EDUVELOP
JOURNAL OF ENGLISH EDUCATION AND DEVELOPMENT
Nationally Accredited Journal Decree No. 10/C/C3/DT.05.00/E5/2025 (Sinta 2)

Volume 9, No. 1, March 2026 • **Page: 329 - 339**
DOI: https://doi.org/v9i1.6207
ISSN 2597-713X (print)
ISSN 2597-7148 (online)

OPEN ACCESS: Research Article

Prompt Engineering to CEFR Alignment: Investigating Generative AI for the Creation of English Listening Assessments

Fikri Asih Wigati[1*], Putri Kamalia Hakim [2], Nia Pujiawati [3], Maya Rahmawati [4]
[1*]English Language Education Study Program, Universitas Singaperbangsa Karawang, Karawang, Indonesia

*Correspondence e-mail: fikri.asihwigati@staff.unsika.ac.id

Received : 2026-01-30	**Accepted :** 2026-03-05
Revision : 2026-03-07	**Published :** 2026-03-31

Abstract
The increasing globalization of higher education has emphasized the need for English listening assessments that are both practical and internationally benchmarked. However, creating high-quality listening test materials can be resource intensive. It requires substantial expertise, time, and infrastructure, particularly in institutions with limited resources. This study aims to investigate the feasibility of using generative artificial intelligence, specifically OpenAI's ChatGPT-4, to support the early-stage development of CEFR-referenced English listening scripts and test items for proficiency-oriented assessment contexts. Adopting an exploratory research design, the study employed an iterative prompt engineering approach to guide ChatGPT-4 in producing a set of listening scripts and multiple choice items developed with reference to CEFR levels A2, B1, B2, and C1. CEFR descriptors served as reference points for guiding linguistic difficulty during material development rather than as a structural framework for test composition or score interpretation. The generated materials were examined through descriptive linguistic analysis using Text Inspector, which focused on script length, readability trends, CEFR oriented lexical profiling, alongside a qualitative review of topical coverage and distractor plausibility. The results suggested that ChatGPT-4 can effectively serve as a supportive drafting tool when integrated into a structured, guided development process guided by human. The generated scripts showed patterned linguistic variation across CEFR reference levels, while also had limitations in managing features of spoken discourse, naturalness at higher levels, and balanced topic representation. In conclusion, generative AI shows potential to reduce reliance on costly external resources in developing listening material, as long as expert human oversight remains central.

Keywords: *Generative AI; ChatGPT 4; CEFR; Listening test; prompt engineering*

DOI: https://doi.org/v9i1.6207

1. Introduction

The growing globalization of higher education has increased the need for clear and widely accepted measures of English language proficiency. Universities are expected to help students develop English skills that are useful for academic purposes and recognized beyond local contexts, as institutions increasingly rely on shared benchmarks to assess language ability for study progression and international engagement (Coleman et al., 2024). Among the language skills assessed, listening comprehension plays a central role, as it supports effective communication and academic participation, especially in settings where English is used as the main language of instruction or interaction (Aryadoust & Luo, 2023). From an assessment perspective, this raises important questions about how listening ability is defined and measured.

Listening assessment is concerned not only with language input, but also with how well test tasks accurately represent the underlying listening construct. According to established frameworks in language assessment, listening involves multiple dimensions, including the ability to process spoken input in real time, interpret meaning at both literal and inferential levels, and manage discourse-level information such as coherence and speaker intent (Field, 2008; Taylor & Geranpayeh, 2011). As a result, the development of listening test materials requires careful control of linguistic features, task design, and input characteristics to ensure that test items reflect the intended construct rather than unrelated factors such as background knowledge or test-taking strategies. This perspective highlights the need to align linguistic difficulty, discourse features, and item design within a clear assessment framework.

In this study, listening ability is understood as a multidimensional construct that includes several core components of academic listening. These include identifying main ideas, interpreting speaker intention, extracting specific information, and making basic inferences based on the listening input (Sawaki et al., 2009; Taylor & Geranpayeh, 2011). In addition, aspects such as discourse processing and coherence recognition are also considered in the design of the listening scripts. However, the study focuses primarily on receptive processing of spoken input rather than interactive or real-time communicative competence.

