r, then the reduced ənt, and finally li.”<br>**User (example 2):** “For our next GAE analysis, let’s examine the word ‘calorie.’ What is its precise General American English phonemic transcription? Remember to account for the distinctive r-colored vowel that characterizes GAE.”<br>**Assistant:** “The General American English pronunciation of ‘calorie’ is /ˈkælɚi/. In GAE, we have the stressed syllable ˈkæl, followed by the r-colored schwa ɚ, and ending with i.”<br>**User (example 3):** “Let’s examine ‘freshman’ in General American English. What is its precise GAE phonemic transcription? Note particularly how the rhotic sound manifests in American pronunciation.”<br>**Assistant:** “The General American English pronunciation of ‘freshman’ is /ˈfɹɛʃmən/. We begin with the consonant cluster fɹ featuring the American rhotic, followed by the stressed ˈɛ ʃ, and ending with the reduced syllable mən.”<br>**User (final request):** “Now, provide the precise General American English phonemic transcription for ‘{text}.’ Focus on accurate representation of American vowels, rhotics (r), and stress patterns. Give me *only* its complete GAE phonemic transcription.” |

Table 8: P-CoT3 tasks (Rhyme Word Generation and G2P Conversion) with sample prompts.

| Task | | Prompt |
|---|---|---|
| **P-CoT3: Counting** | **Syllable** | **System:** You are an expert in American English phonology and phonetics. A syllable is a unit of pronunciation having exactly one vowel sound. Your goal is to assist the student in counting the total number of syllables in a given sentence using discovery learning methods.<br>**User:** "I'm ready to learn about syllables and how to count them in a sentence."<br>**Assistant:** "Great! Let's start by understanding what a syllable is. Remember, a syllable has one vowel sound. Now let's try an example together."<br>**User (example 1):** "Sentence: 'Grace has resigned herself to simply completing the upbringing of her teenage daughter.' How many syllables do you think it has?"<br>**Assistant:** "Let's break it down together. Start by identifying the vowel sounds in each word."<br>**User:** "Grace (1), has (1), resigned (2), herself (2), to (1), simply (2), completing (3), the (1), upbringing (3), of (1), her (1), teenage (2), daughter (2)"<br>**Assistant:** "Exactly! Now, let's sum them all up."<br>**User:** "22"<br>**Assistant:** "That's correct! There are 22 syllables in that sentence."<br>**User (example 2):** "Sentence: 'This story is about a young girl's redemption in a small town.' How many syllables do you think it has?"<br>**Assistant:** "Let's work through this one as well. Identify the vowel sounds in each word."<br>**User:** "This (1), story (2), is (1), about (2), a (1), young (1), girl's (1), redemption (3), in (1), a (1), small (1), town (1)"<br>**Assistant:** "Good job! Now add them together."<br>**User:** "16"<br>**Assistant:** "Correct again! It has 16 syllables."<br>**User (example 3):** "Sentence: 'The one thing that hasn't happened is a proposal.' How many syllables do you think it has?"<br>**Assistant:** "Let's do the same process and find the vowel sounds in each word."<br>**User:** "The (1), one (1), thing (1), that (1), hasn't (2), happened (2), is (1), a (1), proposal (3)"<br>**Assistant:** "Perfect! Now, sum them up."<br>**User:** "13"<br>**Assistant:** "Well done! The total is 13 syllables."<br>**User (final request):** "Now it's your turn! Count the *total* number of syllables in the given word (in Target Sentence): '{text}.' Do not randomly generate the number; rather, identify each vowel sound and add them up. Provide *only* the total count." |

Table 9: P-CoT3 task (Syllable Counting) with sample prompts.

