Using ChatGPT to Evaluate the Methodological Components of Research Proposals: An Experimental Study on Undergraduate English Majors in Vietnam

Tran The Phi[1], Nguyen Thi Hoang Lan[1*]

[1] Faculty of Foreign Languages, Saigon University, Vietnam
*Corresponding author's email: nthlan@sgu.edu.vn
https://orcid.org/0009-0009-8821-0806
https://doi.org/10.54855/callej.252637

Received: 06/10/2024 Revision: 16/06/2025 Accepted: 01/07/2025 Online: 02/07/2025

ABSTRACT

Artificial Intelligence is increasingly applied in education, but its effectiveness in evaluating research methodologies remains underexplored. This study examines the intra- and inter-rater reliability of ChatGPT-4o, employing zero-shot learning, in assessing 37 research proposals from English majors at Saigon University, Vietnam, focusing on Research Title, Questions, Hypotheses, Paradigm, Design, and Techniques. A quantitative quasi-experimental design was used, with two evaluation groups: Control (module lecturers) and Experimental (ChatGPT-4o). ChatGPT-4o followed a structured zero-shot prompt set, with a researcher-designed five-point rubric and the *How to Research* book uploaded for reference to evaluate each proposal twice. The lecturers evaluated proposals independently, discussed and finalized scores. Data collected were analyzed using Quadratic Cohen's weighted Kappa. Results showed moderate to high intra-rater reliability and moderate inter-rater reliability in straightforward areas, but the machine struggled with abstract criteria requiring deeper reasoning, such as evaluating title relevance and the justification of paradigm and design. These findings highlight the limitations of AI in fully capturing the complexities of research methodologies. However, ChatGPT-4o may be a reliable tool in contexts with clear rubrics and minimal training, reducing the need for human intervention. Future studies should expand the sample size and explore different approaches to improve its ability in research evaluation.

Keywords: research proposals, evaluation, ChatGPT, Zero-shot learning, Methodological Components

Introduction

Evaluating research is a crucial process in ensuring academic rigor, with peer review being the most prominent despite criticisms regarding biases, conservatism, reviewer fatigue, and limited individual expertise (Lee et al., 2013; Spaapen et al., 2007). These shortcomings underscore the

CITATION | Tran, T. P., & Nguyen, T. H. L. (2025). Using ChatGPT to Evaluate the Methodological Components of Research Proposals: An Experimental Study on Undergraduate English Majors in Vietnam. *Computer-Assisted Language Learning Electronic Journal (CALL-EJ),* 26(3), 129-227. DOI: https://doi.org/10.54855/callej.252637

need for an indefatigable, consistent, impartial method that can handle the heterogeneous knowledge beyond the capacity of any single assessor.

With technological improvements, AI has advanced research evaluation through tasks like plagiarism detection, formatting check, and statistical and examination of methodological transparency (Checco et al., 2021; Kousha & Thelwall, 2022; Lin et al., 2023). Recent research further used ChatGPT's zero-shot learning – an approach letting a model perform tasks using instructions and pre-trained knowledge without prior examples – to assess research quality (Thelwall, 2024). However, a comprehensive procedure for evaluating the soundness of methodological components in relation to specific research objectives remains insufficiently explored, primarily due to the advanced logical reasoning required (Lin et al., 2023).

This study aims to address this gap by investigating ChatGPT-4o's evaluation of specific methodological components – Paradigm, Design, and Techniques – and their alignment with the Research Title, Questions, and Hypotheses using a predefined rubric. These components require evaluative judgments that go beyond mere factual recall, demanding advanced logical reasoning and contextual understanding, which machine evaluation has not yet demonstrated.

To assess ChatGPT's capacity as an evaluator, its reliability, specifically, intra-rater and inter-rater reliability must first be established (Gwet, 2008; H. D. Brown, 2018). Intra-rater reliability refers to the evaluator's self-consistency across instances, while inter-rater reliability examines whether two or more raters yield consistent scores for the same paper (Gwet, 2008; H. D. Brown, 2018). In this research, ChatGPT's intra-rater reliability will be examined by comparing its evaluations of the same paper at different times, while inter-rater reliability compares ChatGPT's evaluations with those of human raters - the current standard in research assessment.

In Vietnam context, while AI is increasingly used in English education (Pham & Cao, 2025) and has demonstrated potential in supporting not only reading and writing skills (Duong, Tong, & Le, 2024; Duong & Le, 2024; Hoang & Vu, 2024), but also listening (Luu & Doan, 2025) and pronunciation (Nguyen et al., 2025), its ability to perform evaluative tasks remains underexplored. Therefore, the researchers selected Saigon University, Vietnam with its interdisciplinary programs, compliance with Ministry of Education standards, and supportive setting for research (Saigon University, 2018) as the research site.

Literature review

ChatGPT in education

Overview of AI and ChatGPT

According to Russell and Norvig (2020), Artificial Intelligence (AI) encompasses the design of systems that exhibit intelligence by understanding and reasoning about the world through input. These systems utilize techniques such as machine learning, knowledge representation, and problem-solving algorithms to simulate human cognitive functions. Recent advancements in AI and Natural Language Processing (NLP), including ChatGPT, have enabled systems to comprehend and generate human-like text, suggesting that computers are reaching human-like intelligence and capabilities.

ChatGPT is a large-scale transformer-based language model capable of producing original content (Chaudhary & Gupta, 2023; Sabzalieva & Valentini, 2023). It is pre-trained on various large datasets before being fine-tuned using Reinforcement Learning from Human Feedback and refined through Proximal Policy Optimization (OpenAI, 2022). As the model learns from prompts through chat-like environment, small changes in input can significantly affect its responses, highlighting the importance of prompt design (OpenAI, 2022; Thelwall, 2024). In

addition, ChatGPT's versatility enables it to perform complex tasks across various fields with human-like cognitive abilities (T. B. Brown et al., 2020; Chaudhary & Gupta, 2023; Sabzalieva & Valentini, 2023).

ChatGPT-4o, known as "omni", is the latest model as of May 2024, with the ability to integrate text analysis with audio and visual inputs (OpenAI, 2024a). Pre-trained on data up to October 2023 (OpenAI, 2024a), it performs on par with ChatGPT-4 Turbo (large multimodal model with human-level performance) (OpenAI, 2024b) in English text comprehension, reasoning, and coding tasks. It also outperforms ChatGPT-4 Turbo and various AI models on almost all domains of text evaluation (OpenAI, 2024a). This implies that ChatGPT-4o is well-suited for tasks involving high-level reasoning and text analysis, such as research paper evaluation.

AI and ChatGPT in research evaluation

AI in text evaluation

The advancements in AI-powered tools for Automated Essay Scoring (AES) and Automatic Writing Evaluation (AWE) (i.e., evaluating syntax, complexity, and vocabulary range based on databases) have allowed applications like Criterion and Write&Improve to consistently assess writing with human, albeit with some exceptions (Hockly, 2019). Users also praised these tools for saving time in the writing and editing process (Heriyawati & Romadhon, 2025).

Empirically, NLP models (i.e., BERT, XLNet) have been demonstrated to outperform human raters on the Kaggle AES dataset with transformer-based approaches surpassing traditional methods like Bag-of-Words and long short-term memory networks in accuracy and efficiency (Rodriguez et al., 2019). Ormerod et al. (2021) later found that smaller NLP models with fewer parameters can be fine-tuned and combined to outperform larger ones. Studies indicated that fine-tuning NLP models, such as ChatGPT, with domain-specific datasets (student response data and scoring rubrics) can create accurate, fast, reliable, and scalable automatic scoring systems (Latif & Zhai, 2023), even with zero-shot prompting (Wang & Gayed, 2024).