Despite this theoretical understanding, the development of high-quality, standardized English listening tests presents significant challenges, particularly for institutions operating with limited resources. The development of authentic and contextually relevant listening scripts, producing high-fidelity audio recordings, and constructing valid and reliable test items requires considerable amount of time and expertise in applied linguistics, language test design, and often, expensive infrastructure (Nurhayati et al., 2024). Test developers often struggle to secure a sufficient range of speakers, produce naturally paced listening passages, precisely aligning difficulty levels with established frameworks such as the Common European Framework of Reference for Languages (CEFR) (Field, 2008; McKinley & Rose, 2017).

In response to these challenges, recent advancements in artificial intelligence (AI), particularly generative AI models like OpenAI's ChatGPT-4, offer alternative approaches for language assessment development (Aryadoust et al., 2024). Generative AI

DOI: https://doi.org/v9i1.6207

is capable of producing extended written texts that approximate natural discourse patterns, making it a potentially useful tool for drafting listening scripts. Through techniques such as prompt engineering and iterative fine-tuning, this technology can significantly enhance the efficiency, flexibility, and affordability of developing examination materials, potentially in context with limited resources (Mead & Zhou, 2023).

Alongside these developments, the use of generative AI in language assessment has raised important concerns related to validity, reliability, and ethical use. Recent discussions highlight potential risks such as bias in generated content, inconsistencies in output quality, and challenges in maintaining construct validity when AI-generated materials are used without sufficient human oversight. In addition, issues related to fairness, transparency, and test security have become increasingly relevant, particularly in high-stakes assessment contexts. These concerns suggest that the integration of AI in assessment should be approached cautiously, with careful consideration of both its potential benefits and limitations.

Previous research has explored the capabilities of large language models (LLMs) in various educational contexts, ranging from generating narrative texts and summarizing academic articles to answering critical thinking questions (Peláez-Sánchez et al., 2024; Jiang et al., 2024). Models like GPT-3 have demonstrated remarkable text generation abilities, laying the groundwork for more specialized applications (Richardson, 2021). The subsequent development of models such as ChatGPT-4, with enhanced reasoning and content generation capabilities, has enabled more complex applications in presdomains like language assessment (OpenAI, 2023). However, the systematic use of generative AI for constructing CEFR-referenced listening scripts and test items, particularly within the specific resource constraints of Indonesian university contexts, remains limited in the literature, particularly in studies that examine AI-supported development as a structured, assessment-oriented process. This study, therefore, focuses on a human–AI collaborative approach that integrates CEFR-referenced linguistic analysis, iterative prompt engineering, and structured human revision within the material development process. Instead of evaluating the model of autonomous system, the study examines how generative AI can be used as a drafting support tool within a controlled assessment design framework.

In this Listening assessment is not merely concerned study, the CEFR is employed as a reference framework for controlling linguistic difficulty during the material development stage, rather than as a structural blueprint for test construction or score interpretation. Although proficiency tests do not organize items by discrete CEFR levels in their final form, CEFR descriptors are widely used in early-stage test development to guide difficulty targeting and linguistic calibration. Accordingly, CEFR levels in this study function as reference points for specifying lexical, syntactic, and discourse-level features of listening scripts, while empirical validation is reserved for future research. Within this framework, linguistic quality refers to the extent to which the generated listening scripts demonstrate appropriate lexical range, syntactic complexity, readability, and alignment with CEFR-referenced difficulty levels, as well as coherence

DOI: https://doi.org/v9i1.6207

and plausibility in supporting listening comprehension tasks.

The aim of this study is to investigate the feasibility of using ChatGPT-4 to generate CEFR-referenced English listening scripts and test items for proficiency-oriented assessment contexts. The findings are expected to offer practical insights into how generative AI may reduce reliance on costly external resources, such as professional voice actors and recording studios, thereby supporting more accessible and sustainable language assessment development. Furthermore, this research aligns with institutional priorities in higher education concerning digital learning technologies and Industry 4.0 competencies, contributing to the preparation of graduates for global contexts. Based on this background, this study is designed to address the following research questions:

1. How can ChatGPT-4, as a generative AI model, be utilized in the development of CEFR-based English listening scripts and test items?
2. To what extent do the linguistic quality and topical diversity of ChatGPT-4 outputs meet the established standards for listening test item construction?