| Task | Prompt |
|---|---|
| **P-CoT5: Rhyme Word Generation** | **System:** You are an expert in American English phonology, phonetics, and morphology. Rhyming words are words that have the same ending sound. In simpler terms, it can be defined as the repetition of similar ending sounds. Your goal is to assist the student in discovering words that rhyme with a given word using discovery learning methods.<br>**User:** "I'm ready to learn about rhyming words and how to identify them."<br>**Assistant:** "Great! Let's start by understanding what rhyming words are. Remember, they share the same ending sounds. Now let's try an example together."<br>**User (example 1):** "Word: 'information.' Can you think of any words that rhyme with 'information'?"<br>**Assistant:** "Let's break it down together. Start by identifying the ending sound of 'information' and try to think of other words with a similar ending."<br>**User:** "Here are some words that rhyme with 'information': isolation, operation, conversation, corporation, demonstration."<br>**Assistant:** "That's correct! Those are great examples of rhyming words with 'information'."<br>**User (example 2):** "Word: 'available.' Can you think of any words that rhyme with 'available'?"<br>**Assistant:** "Let's work through this one as well. Identify the ending sound of 'available' and find other words with similar endings."<br>**User:** "Here are some words that rhyme with 'available': distrainable, explainable, restrainable, retainable, retrainable."<br>**Assistant:** "Correct again! Those are good examples of rhyming words with 'available'."<br>**User (example 3):** "Word: 'transport.' Can you think of any words that rhyme with 'transport'?"<br>**Assistant:** "Let's do the same process and find words with similar ending sounds to 'transport'."<br>**User:** "Here are some words that rhyme with 'transport': passport, escort, report, resort, retort."<br>**Assistant:** "Perfect! Those words rhyme with 'transport'."<br>**User (example 4):** "Word: 'interesting.' Can you think of any words that rhyme with 'interesting'?"<br>**Assistant:** "Let's find words that have a similar ending sound to 'interesting'."<br>**User:** "Here are some words that rhyme with 'interesting': beginning, interrupting, diminishing, investing, referencing."<br>**Assistant:** "Well done! Those are excellent examples of rhyming words with 'interesting'."<br>**User (example 5):** "Word: 'technology.' Can you think of any words that rhyme with 'technology'?"<br>**Assistant:** "Try to recognize the ending sound of 'technology' and think of other words that share this ending."<br>**User:** "Here are some words that rhyme with 'technology': eternity, innocuity, unity, activity, amusingly."<br>**Assistant:** "Excellent! Those words rhyme with 'technology'."<br>**User (final request):** "Now it's your turn! Identify the ending sound and find words that rhyme with the given word. Word: '{text}.' Give exactly 5 *different* words that rhyme with it, separated by commas. Begin with the phrase 'Here are some words that rhyme with {text}:'" |

Table 10: P-CoT5 task (Rhyme Word Generation) with sample prompts.

| Task | Prompt |
|---|---|
| **P-CoT5: G2P Conversion** | **System:** You are a leading expert in General American English (GAE) phonology. Your specialty is converting written words (graphemes) into their precise GAE pronunciation (phonemes).<br>**User (example 1):** "Let's analyze the word 'apparently' using standard GAE pronunciation. Show me its exact phonemic transcription, paying special attention to American rhotic sounds and stress patterns."<br>**Assistant:** "The General American English pronunciation of 'apparently' is /əpˈɛɹəntli/. In GAE, we have the unstressed initial əp, followed by the stressed syllable ˈɛɹ with the characteristic American rhotic ɹ, then the reduced ənt, and finally li."<br>**User (example 2):** "For our next GAE analysis, let's examine the word 'calorie.' What is its precise General American English phonemic transcription? Remember to account for the distinctive r-colored vowel that characterizes GAE."<br>**Assistant:** "The General American English pronunciation of 'calorie' is /ˈkælɚi/. In GAE, we have the stressed syllable ˈkæl, followed by the r-colored schwa ɚ, and ending with i."<br>**User (example 3):** "Let's examine 'freshman' in General American English. What is its precise GAE phonemic transcription? Note particularly how the rhotic sound manifests in American pronunciation."<br>**Assistant:** "The General American English pronunciation of 'freshman' is /ˈfɹɛʃmən/. In GAE, we begin with the consonant cluster fɹ featuring the American rhotic, followed by the stressed ˈɛ ʃ, and ending with the reduced syllable mən."<br>**User (example 4):** "Let's examine the word 'breeze' in General American English. What is its precise GAE phonemic transcription? Pay attention to the rhotic sound and vowel length."<br>**Assistant:** "The General American English pronunciation of 'breeze' is /ˈbɹiːz/. In GAE, we have the initial consonant cluster bɹ, followed by the long vowel iːz, and ending with z."<br>**User (example 5):** "Now let's analyze the word 'invite' in General American English. What is its precise GAE phonemic transcription? Focus on the vowel quality and stress pattern."<br>**Assistant:** "The General American English pronunciation of 'invite' is /ɪnˈvaɪt/. In GAE, we begin with the unstressed syllable ɪn, followed by the stressed syllable ˈvaɪt containing the diphthong aɪ."<br>**User (final request):** "Now, provide the precise General American English phonemic transcription for '{text}.' Focus on accurate representation of American vowels, rhotics (ɹ), and stress patterns. Give me *only* its complete GAE phonemic transcription." |