The aforementioned research underscores AI's competence in evaluating creative writing, with fine-tuning using domain-specific datasets significantly improving the reliability and accuracy of the evaluations, enabling performance that can rival that of humans. This highlights AI's potential to assess more complex forms of writing, such as research manuscripts.

AI in research evaluation

In fact, various AI tools have been developed for scientific manuscript review process. As listed by Kousha and Thelwall (2022), these tools can assist in initial screening tasks such as plagiarism detection (e.g., iThenticate), statistical data cross-checking (e.g., StatReviewer), and reference checking (e.g., Recite). Notably, a tool developed by Menke, Roelandse, Ozyurt, Martone, and Bandrowski, called SciScore, can even extract and evaluate the rigor and transparency of medical science methods using criteria such as blinding, randomization, power analysis, and resource identifiers (Menke et al., 2020). A more comprehensive concept of research paper evaluation, Automated Scholarly Paper Review (ASPR), was introduced by Lin et al. (2023). The ASPR pipeline integrates AI-powered tools for format examination, plagiarism detection, machine-generated content detection, article type recognition, scope evaluation, as well as assessments of originality, quality, clarity, and significance.

However, there is a noticeable scarcity of instruments that critically evaluate whether the methodology can achieve specific research objectives because current tools mostly focus on transparency and reproducibility. Lin et al. (2023) emphasized that this gap remains a

significant challenge since evaluating research methodology requires advanced logical reasoning capabilities – a feature that ASPR must evolve to achieve in the future.

ChatGPT in research evaluation

ChatGPT supports many research tasks in reviewing studies. Although fine-tuning LLMs is usually popular for task alignment (Wang & Gayed, 2024), Kojima et al. (2023) suggested that with well-designed prompts, models can perform high-level reasoning across diverse tasks with zero-shot prompting as well, sparking interest in this fast and flexible approach.

One study by Syriani et al. (2023) created a universal prompt for article screening by manually crafting and refining the prompt before automating the process using an API-based approach (dataset sampling and hyperparameter tuning). There are three components in the prompt: Context (explaining the purpose and focus), Instructions (describing the task), and Task input (providing the article title and abstract). It was found that ChatGPT could yield article screening results comparable to traditional classifiers without being retrained. The results were reliable across multiple test runs and datasets. Although it cannot completely replace manual screening, it reduces workload for staff and enhances systematic review efficiency. In addition, evidence showed that effective prompting can enhance ChatGPT's screening ability without additional training.

In another study, Liang et al. (2023) extracted paper content from PDFs and constructed specific prompts for ChatGPT-4 using zero-shot learning. The prompts were iteratively refined to ensure detailed, constructive, and multi-point feedback, and incorporated the paper's title, abstract, figure and table captions, and main text. The researchers also instructed it to be as "*specific and as detailed as possible*". In a single operation, ChatGPT-4 generated feedback with justifications on the significance and novelty, reasons for acceptance, reasons for rejection, and suggestions for improvement. The results showed that the overlap in the scores by ChatGPT-4 and by human reviewers slightly exceeded the overlap between two human reviewers, suggesting a high consistency with expert feedback. Even so, it had certain limitations including the lack of deep subject expertise and nuanced contextual understanding, hindering its evaluation ability.

More recently, Thelwall (2024) designed the customized chatbot ChatGPT 4.0 REF D using zero-shot learning, meaning that the chatbot was not trained on any predefined "correct" scores or labeled data for evaluation. The configuration of the chatbot included detailed scoring criteria from UK REF Main Panel D 2021 (Research Excellence Framework), but the prompt was minimal ("score this"), leaving ChatGPT to interpret the assessment process independently. It evaluated research articles on their rigour, originality, and significance on a scale from 1 to 4 by running 15 evaluations independently. Even though ChatGPT-4 was able to generate structured evaluations and plausible justifications, it assigned different ratings to the same paper across different attempts, suggesting low intra-rater reliability. In addition, its individual scores showed weak correlations with the author's self-assessments, indicating low inter-rater reliability. Moreover, the model displayed a strong bias toward 3-star ratings and never used the lowest score (1). It also scored a fabricated paper involving "squirrel surgeons" as 4 stars, but later stated that squirrels could not write research papers, showing its inability to harness its own wider knowledge to critically evaluate research. It appears that ChatGPT-4 can assist in summarizing research, yet is not fully reliable for quality assessments without human oversight. However, giving more detailed prompts that explicitly guide the evaluation process could potentially improve its accuracy and consistency.

While studies like Syriani et al. (2023) and Liang et al. (2023) demonstrated how ChatGPT

supports systematic scientific paper evaluation, they also flag that the system needs specialized domain knowledge to provide critical assessments effectively. As the model appears sensitive to the clarity and depth of the instructions given, incorporating detailed prompts and materials providing domain-specific knowledge, such as rubrics, would serve as fundamental strategies to address the existing gap.

Methodological components in research proposals

Overview of research proposals

According to Creswell (2015), research is a process of collecting and analyzing information to increase the understanding of an issue. This consists of posing questions, collecting data and presenting answers to those questions. However, writing a thesis or dissertation begins with a proposal, identifying the investigated subject, and methods used, proving their appropriateness (Locke et al., 2007). Although formats vary, proposals should meet the formatting and communication expectations of the submitting institution (Denscombe, 2020; Locke et al., 2007). In this study, undergraduates' proposals followed the format required by the module. It is adapted from a simple, traditional thesis and consists of the Introduction, Literature Review, Methodology, Proposed Chapter Outline, and Timeline.

The Methodology chapter and Methodological components in a research proposal

The Methodology chapter often includes not only the *Research Design, Setting, Informants and Texts,* and *Data collection and Analysis* (Paltridge & Starfield, 2020) but also the *research aims/questions.* The term Methodology usually refers to the paradigm that underpins the whole research (Blaxter et al., 2010), and this ensures the credibility and trustworthiness of the results (Paltridge & Starfield, 2020). In journal articles, this section is often compressed into a "Methods" section (Swales, 2004, p. 86, as cited in Paltridge & Starfield, 2020). However, to comprehensively address the research paradigm, procedure, and rationale, the current research adopts the term Methodology.

In most cases, it consists of the population description, sampling method, research design, techniques and procedures for data collection, data analysis technique, and contingency plans in case of problems (Locke et al., 2007). However, the items discussed in this study are research design (termed as Design) and data collection techniques (termed as Techniques).

Evaluating the research proposal's methodological components

To properly evaluate the Methodology section, research title is the first component to consider as it summarizes the main focus and purpose of the study (Thomas, 2003). A focused title reflects a well-defined research problem and guides the methodological approach (Tcherni-Buzzeo & Pyrczak, 2024; Thomas, 2003), offering insight into the appropriateness of the research problem and the efficacy of the proposed methodology. Subsequently, the research questions or hypotheses and the justifications for the Methodology employed should be evaluated (Paltridge & Starfield, 2020), especially in terms of the logic of arguments, the quality of the research questions or hypotheses, and the alignment of the methodology with these inquiries (Cadman, 2002 as cited in Paltridge & Starfield, 2020; Creswell, 2015). Consequently, examining the research title before delving into the research questions or hypotheses allows assessors to check if the proposed methodology is in line with the overarching objectives of the research.