2. Methods

This study used an iterative exploratory research design that focused on early-stage material development and descriptive review of generative AI outputs. The purpose of the study was to examine whether generative AI could be used to support the development of CEFR-referenced English listening scripts and test items. OpenAI's ChatGPT-4, a large language model trained on large text datasets, was used as the main tool for generating materials. The exploratory design made it possible to revise prompts continuously and to review the generated outputs qualitatively, with CEFR descriptors serving as reference points for guiding decisions about linguistic difficulty during the development process.

2.1. Prompt Engineering Procedures

The main methodological approach involved Progressive-Hint Prompting (PHP) (Zheng et al., 2023), an iterative technique in which a general instruction is followed by more specific prompts. This approach allowed the researcher to guide the AI step by step toward the intended linguistic and contextual features. PHP was applied to shape key aspects of listening script generation, including the CEFR reference level, genre, topic area, and selected spoken language features. All interactions with ChatGPT-4 took place within a continuous conversational setting, enabling each script and set of items to be developed in a single context and refined gradually based on the characteristics of the generated output.

2.2. Listening Script Generation

A total of 20 English listening scripts was produced, distributed across four CEFR reference levels (A2, B1, B2, and C1), with five scripts developed for each level. These CEFR levels served as reference points during script development, guiding decisions related to vocabulary range, syntactic complexity, discourse density, and overall text length.

To ensure topical breadth and relevance to listening assessment contexts, script topics were systematically selected from 24 domains commonly represented in

DOI: https://doi.org/v9i1.6207

proficiency-oriented listening tests (Aryadoust, et al., 2024). During script generation, the prompting environment specified the CEFR reference level, target length (ranging from 250 to 400 words), genre, and topic, allowing for consistent control of content features while supporting iterative refinement.

2.3. Prompt Fine-tuning for Spoken Language Features

To enhance the authenticity of the generated scripts, a specific step was designed to refine the prompt with the aim of incorporating features typical of spoken language. These features included the strategic use of disfluencies (e.g., uh, um, you know), discourse markers (e.g., anyway, right), repetitions, reformulations, parenthetical clauses, and turn-taking patterns typical of spoken language (Clark & Fox Tree, 2002). The aim of this step was not to model spontaneous speech in all its complexity but to approximate authentic listening conditions typical of academic settings. Following script generation, each script was reviewed by the researcher using a set of predefined criteria (Field, 2008; Taylor & Geranpayeh, 2011). The evaluation focused on three main aspects: (1) the appropriateness of spoken language features in relation to the target CEFR level, (2) the balance between naturalness and clarity, and (3) the overall coherence of the script as a listening text. A script was considered acceptable when it met the following conditions:

a) spoken features were present but not excessive,
b) the inclusion of disfluencies and discourse markers did not reduce comprehensibility, and
c) the script maintained a clear and coherent flow suitable for listening comprehension tasks.

If a script did not meet these criteria, follow-up prompts were used to revise specific aspects of the text. For example, prompts were adjusted to reduce the number of disfluencies, simplify sentence structures, or improve clarity while maintaining naturalness. This process was repeated until the script met the predefined criteria. This iterative evaluation and revision process was applied consistently across all scripts, allowing the development procedure to remain systematic and replicable.

2.4. Item Writing Procedure

Following the finalization of each listening script, ChatGPT-4 was prompted to generate four multiple-choice comprehension items per script. Items were designed to target core academic listening skills, including identification of main ideas, understanding speaker purpose, extraction of specific information, and inferential comprehension, in line with established listening assessment frameworks (Sawaki et al., 2009; Taylor & Geranpayeh, 2011).

The item prompts provided explicit instructions to the AI to produce questions that are based on the CEFR levels of difficulty and that contain distractors consisting only of the information provided in the listening text. Progressive Hint Prompting was utilized to improve the items and distractors, especially to reduce the need for general world knowledge and increase the plausibility while ensuring a single correct answer. The plausibility of distractors was not evaluated solely based on their surface appearance. Instead, each option was reviewed by the researcher to determine whether it could

DOI: https://doi.org/v9i1.6207

reasonably be selected by test takers based on partial or misinterpreted information from the script, while still remaining unsupported as the correct answer. When distractors were found to rely on general knowledge or lacked a clear connection to the text, they were revised through follow-up prompts.