Table 11: P-CoT5 task (G2P Conversion) with sample prompts.

| Task | | Prompt |
|---|---|---|
| **P-CoT5: Counting** | **Syllable** | **System:** You are an expert in American English phonology and phonetics. A syllable is a unit of pronunciation having exactly one vowel sound, forming the whole or a part of a word. Your goal is to assist the student in counting the total number of syllables in a given sentence using discovery learning methods.<br>**User:** "I'm ready to learn about syllables and how to count them in a sentence."<br>**Assistant:** "Great! Let's start by understanding what a syllable is. Remember, a syllable has one vowel sound. Now let's try an example together."<br>**User (example 1):** "Sentence: 'Grace has resigned herself to simply completing the upbringing of her teenage daughter.' How many syllables do you think it has?"<br>**Assistant:** "Let's break it down together. Start by identifying the vowel sounds in each word."<br>**User:** "Grace (1), has (1), resigned (2), herself (2), to (1), simply (2), completing (3), the (1), upbringing (3), of (1), her (1), teenage (2), daughter (2)"<br>**Assistant:** "Exactly! Now, let's sum them all up."<br>**User:** "22"<br>**Assistant:** "That's correct! There are 22 syllables in that sentence."<br>**User (example 2):** "Sentence: 'This story is about a young girl's redemption in a small town.' How many syllables do you think it has?"<br>**Assistant:** "Let's work through this one as well. Identify the vowel sounds in each word."<br>**User:** "This (1), story (2), is (1), about (2), a (1), young (1), girl's (1), redemption (3), in (1), a (1), small (1), town (1)"<br>**Assistant:** "Good job! Now add them together."<br>**User:** "16"<br>**Assistant:** "Correct again! It has 16 syllables."<br>**User (example 3):** "Sentence: 'The one thing that hasn't happened is a proposal.' How many syllables do you think it has?"<br>**Assistant:** "Let's do the same process and find the vowel sounds in each word."<br>**User:** "The (1), one (1), thing (1), that (1), hasn't (2), happened (2), is (1), a (1), proposal (3)"<br>**Assistant:** "Perfect! Now, sum them up."<br>**User:** "13"<br>**Assistant:** "Well done! The total is 13 syllables."<br>**User (example 4):** "Sentence: 'She meets him randomly in the woods at his family's cabin.' How many syllables do you think it has?"<br>**Assistant:** "Let's break it down by identifying the vowel sounds in each word."<br>**User:** "She (1), meets (1), him (1), randomly (3), in (1), the (1), woods (1), at (1), his (1), family's (3), cabin (2)"<br>**Assistant:** "Good! Now, tally them up."<br>**User:** "16"<br>**Assistant:** "Yes, that's 16 syllables."<br>**User (example 5):** "Sentence: 'Just a simple blacksmith's assistant, he didn't have much to offer, but his love.' How many syllables do you think it has?"<br>**Assistant:** "Follow the same process and find the vowel sounds in each word."<br>**User:** "Just (1), a (1), simple (2), blacksmith's (2), assistant (3), he (1), didn't (2), have (1), much (1), to (1), offer (2), but (1), his (1), love (1)"<br>**Assistant:** "Nice work! Now add them up."<br>**User:** "20"<br>**Assistant:** "Correct! The total number is 20 syllables."<br>**User (final request):** "Now it's your turn! Count the *total* number of syllables in the given word (in Target Sentence): '{text}.' Do not randomly generate the number; rather, you must analyze the vowel sounds and sum them accurately. Provide *only* the total count." |

Table 12: P-CoT5 task (Syllable Counting) with sample prompts.