Evaluating Research Title

According to Blaxter et al. (2010), titles need to be as short as possible and should help focus the subsequent work. A typical title is concise, consisting of about 10 to 15 words, and names

the major variable(s) (Tcherni-Buzzeo & Pyrczak, 2024), announces the topic and communicates the research's conceptual framework (Turabian et al., 2018). Tcherni-Buzzeo and Pyrczak (2024) have put forth a list of questions detailing specific features necessary for a good research title. These questions can be answered either by using a dichotomous scale (Yes/No) or a 5-point scale ranging from 1 (Very unsatisfactory) to 5 (Very satisfactory). Applying this rubric, the present study evaluates research titles based on ten key criteria, ensuring that they meet academic and methodological standards. These criteria assess the title's specificity, clarity, conciseness, objectivity, inclusion of key variables, and overall alignment with the study's conceptual framework.

Evaluating Research Questions and Hypotheses

A research problem is usually initially posed as a guiding question that clarifies the research type and paradigm (qualitative, quantitative, or mixed method) (Fraenkel et al., 2023). A crucial feature affecting the design of the study is research questions' feasibility (Andrews, 2003; L. Cohen et al., 2018). This means it can be investigated within the resource constraints (Fraenkel et al., 2023).

Specifying participants, the research site, and focusing on a single concept make it easier to review the project's feasibility (Creswell, 2015). In addition, the question must be clear with all terms easily explained and defined for measurement (Fraenkel et al., 2023). The significance of questions is important as it determines whether the questions contribute valuable knowledge, while ethics ensure that they do not bring any harm to the subjects (Fraenkel et al., 2023). In addition, Creswell and Creswell (2018) highlighted that the research title and the research questions must align closely to ensure clarity and relevance for the audience.

If it suits, the research questions might be presented in the form of hypotheses, which serve to predict the outcomes, especially in quantitative studies (Creswell, 2015). To ensure researchers can design proper methods, a hypothesis should be precise and specific by focusing on a single topic, specifying the participants and research sites. In addition, the hypothesis should consist of variables that can be accurately defined and measured so that it can be empirically tested (Andrews, 2003; Blaxter et al., 2010).

Building on these theoretical foundations, the evaluation of both research questions and hypotheses in this research is guided by a set of analytical criteria. First, feasibility is determined by assessing whether the research questions can be answered and the hypotheses can be tested within the constraints of time, budget, and available resources specified by the module lecturers. Secondly, the research questions and hypotheses ought to be specific, clearly identifying the participants, research site, and the central phenomenon or concept. Equally important is clarity, which ensures that the terms used are precise, understandable, and measurable. Moreover, the research questions/hypotheses must show that the research offer valuable contribution to English language education, posing no risks to participants, and show strong alignment with the title.

Evaluating Research Design

In this paper, the research design follows the framework outlined in *How to Research* by Blaxter, Hughes, and Tight (2010). It refers to the plan for a study, used as a guide in collecting and analyzing data. The four basic designs for social sciences are action research, case studies, experiments, and surveys. First, it is imperative that the research design aligns with the research questions (Tcherni-Buzzeo & Pyrczak, 2024). The specification, justification for the chosen design, and its effectiveness in collecting data that can address the research questions or hypotheses are to be examined. A scale of 1 to 5, ranging from "Very Unsatisfactory" to "Very

Satisfactory" (Tcherni-Buzzeo & Pyrczak, 2024), is adopted for the evaluation. Since research design is guided by paradigms (quantitative, qualitative, or mixed) (Blaxter et al., 2010), we must also evaluate the paradigm to better examine how the design aligns with the research problem. In line with the research design, research paradigm is assessed in three aspects: its specification, justification and effectiveness in generating data for the study. *How to Research* by Blaxter, Hughes, and Tight (2010), the Research Methods module's coursebook, serves as the main guide for evaluation.

Evaluating Research Techniques

The present research only sets out to evaluate the four basic techniques listed in the work of Blaxter et al. (2010), namely, document study, interviews, observations, and questionnaires. Similar to the evaluation of design and paradigm, it is also crucial to identify whether the techniques align with the research questions (Tcherni-Buzzeo & Pyrczak, 2024). The specification of these techniques and the justification for their selection should be examined on a scale of 1 to 5. In the same vein, the present research also aims to evaluate whether the techniques are effective in collecting data required to address the research questions or hypotheses.

Research Questions

The primary objective is to assess whether ChatGPT-4o can serve as a viable alternative to human evaluators. To achieve this, it is essential to first examine the reliability of ChatGPT-4o as an assessment tool. Thus, this study will focus on measuring the reliability of ChatGPT-4o by answering the following questions:

1. What is the level of intra-rater reliability demonstrated by ChatGPT-4o, using zero-shot learning, when evaluating proposals' Research Titles, Questions, Hypotheses, Paradigm, Design, and Techniques using the same rubric across different instances?
2. What is the level of inter-rater reliability demonstrated by ChatGPT-4o, using zero-shot learning, in comparison to human raters (lecturers) when evaluating proposals' Research Titles, Questions, Hypotheses, Paradigm, Design, and Techniques using the same rubric?

Methods

Research objectives and design

This study measures the reliability of ChatGPT-4o in evaluating research proposals by English majors at Saigon University using the same rubric and resources. Specifically, it identifies discrepancies between evaluations generated by ChatGPT-4o itself (Intra-rater reliability) as well as those between ChatGPT-4o and lecturers of Research Method module, across components like Research Title, Questions/Hypotheses, Paradigm, Design, and Techniques. A quantitative Quasi-experimental design is adopted to assess whether a specific intervention (ChatGPT) influences an outcome (evaluation of research proposals) (Creswell & Creswell, 2018). This design enables comparison between evaluation scores from two groups: the Control group (lecturers) and the Experimental group (ChatGPT-4o).

Research participants and procedures

The experiment itself comprises three stages: preparation, implementation, and data collection. These stages are repeated for the two groups (Control and Experimental). The Control group refers to Research Methods module lecturers and the Experimental group refers ChatGPT-4o.

Stage 1: Preparation

In this stage, three resources are required for the evaluation are designed, namely, Scoring rubric, ChatGPT prompt and Student proposals.

Scoring rubric (Appendix A): An analytic rubric with a five-point scale is employed to evaluate the proposals in a structured and transparent framework. As it allows for separate evaluation of each criterion, the analytic format was chosen (Brookhart, 2013) to evaluate the Title, Questions, Hypotheses, Paradigm, Design and Techniques of each proposal. As these proposals were produced in partial fulfilment of the module *Research Methods*, this rubric is based mainly on the book *How to Research* by Blaxter, Hughes, and Tight because it is the module's coursebook. Supplementary documents mentioned above (Andrews, 2003; Creswell, 2015; Creswell & Creswell, 2018; Fraenkel et al., 2023; Tcherni-Buzzeo & Pyrczak, 2024) and lecturers' opinions were also incorporated to ensure that the rubric can address both practical and theoretical issues in research.

The scoring rubric (henceforth referred to as Rubric) includes 3 columns. The first column (Sections to evaluate) addresses the section title. The second (Evaluating questions) includes the questions that ought to be answered when evaluating. Column 3 (Evaluation scale) details performance descriptions of each point in the scale for each question, ranging from "Very Unsatisfactory" to "Very Satisfactory". This five-point scale was adopted to balance between detailed evaluation and usability (Brookhart, 2013). To ensure each point is distinct from one another, the description for each point was generated by ChatGPT-4o before being continuously refined through a continuum-based approach (Brookhart, 2013) with input from lecturers gathered via informal discussions and the researchers themselves.