2.5. Linguistic and Topical Analyses

The linguistic features of the AI-generated listening scripts were analyzed using a combination of quantitative and qualitative approaches within an exploratory framework. Text Inspector was used to obtain quantitative indicators, including total word count, average sentence length, readability levels, and CEFR-referenced lexical profiles. These measures provided an initial overview of how the scripts varied across different CEFR reference levels. In addition to these quantitative indicators, a qualitative, expert-informed review was conducted to examine discourse-level features and item quality. This review focused on aspects such as coherence, clarity, the appropriateness of spoken language features, and the plausibility of distractors in the multiple-choice items. The evaluation was guided by predefined criteria derived from established principles in listening assessment.

The combination of quantitative linguistic profiling and qualitative review allowed for a more comprehensive analysis of the generated materials. While this approach does not constitute a formal triangulation design, it provides a complementary analytical perspective suitable for an exploratory study focused on early-stage material development.

3. Results and Discussions

3.1. The Utilization of ChatGPT-4 to Develop CEFR-Referenced English Listening Scripts and Test Items

This section addresses the first research question by examining how ChatGPT-4 was used in the development of CEFR-referenced English listening scripts and corresponding test items through a structured, human-guided prompt engineering process. The structured application of the PHP approach proved effective in supporting the generation of a corpus of 20 English listening scripts calibrated to predefined difficulty targets (Zheng et al., 2023). Importantly, ChatGPT-4 was not treated as an autonomous test developer, but rather as a generative drafting tool whose outputs were continuously shaped, evaluated, and refined by the researcher.

The process of utilization started with generalized prompts that identified the communicative context, genre, length, and reference CEFR level of each script. For instance, a prompt such as “American English listening script, reference CEFR level A2, informal conversation regarding weekend travel plans, and around 250 words” was used to initiate the process of script generation. At this stage, the model generally created a script that was coherent and broadly matched to the identified context. However, there was a need to refine the script to make it more comparable to the identified level. The use of iterative prompting enabled the development process to be broken down into smaller, more controlled steps. This helped maintain consistency across scripts while reducing variability in output (Wei et al., 2022). However, the need for repeated

DOI: https://doi.org/v9i1.6207

refinement also indicates that alignment with CEFR levels did not occur automatically, but depended on careful human intervention. This suggests that AI-generated content does not inherently reflect target proficiency levels and must be guided through structured input and evaluation.

The prompt engineering strategy was similarly employed in the development of the multiple-choice listening comprehension questions. For the listening scripts, four questions were generated using ChatGPT-4, which focused on the basic listening skills required for academic listening, such as the identification of the main idea, the speaker's purpose, specific information, and inferential comprehension, based on the listening frameworks (Sawaki et al., 2009; Taylor & Geranpayeh, 2011).

During the review of the initial items, several recurring problems were observed. In some cases, distractors were clearly incorrect and could be easily eliminated without careful listening. In other cases, distractors appeared reasonable but were based on general background knowledge rather than information provided in the script. An example is shown in Table 1, where some answer options could be chosen based on common assumptions rather than what the speaker explicitly stated.

Table 1. Distractor efficiency

Item	Question	Options	Answer	Issue
LS_B2_01_Q1	According to the speaker, what is the primary benefit of investing in renewable energy?	A) It creates many job opportunities. B) It reduces greenhouse gas emissions. C) It makes countries energy independent. D) It is cheaper than fossil fuels in the long run.	B	Option A may seem reasonable based on general knowledge, even though it is not clearly stated as the main benefit in the script. Distractor D could also be a generally accepted fact. Careful refinement is needed to ensure distractors are plausible but incorrect based only on the listening text.

This issue can be interpreted in relation to established principles of item writing, particularly the concept of construct-irrelevant variation. When distractors are based on general knowledge instead of the listening input, they may introduce additional cognitive demands that are not directly related to the listening construct. As a result, test takers may rely on background knowledge or test-wise strategies rather than actual listening comprehension. This highlights the importance of controlling distractor design to ensure that items accurately reflect the intended construct. Overall, these findings suggest that while generative AI can support the early stages of test development, the quality of assessment materials remains dependent on human judgment. The development process should therefore be understood as a collaborative interaction

DOI: https://doi.org/v9i1.6207

between AI-generated output and expert evaluation, rather than as an automated procedure (Nasr et. al., 2025).