In general, the Very Unsatisfactory (1 point) level indicates that major components either disappear completely (except for the case of Hypotheses) or are inappropriate. At the Unsatisfactory (2 points) level, relevant aspects can be found yet lack precision, details or completeness. The Neutral (3 points) level indicates that the specification and completeness exist at a minimal level and require improvement in precision, coherence, or depth. The Satisfactory (4 points) level demonstrates clear, well-structured, and effective performance but may lack extra depth or explanation. The Very Satisfactory (5 points) reflects a product that is not only clear, well-structured, and precise, but also well-supported and explained with strong evidence.

ChatGPT prompt (Appendix B):

According to Brown et al. (2020), adapting language representations in NLP systems to different tasks now does not require specific customization. Even though models like ChatGPT use large datasets to perform certain tasks, these big datasets are no longer a necessity in NLP training thanks to "zero-shot training", in which the model is expected to perform a task without any specific examples and rely solely on its pre-training knowledge to respond to natural language instructions (T. B. Brown et al., 2020). Even though this approach requires fewer resources than others (i.e., few-shot and one-shot), its products are comparable and even superior to those of other approaches (Reynolds & McDonell, 2021).

Reynolds and McDonell (2021) further looked into methods to make prompts more reliable in creating desired outcomes for a specific task. The prompts used in this paper were designed using d*irect task specification* and *behavior constraints*, which create metaprompts - seeds encapsulating a more general intention that will unfold into a specific prompt when combined with additional information, such as the task questions (Reynolds & McDonell, 2021). The metaprompt provides context to clarify the background and specifics of the task.

The prompt design process began with identifying the signifiers to guide the intended behaviors and actions of the model. These signifiers were inferred from the practices of human assessors during the evaluation process (the researchers themselves). After the basic prompt was designed, it was adjusted multiple times during the design process. However, only two versions were deployed on ChatGPT before the final version was produced. Initially, the first two versions of the prompt set (v.1 and v.2) did not include a role description of ChatGPT and demanded the Evaluation template in the form of downloadable Excel file, which added complications to the process.

In terms of language use, prompt set v.1 featured suggestive phrasing, such as " *Use the 10 questions from the "Evaluating Questions" column of the rubric*," which was open to interpretation. In the final version, this *became " Extract the verbatim of 10 questions from the "Evaluating Questions" column of the Rubric file (from T1 to T10). Avoid changing the Evaluating questions.*" explicitly prohibiting modifications to maintain rubric integrity.

Similarly, the handling of assumptions became stricter; while the first version simply stated "*Base your evaluation on the text in the proposal*," the final version reinforced this by stating, *"Your evaluation must be strictly based on evidence (text) from the proposal. You cannot make any assumption, implication, or change*." Additionally, the handling of missing hypotheses evolved from "*If the hypotheses are not clearly stated, mark as N/A,*" to "*If the hypotheses are NOT explicitly stated under a heading, mark this section as N/A*," reducing assumptions made on implied hypotheses.

These modifications were made based on results from pilot evaluations of 5 proposals on both versions 1 and 2 which indicated assumptions and modifications to the content of the proposals and Rubric as well as the extended time spent on generating the Excel file. The revision stopped when the evaluation from ChatGPT was no longer subject to assumptions, or unauthorized modifications.

At the conclusion of this process, a final set of prompts for ChatGPT, subsequently referred to as **Prompt** was designed. This Prompt states the primary responsibility of ChatGPT as proposal reviewer, while also outlining a specific requirement for strict adherence to the Rubric and prohibiting any changes to the content. The evaluation process involves the individual evaluations of 6 sections following a fixed sequence of actions:

1. Extract text: For each section (Research Title, Research Questions, Research Hypotheses, Research Paradigm, Research Design, Research Technique), identify and extract the exact text from the proposal.
2. Summarize theories: Summarize relevant theories from the book "How to Research".
3. Evaluate using rubric: Apply the specific evaluating questions from the Rubric to assess each section. Ensure that the questions are used exactly as provided, without alteration.
4. Compare with theories: Compare the text with theories from *How to Research* (for Paradigm, Design and Technique sections).
5. Assign scores: Assign scores using the descriptions in the "Evaluation Scale" column of the Rubric file.
6. Calculate average score: Compute the average score for each section.
7. Provide comments: For each evaluating question, include the extracted text and detailed comments on the evaluation.

8. Generate Evaluation template: Create a table with columns for Section, Details, Evaluating Questions, Score, and Comments. Include all necessary details and leave rows blank where appropriate.
9. Mark N/A: If hypotheses are not explicitly stated under a heading, mark the section as N/A.

Students' proposals:

Thirty-seven research proposals were collected from English majors as the final project for the Research Methods course at Saigon University, Vietnam. These proposals were developed based on specific guideline established by the Faculty of Foreign Languages and in this research, they were evaluated using a predefined rubric (Rubric). Permission to access and analyze the proposals was obtained from the Dean of the faculty, an author of this study. The proposals had been previously scored by the module lecturers, and all grades were finalized prior to the conduct of this research. This ensured that the study had no influence on the students' academic outcomes, and that there was no potential for harm or benefit to the students as a result of their participation in the research.

Stage 2: Implementation

Control Group Evaluation Procedure

1. Distribution of Proposals: the proposals were distributed to 2 lecturers responsible for the Research Methods module at Saigon University. One lecturer has a Ph.D. in Contrastive Linguistics and an M.A in TESOL while the other has an M.A in TESOL. They were both familiar with the module requirements. As they both took part in the design and had clear understanding of the Rubric, only brief training on how to record their evaluation in the Evaluation template was needed to ensure uniform score presentation.
2. Independent Evaluation: Each lecturer assessed the proposals separately to ensure objectivity, following the Rubric criteria.
3. Discrepancy Analysis: The lecturers' evaluation scores were compared for discrepancies.
4. Consensus Meeting: The lecturers had a face-to-face meeting to reconcile differences and agree on a final evaluation.
5. Finalization of Evaluation templates: The consensus scores for each proposal were recorded on a single Evaluation template.

Experimental Group Evaluation Procedure

1. Rubric input and evaluation question extraction: The Rubric was input into the ChatGPT-4o chat interface with the instructions to extract all Evaluating questions.
2. Proposal input and Evaluation: Following the extraction, the proposal document, the book How to Research, and the Prompt were uploaded.
3. Results recording and storage: The evaluation results were recorded using the paste function and compiled into Excel sheets.
4. Proposal re-evaluation: Each proposal underwent two evaluations by ChatGPT-4o using the Rubric. New entries were opened for different proposals. The first-round evaluation of 37 proposals was completed before the second round began. Due to

ChatGPT-4o's limitations on file uploads, the evaluations were conducted in batches rather than simultaneously.

5. Data collection: The results from the two evaluations of each proposal were recorded in the Evaluation template.

Stage 3: Data collection

All evaluations ended with the record of evaluation scores (1 to 5) for all Evaluating questions. The Control Group procedure concluded with the production of a single Evaluation template for each proposal. In contrast, the Experiment Group procedure involved two rounds of evaluation, with the results from each round recorded on separate Evaluation templates. To facilitate comparison across evaluations, the scores from each Evaluation template were compiled into a final form. This process resulted in 37 final forms, corresponding to the 37 proposals evaluated, each with three sets of scores.

Data analysis

To identify the intra- and inter-rater reliability of evaluations from ChatGPT-4o and human raters, we employed Cohen's weighted Kappa statistics. Weighted Kappa (J. Cohen, 1968) measures the reliability of ordinal scores while factoring in the severity of disagreements. It assigns larger penalties for greater degrees of disagreement (Maclure & Willet, 1987, as cited in Berry et al., 2008; Berry et al., 2008). For example, the penalty for 1 and 5 points is weighted more than for 4 and 5. Given our five-point scale Rubric, this statistical test can accurately capture the severity of disagreement. To further highlight the discrepancies between raters, we employed Quadratic Weighted Kappa statistics where the weights are increased quadratically (Sim & Wright, 2005) and categorized the scores based on Landis and Koch (1977).

The evaluating questions were labeled as follows: T1–T10 (Titles), Q1–Q10 (Questions), H1–H10 (Hypotheses), P1–P3 (Paradigm), D1–D3 (Design), and Te1–Te3 (Techniques). Moreover, "Human" refers to lecturers' evaluation scores, "GPT 1" to the first evaluation done by ChatGPT-4o and "GPT 2" to the second. We calculated the Quadratic Weighted Kappa to compare (1) GPT 1 with GPT 2, (2) Human with GPT 1, and (3) Human with GPT 2. In the Results section, **k** denotes individual Kappa value. We also calculated the average Kappa value for the Human and GPT evaluation scores for each item (denoted as **K**). However, no specific symbol is used to represent the overall average Kappa value across all items.

Results

Research Title

Overall, the average agreement between the two rounds of ChatGPT evaluations, representing intra-rater reliability, was fair, with a mean Kappa value of 0.354. However, the Kappa statistics for inter-rater reliability were notably lower, with mean values of 0.248 and 0.099.

Questions T3, T5, and T9 yielded invalid Kappa values due to identical scores across all assessments, resulting in 100% agreement. This suggests that human assessors and ChatGPT achieved complete intra-rater and inter-rater reliability when assessing whether a title was phrased as a yes-no question, whether it described the results, and whether the titles and whether the subtitles provided relevant information.

Table 1.

Comparison of Intra-Rater and Inter-Rater Reliability Between Human and ChatGPT Evaluations of Research Title

	GPT 1 vs GPT 2		Human vs GPT 1		Human vs GPT 2		K value average of Human and GPT evaluations
Items	**Weighted Kappa[a]**	**Sig.**	**Weighted Kappa[a]**	**Sig.**	**Weighted Kappa[a]**	**Sig.**	
T1	0.439	0.001	0.042	0.316	-0.052	0.337	-0.005
T2	0.358	0.005	0.340	0.001	0.219	0.025	0.280
T3	.e	.e	.e	.e	.e	.e	.e
T4	0.298	0.070	0.181	0.109	0.132	0.248	0.157
T5	.e	.e	.e	.e	.e	.e	.e
T6	0.317	0.011	0.484	0.000	0.230	0.067	0.357
T7	0.268	0.047	0.297	0.008	0.141	0.212	0.219
T8	0.294	0.048	0.299	0.011	-0.032	0.783	0.134
T9	.e	.e	.e	.e	.e	.e	.e
T10	0.501	0.000	0.092	0.162	0.054	0.336	0.073
Average T1 to T10	**0.354**		**0.248**		**0.099**		**0.173**

The agreement between Human and GPT 1 on T6 was the highest (k = 0.484), but the Human vs. GPT 2 and GPT 1 vs. GPT 2 comparisons yielded significantly lower results, indicating only a "fair" level of reliability in identifying key variables. T2 also demonstrated a "fair" level of reliability in evaluating the brevity of the titles, with all Kappa values between 0.2 and 0.4. T7's Kappa values displayed fair agreement, except for Human vs. GPT 2, which fell into the "slight" range, suggesting variability in GPT's evaluation of participants specification. Similarly, while a similar phenomenon was observed in T8, its Human vs. GPT 2 comparison exhibited a slight negative agreement (k = -0.032) with a high p-value (p = 0.783), suggesting that the titles' uniqueness may lead to varying interpretations by the machine.

Low inter-rater agreement was observed in T1 (K = -0.005), and in T10 (K = 0.073), with non-significant p-values, indicating that the specificity and relevance of the titles might be interpreted and evaluated differently across raters. However, moderate intra-rater reliability was observed for these questions, with statistically significant values, indicating that the machine evaluated these aspects consistently. Meanwhile, T4's inter-rater reliability was low (k < 0.2), and its intra-rater reliability was only fair (k = 0.298), indicating that the criteria involving jargon and acronyms varied in interpretation.

Research Questions

Table 2.

Comparison of Intra-Rater and Inter-Rater Reliability Between Human and GPT Evaluations of Research Questions

	GPT 1 vs GPT 2		Human vs GPT 1		Human vs GPT 2		K value average of Human and GPT evaluations
Items	**Weighted Kappa[a]**	**Sig.**	**Weighted Kappa[a]**	**Sig.**	**Weighted Kappa[a]**	**Sig.**	
Q1	0.218	0.025	0.133	0.007	0.067	0.185	0.100
Q2	0.170	0.271	0.057	0.198	-0.035	0.314	0.011
Q3	0.424	0.001	0.268	0.000	0.182	0.005	0.225
Q4	0.768	0.000	0.348	0.000	0.380	0.000	0.364
Q5	0.140	0.234	0.135	0.039	0.034	0.283	0.085
Q6	0.151	0.306	0.126	0.051	0.067	0.193	0.097
Q7	0.267	0.024	0.022	0.586	0.049	0.173	0.036
Q8	0.528	0.000	0.023	0.383	-0.026	0.525	-0.002
Q9	.e	.e	.e	.e	.e	.e	.e
Q10	0.403	0.014	0.020	0.694	0.113	0.027	0.067
Average Q1 to Q10	**0.341**		**0.126**		**0.092**		**0.109**

The analysis showed a "fair" level of intra-rater reliability (K = 0.341) in GPT's evaluations. In contrast, its inter-rater reliability was lower, with average Kappa values of 0.126 and 0.092 for the 2 rounds. Similar to T3, T5, and T9, Q9 had no statistical output due to its uniform scores across three evaluations, indicating a strong reliability regarding ethical considerations.

Despite remaining in the "fair" range, Q4 had the highest human-GPT agreement with statistically significant Kappa values for both rounds and an average K of 0.364 as well as the highest agreement between GPT rounds ($k = 0.768$, $p < 0.001$). These statistics indicated that the model's intra-rater reliability in assessing the research site specification was higher than its inter-rater reliability. A similar pattern appeared in Q3, reflecting a "fair" level of reliability in evaluating the specification of participants in the research questions.

Q1, Q2, Q5, Q6, Q7, and Q10 showed "slight" inter-rater reliability ($K < 0.1$). Human vs. GPT 1 for Q1, Q5, and Q6 showed slightly higher agreement ($K > 0.1$), unlike GPT 2. The differences in p-values further indicated that the observed higher consistency in Human vs. GPT 1 was meaningful, whereas the low consistency in GPT 2 might be due to chance. Intra-rater reliability was generally higher for Q1, Q5, Q6, and Q10, with Q10 at $k = 0.403$. These results showed that ChatGPT yielded slight inter-rater reliability regarding the research questions' answerability, focus, clarity and alignment with titles, but its intra-rater reliability was higher, albeit still within a "fair" range.