3.2. The Linguistic Quality and Topical Diversity of ChatGPT-4 Outputs

To address the second research question, the linguistic quality and diversity of the generated listening materials were analyzed using both quantitative and qualitative approaches. A linguistic analysis of the listening materials using Text Inspector showed that there was a discernible pattern of variation in script characteristics across the CEFR reference levels. As can be seen in Table 2, scripts generated for higher CEFR reference levels tended to have longer overall length and lower readability scores.

Table 2.
Descriptive Linguistic Indicators of Scripts across CEFR Reference Levels

CEFR Reference Level	Avg. Word Count	Readability Trend	Lexical Profile
A2	250	High	Predominantly A2–B1 vocabulary
B1	300	High-Moderate	B1 with emerging B2 items
B2	350	Moderate	Increased B2–C1 items
C1	400	Lower	Predominantly C1 vocabulary

The lexical analysis showed clear differences across CEFR reference levels. Scripts developed for lower CEFR levels mainly used high-frequency, general-purpose vocabulary, which is commonly linked to easier processing for less proficient listeners. In contrast, scripts developed for higher CEFR reference levels contained a larger number of B2–C1 vocabulary items, reflecting more specific and abstract word use. These observations indicate that CEFR descriptors provided useful guidance for shaping lexical features during script development, even though difficulty was not tested empirically at this stage.

Regarding discourse features, spoken elements such as fillers and hesitations were included to reflect characteristics of spoken language. However, the analysis showed that the use of disfluencies was not always well balanced. In some scripts, especially those developed for lower reference levels, disfluencies appeared too frequently and had the potential to reduce message clarity. This observation emerged directly from the analysis and suggests that while generative AI can imitate features of spoken discourse, human review after generation is still needed to adjust these features so that they support naturalness without reducing comprehensibility (Clark & Fox Tree, 2002). This limitation reflects the broader constraint of generative AI in simulating spoken discourse, where surface-level features can be reproduced, but their functional use in assessment contexts still requires expert judgment.

DOI: https://doi.org/v9i1.6207

Scripts developed with reference to higher CEFR levels, especially C1, showed several issues related to how natural the language sounded in use. While the sentence structures often reflected advanced language use, some parts sounded awkward or too compact when read as spoken discourse. This suggests that increased linguistic complexity does not always lead to more authentic or effective listening input. From a construct perspective, this indicates that higher-level listening tasks require not only more complex language, but also more appropriate discourse organization and pragmatic use (Mead & Zhou, 2023).

In terms of topic coverage, the generated scripts included a range of academic and everyday themes that matched the study's design goals. The topics reflected areas commonly found in listening assessments, showing that ChatGPT-4 can produce materials across different domains. However, closer review showed that some topics appeared more frequently than others, particularly general social issues and widely discussed scientific themes. Topics involving specialized terminology or specific cultural contexts were less developed. This observation indicates that topic variety does not occur automatically and needs to be guided through careful prompt design. Future work may improve balance across topics by providing more detailed topic guidance or using external reference materials during script development.

Overall, these findings indicate that AI-generated listening materials can approximate CEFR-referenced difficulty levels, but their effectiveness depends on how well linguistic features, discourse characteristics, and item design are aligned with the underlying listening construct. This reinforces the idea that AI can support material development, but cannot replace the role of expert judgment in ensuring assessment quality.

4. Conclusion

This study aims to examine the potential use of generative AI model, ChatGPT-4, in the early-stage development of CEFR-referenced English listening scripts and test items. The focus was on how the technology can be effectively integrated into assessment-oriented processes. The findings show that ChatGPT-4 can be used as a supportive drafting tool in the development of English listening scripts and test items if appropriately guided by the use of structures prompt engineering and continuous human supervision.

From a linguistic perspective, the generated scripts showed a clear pattern of variation across CEFR reference levels, suggesting that CEFR descriptors can serve as potential basis for controlling input level when developing materials. At the same time, the analysis revealed clear limitations, particularly in the calibration of spoken discourse features, pragmatic naturalness at higher levels, and balanced topical coverage. In particular, while generative AI was able to imitate features of spoken language, these were not always used in a way that supports listening comprehension. This indicates that such features still require careful human adjustment to maintain both naturalness and clarity. These results showed that the quality of AI-generated listening materials emerges through iterative human–AI interaction, not through automated generation alone.