While Q2 and Q7 had considerably low inter-rater reliability ($K < 0.05$, $p > 0.1$), their intra-rater reliability was higher. Q8 followed this trend with negative statistics for inter-rater reliability ($K = -0.002$) and a moderate one for intra-rater reliability ($k = 0.528$, $p < 0.001$). With statistically significant statistics, it can be concluded that the model had intra-rater reliability in evaluating the measurability of the terms in the research questions (Q7) and the value of these questions (Q8). Nonetheless, the remaining consistency was inconclusive due to their high p-values ($p > 0.05$).

Research Hypotheses

Table 3.

Comparison of Intra-Rater and Inter-Rater Reliability Between Human and GPT Evaluations of Research Hypotheses

	GPT 1 vs GPT 2		Human vs GPT 1		Human vs GPT 2		K value average of Human and GPT evaluations
Items	**Weighted Kappa[a]**	**Sig.**	**Weighted Kappa[a]**	**Sig.**	**Weighted Kappa[a]**	**Sig.**	
H1	1.000	0.000	0.667	0.134	0.667	0.134	0.667
H2	1.000	0.000	0.500	0.134	0.500	0.134	0.500
H3	1.000	0.000	0.250	0.190	0.250	0.190	0.250
H4	0.000	0.236	0.063	0.743	0.750	0.089	0.406
H5	1.000	0.000	0.757	0.083	0.757	0.083	0.757
H6	1.000	0.000	0.500	0.134	0.500	0.134	0.500
H7	0.400	0.392	-0.500	0.083	-0.154	0.386	-0.327
H8	0.667	0.210	0.100	0.386	0.182	0.134	0.141
H9	.e	.e	.e	.e	.e	.e	.e
H10	.e	.e	.e	.e	.e	.e	.e
Average H1 to H10	**0.758**		**0.292**		**0.431**		**0.362**

Overall, the two rounds of GPT evaluations of research hypotheses showed high intra-rater reliability, with an average Kappa value of 0.758. Notably, the first GPT evaluations had considerably lower inter-rater reliability compared to the second. However, none of the GPT-human comparisons were statistically significant ($p > 0.05$), suggesting the agreement may be due to chance.

No statistical outputs for H9 and H10 were found due to complete agreement across evaluations, reflecting substantial reliability in evaluating ethical considerations of the hypotheses and their alignment with the research titles.

Despite H1 and H5 yielding "substantial" Kappa values for inter-rater reliability, their high p-values prevented a definitive conclusion. However, H1, H2, H3, H5, and H6, with Kappa values of 1.0 ($p < 0.001$), showed that perfect intra-rater reliability was reached in evaluating the testability, feasibility, operationalizability, research site specification, and focus of the hypotheses. Conversely, while both intra- and inter-rater reliability of H7 was not high with Human vs. GPT yielding a negative average Kappa value of -0.327, its statistics were not significant, suggesting inconsistency in evaluating hypothesis accessibility.

Research Paradigm, Design and Techniques

Table 4.

Comparison of Intra-Rater and Inter-Rater Reliability Between Human and GPT Evaluations of Research Paradigm, Design and Techniques

	GPT 1 vs GPT 2		Human vs GPT 1		Human vs GPT 2		K value average of Human and GPT evaluations
Ratings	**Weighted Kappa[a]**	**Sig.**	**Weighted Kappa[a]**	**Sig.**	**Weighted Kappa[a]**	**Sig.**	
P1	0.742	0.000	0.567	0.000	0.546	0.000	0.556
P2	0.641	0.000	0.126	0.047	0.088	0.179	0.107
P3	0.726	0.000	0.389	0.000	0.345	0.000	0.367
Average P1 to P3	**0.703**		**0.361**		**0.326**		**0.344**
D1	0.410	0.001	0.383	0.000	0.239	0.009	0.311
D2	0.420	0.000	0.208	0.002	0.144	0.016	0.176
D3	0.431	0.000	0.215	0.009	0.284	0.002	0.250
Average D1 to D3	**0.420**		**0.269**		**0.223**		**0.246**
Te1	0.601	0.000	0.415	0.000	0.208	0.075	0.312
Te2	0.486	0.000	0.058	0.127	0.036	0.196	0.047
Te3	0.252	0.017	0.086	0.133	0.106	0.089	0.096
Average Te1 to Te3	**0.446**		**0.186**		**0.117**		**0.152**

In general, the inter-rater reliability for both research paradigm and design was fair while that for technique was slight, marked by a considerably low average Kappa value of 0.152. Notably, ChatGPT's initial evaluations aligned more closely with human ratings than subsequent rounds. Moreover, P2, D2, and Te2, which required justifications for choice of paradigms, designs, and techniques, yielded the lowest inter-rater reliability in their respective groups with average Kappa values of 0.107, 0.176, and 0.047, respectively.

Further analysis showed that research paradigm had the highest intra-rater reliability, with an average Kappa of 0.703. However, its inter-rater reliability remained in the "fair" range. P1 had the highest agreement with humans, with both Kappa values exceeding 0.5 ($p < 0.05$). It also achieved a "substantial" Kappa of 0.742 in GPT-to-GPT evaluations, indicating substantial internal consistency in GPT's evaluation of paradigm specification.

In contrast, the evaluation of research design demonstrated less consistency, with average Kappa values of 0.420 between GPT evaluations and 0.246 between human and GPT evaluations. Nevertheless, all Kappa values for this section were statistically significant ($p < 0.05$), indicating that the observed reliability was not due to random variation.

For research techniques, item Te1 had a "fair" inter-rater Kappa ($K = 0.312$), while Te2 and Te3 (justification and appropriateness) were marked by "slight" agreement ($K = 0.047$ and 0.096) with non-significant statistics ($p > 0.05$). Conversely, there was a high level of intra-rater reliability for Te1 ($k = 0.601$), and a "fair" level for Te2 and Te3 ($k = 0.486$, $k = 0.252$), indicating that GPT was more consistent with itself when evaluating research techniques in the

proposals.

Discussion

Overall, employing a zero-shot learning approach, ChatGPT-4o demonstrated a higher level of intra-rater reliability between its evaluation rounds compared to its concordance with human assessors, indicating a lower level of inter-rater reliability with human judgments.

The statistics revealed that most questions fell within the fair agreement range, with Kappa values between 0.2 and 0.4, indicating a fair level of intra-rater reliability. On average, ChatGPT's intra-rater reliability for evaluating Research Titles and Questions was in the upper half of the "fair" range, significantly higher than its inter-rater reliability. This is in line with the findings from Syriani et al. (2023), who concluded that ChatGPT could screen research titles with more self-consistency than traditional classifiers. Notably, criteria related to Hypotheses and Research Paradigm, Design and Techniques showed higher intra-rater reliability. Specifically, the testability, resource allocation, specificity, research site, and central focus of Research Hypotheses displayed perfect intra-rater reliability, while all questions involving Paradigm were at a substantially high level. ChatGPT's intra-rater reliability in evaluating the justifications and appropriateness of Research Design and Techniques was also moderate. These findings align with Thelwall (2024), who observed that ChatGPT tends to produce plausible evaluations but struggles with fine-grained distinctions in research quality when asked to evaluate research based on the REF criteria. ChatGPT also gave different scores to 50 out of 51 articles, with the remaining scored as 3* all 15 times (Thelwall, 2024). This phenomenon resulted in reduced intra-rater reliability, which was also observed in the present research. However, the current research did not record ChatGPT's tendency to assign a default score for articles as noticed by Thelwall (2024). This might be attributed to the use of the Rubric and detailed performance descriptions of each score. Moreover, it is also believed that, compared to the "*score this*" prompt in Thelwall's work, the constraints and restrictions in this paper's Prompt contributed to the avoidance of this issue by guiding the model to adhere more strictly to the Rubric.

ChatGPT shows complete consistency with human assessors in specific areas, such as evaluating ethical considerations of Research Questions and Hypotheses and assessing Research Titles for yes-no phrasing and result descriptions. This high level of reliability is likely due to the straightforward nature of these criteria, which require minimal in-depth analysis. As observed by Thelwall (2024), ChatGPT seems to primarily extract and reword the content from the articles evaluated and use it as factual judgments, which obstructs the evaluation of significance, rigor, and originality. This is why ChatGPT can reliably detect simple, well-defined features but struggles with complex judgments requiring external validation. It is possible that this is the feature that allows for reliable evaluations of straightforward areas, as these rely on pattern and textual cues. Notably, ChatGPT-4o's ability to identify potential ethical issues in research projects suggests that the model can approximate human-like reasoning in some cases, contrasting slightly with findings from Lin et al. (2023).

Fair to moderate inter-rater reliability was observed in evaluating the feasibility of research hypotheses, participant descriptions, resource allocation, title brevity, research site aspects, and the clarity, focus, and answerability of research questions. The weak agreement in these areas aligns with findings from Liang et al. (2023) and Thelwall (2024). This suggests that while ChatGPT can approximate human evaluations in areas requiring moderate analysis, it still falls short of achieving full concordance.

Significant unreliability emerged in critical areas such as interpreting the specificity and relevance of titles related to English language learning, understanding jargon and acronyms, evaluating hypothesis accessibility, and justifying paradigms, designs, and techniques. These discrepancies reflect those observed by various authors (i.e., Liang et al., 2023; Lin et al., 2023; Syriani et al., 2023; Thelwall, 2024) when ChatGPT struggled with broader and more abstract criteria that require human logical reasoning. A possible explanation for the phenomenon can be the model's dependence on text extraction methods at a surface level, as observed by Thelwall (2024). The model is bound to face difficulty in determining if a research title is appropriately specific or relevant because it does not possess comparison capabilities for existing literature to identify the norms in specificity and relevance. In addition, the shortcomings in evaluating the justification of Design and Techniques might derive from this feature, in which it treats textual claims as facts, hindering its ability to truly verify whether the justification is sufficient or logical. A reasonable assumption can be made regarding the evaluation of jargon and acronyms, as ChatGPT may struggle to determine what qualifies as "jargon" due to its extensive knowledge base. Moreover, in concordance with the observations from Thelwall (2024), this research also recorded ChatGPT's limited ability to evaluate the significance of the research questions. As explained by Thelwall (2024), this issue arises from the lack of external knowledge of the field or independent reasoning when evaluating research significance.

This suggests that while ChatGPT successfully recognizes recurring critiques and aligns with human feedback in areas requiring moderate analysis, it still falls short of full concordance, particularly in evaluating more nuanced methodological concerns. These findings are in concordance with Liang et al. (2023), who found ChatGPT's agreement with human reviewers was higher for weaker papers, indicating that it is more effective at identifying common flaws and surface-level issues. Liang et al. (2023) also noted that ChatGPT-4 tends to focus on commonly emphasized aspects of feedback, such as suggesting additional datasets for experiments, while struggling with deeper critiques of research design and novelty.

In general, the inconsistencies in evaluating these subjective criteria suggest that while ChatGPT can assist in research evaluation, human oversight remains essential to address the nuances that AI may miss (Liang et al., 2023). Nonetheless, it is worth noting that for straightforward criteria, ChatGPT has achieved a substantially high level of both intra- and inter-rater reliability. This underscores its strong capability to evaluate components of research proposals using a given rubric with minimal training, making it comparable to human assessors in certain contexts.

Conclusion

Overall, the study highlights ChatGPT-4o's potential in employing a zero-shot learning approach to evaluate methodological components against a rubric. In straightforward areas, it yielded moderate to high intra-rater reliability and moderate inter-rater reliability. However, the model struggled with abstract criteria requiring deeper reasoning. This suggests that while ChatGPT can be reliably utilized in contexts where clear, detailed prompts and objective rubrics are in place and minimal training is available, human intervention remains necessary.

A few limitations should be noted in the present study. First, the small sample size of 37 proposals from the English language field may limit the generalizability of the findings. Second, the study focuses solely on quantitative data, potentially overlooking important contextual insights that qualitative feedback could provide such as explanations for the scores given, which

enhances understanding of the evaluation process. Finally, the study addresses only somewhat superficial areas of research methodology that require limited professional expertise, which may not fully capture the complexities and nuances of comprehensive research methodology evaluation.

Future research should increase the sample size with papers from various fields, include qualitative data for deeper insights, and evaluate more complex aspects of research methodology to better assess AI's evaluation capabilities. Since ChatGPT-4o cannot evaluate nuanced methodological aspects and subjective criteria well, it makes sense to further study how refining prompts and rubrics might close the gap between the model and human evaluators. There is potential in examining other methods such as one-shot or few-shot learning to improve how well the model performs in evaluation tasks, especially in areas that require complex reasoning. In addition, by experimenting with other systems such as Gemini or Claude, we can potentially discover better evaluative systems than ChatGPT.

References

Andrews, R. (2003). *Research questions*. Continuum.

Berry, K. J., Janis E. Johnston, & Paul W Miele, Jr. (2008). Weighted Kappa for Multiple Raters. *Perceptual and Motor Skills*, *107*(7), 837–848. https://doi.org/10.2466/PMS.107.7.837-848

Blaxter, L., Hughes, C., & Tight, M. (2010). *How to research* (4th ed.). McGraw-Hill/Open University Press.

Brookhart, S. M. (2013). *How to Create and Use Rubrics for Formative Assessment and Grading*. ASCD.

Brown, H. D. (2018). *Language assessment: Principles and classroom practices*. Pearson.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). *Language Models are Few-Shot Learners* (No. arXiv:2005.14165). arXiv. https://doi.org/10.48550/arXiv.2005.14165

Cadman, K. (2002). English for Academic Possibilities: the research proposal as a contested site in postgraduate genre pedagogy. *Journal of English for Academic Purposes, 2*(2), 85-104.

Chaudhary, S., & Gupta, P. (2023). A Comprehensive Study on Chat GPT. *Journal of Emerging Technologies and Innovative Research*, *10*(10), 196–201.

Checco, A., Bracciale, L., Loreti, P., Pinfield, S., & Bianchi, G. (2021). AI-assisted peer review. *Humanities and Social Sciences Communications*, *8*(1), 25. https://doi.org/10.1057/s41599-020-00703-8

Cohen, J. (1968). Weighted Kappa: Nominal Scale Agreement with Provision for Scaled Disagreement or Partial Credit. *Psychological Bulletin*, *7*(4), 213–220.

Cohen, L., Manion, L., & Morrison, K. (2018). *Research Methods in Education*. Routledge.

Creswell, J. W. (2015). *Educational Research: Planning, Conducting, and Evaluating Quantitative and Qualitative Research*. Pearson.