The significance of this study lies in its practical implications for educational contexts with limited resources. By clarifying both the potential and the boundaries of generative AI in listening material development, this research offers a realistic pathway for institutions seeking more accessible and sustainable assessment practices. Importantly, the study adopts an empathetic stance toward assessment practitioners, recognizing that AI is most valuable when it supports, rather than replaces, professional expertise and judgment. Future research should build on this foundation by conducting empirical validation with test taker data to determine how AI-assisted materials function in operational assessment settings.

Acknowledgement

The authors would like to express sincere gratitude to the Institute for Research and Community Service (LPPM) of Universitas Singaperbangsa Karawang for the institutional support provided in facilitating this research.

References

Aryadoust, V., & Luo, L. (2023). The typology of second language listening constructs: A systematic review. *Language Testing*, 40(2), 375–409. https://doi.org/10.1177/02655322221126604

Aryadoust, V., Zakaria, A., & Jia, Y. (2024). Investigating the affordances of OpenAI's large language model in developing listening assessments. *Computers and Education:Artificial Intelligence*, 6, 100204. https://doi.org/10.1016/j.caeai.2024.100204

Clark, H. H., & Fox Tree, J. E. (2002). Using uh and um in spontaneous dialog. *Cognition*, 84(1), 73–111. https://doi.org/10.1016/S0010-0277(02)00017-3

Coleman, H., Ahmad, N. F., Hadisantosa, N., Kuchah, K., Lamb, M., & Waskita, D. (2024). Common sense and resistance: EMI policy and practice in Indonesian universities. *Current Issues in Language Planning*, 25(1), 23–44. https://doi.org/10.1080/14664208.2023.2205792

Field, J. (2008). *Listening in the language classroom*. Cambridge University Press. Jiang, Y., et al. (2024). Evaluating the critical thinking of large language models:

Insights and limitations. *Journal of Pacific Rim Psychology*. https://doi.org/10.1177/18344909251406111

McKinley, J., & Rose, H. (2017). *The Routledge handbook of English language teaching*. Routledge.

Mead, A. D., & Zhou, C. (2023). Evaluating the quality of AI-generated items for a certification exam. *Journal of Applied Testing Technology*, 24(Special Issue), 1–14.

Nasr, N. R., Tu, C.-H., Werner, J., Bauer, T., Yen, C.-J., & Sujo-Montes, L. (2025). Exploring the impact of generative AI ChatGPT on critical thinking in higher education: Passive AI-directed use or human–AI supported collaboration?

Education Sciences, 15(9), 1198. https://doi.org/10.3390/educsci15091198

Nurhayati, N., Setiawaty, P. W., & Nur, S. (2024). EFL teachers' challenges in designing assessment material for students' listening skills. *English Franca: Academic Journal of English Language and Education*, 8(2), 409–422. https://doi.org/10.29240/ef.v8i2.12053

OpenAI. (2023). *ChatGPT (GPT-4 version)* [Large language model]. https://chat.openai.com

Peláez-Sánchez, I. C., Velarde-Camaqui, D., & Glasserman-Morales, L. D. (2024). The impact of large language models on higher education: Exploring the connection between AI and Education 4.0. *Frontiers in Education*, 9, 1392091. https://doi.org/10.3389/feduc.2024.1392091

Richardson, A. (2022). Advances in OpenAI's GPT-3 applications. International *Journal of Artificial Intelligence and Machine Learning in Engineering*, 21(1), 742–749.

Sawaki, Y., Kim, H.-J., & Gentile, C. (2009). Q-matrix construction: Defining the link between constructs and test items in large-scale reading and listening comprehension assessments. *Language Assessment Quarterly*, 6(3), 190–209. https://doi.org/10.1080/15434300902801917

Sung, H., Chang, T., & Huang, J. (2015). Factors affecting item difficulty in English listening comprehension tests. *Universal Journal of Educational Research*, 3(7), 451–459. https://doi.org/10.13189/ujer.2015.030704

Taylor, L., & Geranpayeh, A. (2011). Assessing listening for academic purposes: Defining and operationalising the test construct. *Journal of English for Academic Purposes*, 10(2), 89–101. https://doi.org/10.1016/j.jeap.2011.03.002

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS 2022)*. https://arxiv.org/abs/2201.11903

Zheng, C., Liu, Z., Xie, E., Li, Z., & Li, Y. (2023). Progressive-hint prompting improves reasoning in large language models. *arXiv*. https://arxiv.org/abs/2304.09797