Creswell, J. W., & Creswell, J. D. (2018). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. SAGE Publications, Inc.

Denscombe, M. (2020). *Research proposals: A practical guide* (2nd ed). Open University Press.

Duong, N., Tong, T., & Le, D. (2024). *Utilizing ChatGPT in checking academic writing for postgraduate students*. In Proceedings of the AsiaCALL International Conference (pp. 193–203). https://doi.org/10.54855/paic.24614

Duong, T., & Le, T. (2024). *Utilizing artificial intelligence in writing feedback: Benefits and challenges for first-year students at Hanoi University of Industry*. In Proceedings of the AsiaCALL International Conference (pp. 238–249).

Fraenkel, J. R., Wallen, N. E., & Hyun, H. (2023). *How to Design and Evaluate Research in Education*. McGraw Hill.

Gwet, K. L. (2008). *Intrarater Reliability*. Wiley Encyclopedia of Clinical Trials.

Heriyawati, D. F., & Romadhon, M. G. E. (2025). "Can AI Be Trusted for My Thesis?" The Voices of Indonesian Higher Education Levels About ChatGPT in Automated Writing Evaluation (AWE). *Computer-Assisted Language Learning Electronic Journal*, *26*(1), 58–75. https://doi.org/10.54855/callej.252614

Hoang, T., & Vu, T. (2024). *Khám phá vai trò của trí tuệ nhân tạo AI – ChatGPT trong kỷ nguyên chuyển đổi số để ứng dụng vào giảng dạy tiếng Anh (ELT) tại Đại học Kinh tế - Kỹ thuật Công nghiệp (UNETI)*. In Kỷ yếu hội thảo khoa học quốc gia: Ngôn ngữ học tính toán – những xu hướng mới, triển vọng và thách thức (pp. 31–39)

Hockly, N. (2019). Automated writing evaluation. *ELT Journal*, *73*(1), 82–88. https://doi.org/10.1093/elt/ccy044

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2023). *Large Language Models are Zero-Shot Reasoners* (No. arXiv:2205.11916). arXiv. https://doi.org/10.48550/arXiv.2205.11916

Kousha, K., & Thelwall, M. (2022). *Artificial intelligence technologies to support research assessment: A review*. Statistical Cybermetrics and Research Evaluation Group, University of Wolverhampton.

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Latif, E., & Zhai, X. (2023). *Fine-tuning ChatGPT for Automatic Scoring* (No. arXiv:2310.10072). arXiv. https://doi.org/10.48550/arXiv.2310.10072

Lee, C. J., Sugimoto, C. R., Zhang, G., & Cronin, B. (2013). Bias in peer review. *Journal of the American Society for Information Science and Technology*, *64*(1), 2–17. https://doi.org/10.1002/asi.22784

Liang, W., Zhang, Y., Cao, H., Wang, B., Ding, D., Yang, X., Vodrahalli, K., He, S., Smith, D., Yin, Y., McFarland, D., & Zou, J. (2023). *Can large language models provide useful feedback on research papers? A large-scale empirical analysis* (No. arXiv:2310.01783). arXiv. https://doi.org/10.48550/arXiv.2310.01783

Lin, J., Song, J., Zhou, Z., Chen, Y., & Shi, X. (2023). Automated scholarly paper review: Concepts, technologies, and challenges. *Information Fusion*, *98*, 101830.

https://doi.org/10.1016/j.inffus.2023.101830

Locke, L. F., Spirduso, W. W., & Silverman, S. J. (2007). *Proposals That Work: A Guide for Planning Dissertations and Grant Proposals*. SAGE Publications, Inc.

Luu, T. M. V., & Doan, Q. V. (2025). ChatGPT's Impact on Listening Comprehension: Perspectives from Vietnamese EFL University Learners. *Computer-Assisted Language Learning Electronic Journal*, *26*(3), 43–63. https://doi.org/10.54855/callej.252633

Maclure, M., & Willet, W. C. (1987). Misinterpretation and misuse of the kappa statistic. *American Journal of Epidemiology, 126*(2), 161–169.

Menke, J., Roelandse, M., Ozyurt, B., Martone, M., & Bandrowski, A. (2020). The Rigor and Transparency Index Quality Metric for Assessing Biological and Medical Science Methods. *iScience*, *23*(11), 101698. https://doi.org/10.1016/j.isci.2020.101698

Nguyen, T. S., Nguyen, T. D. T., Hoang, N. Q. N., & Do, T. K. H. (2025). How AI-Powered Voice Recognition Has Supported Pronunciation Competence among EFL University Learners. *Computer-Assisted Language Learning Electronic Journal*, *26*(3), 64–83. https://doi.org/10.54855/callej.252634

OpenAI. (2022, November 30). *Introducing ChatGPT*. https://openai.com/index/chatgpt/

OpenAI. (2024a). *GPT-4o System Card* (p. 32).

OpenAI. (2024b, May 13). *Hello GPT-4o*. https://openai.com/index/hello-gpt-4o/

Ormerod, C. M., Malhotra, A., & Jafari, A. (2021). *Automated essay scoring using efficient transformer-based language models* (No. arXiv:2102.13136). arXiv. https://doi.org/10.48550/arXiv.2102.13136

Paltridge, B., & Starfield, S. (2020). *Thesis and dissertation writing in a second language: A handbook for students and their supervisors* (2nd ed). Routledge.

Pham, M. T., & Cao, T. X. T. (2025). The Practice of ChatGPT in English Teaching and Learning in Vietnam: A Systematic Review. *International Journal of TESOL & Education*, *5*(1), 50–70. https://doi.org/10.54855/ijte.25513

Reynolds, L., & McDonell, K. (2021). *Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm* (No. arXiv:2102.07350). arXiv. https://doi.org/10.48550/arXiv.2102.07350

Rodriguez, P. U., Jafari, A., & Ormerod, C. M. (2019). *Language models and Automated Essay Scoring* (No. arXiv:1909.09482). arXiv. https://doi.org/10.48550/arXiv.1909.09482

Russell, S., & Norvig, P. (2020). *Artificial Intelligence: A Modern Approach* (4th ed.). Pearson.

Sabzalieva, E., & Valentini, A. (2023). *ChatGPT and artificial intelligence in higher education: Quick start guide – UNESCO Digital Library*. United Nations Educational, Scientific and Cultural Organization. https://unesdoc.unesco.org/ark:/48223/pf0000385146.

Saigon University. (2018, January). *Chiến lược phát triển Trường Đại học Sài Gòn đến năm 2025 và tầm nhìn.* Saigon university.

Sim, J., & Wright, C. C. (2005). The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. *Physical Therapy*, *85*(3), 257–268.

https://doi.org/10.1093/ptj/85.3.257

Spaapen, J. B., Dijstelbloem, H., & Wamelink, F. J. M. (2007). *Evaluating research in context: A method for comprehensive assessment* (2nd ed). Consultative Committee of Sector Councils for Research and Development (COS).

Swales, J. (2004). *Research genres: Explorations and applications.* Cambridge University Press.

Syriani, E., David, I., & Kumar, G. (2023). *Assessing the Ability of ChatGPT to Screen Articles for Systematic Reviews*.

Tcherni-Buzzeo, M., & Pyrczak, F. (2024). *Evaluating Research in Academic Journals*. Routledge.

Thelwall, M. (2024). Can ChatGPT evaluate research quality? *Journal of Data and Information Science*, *9*(2), 1–21. https://doi.org/10.2478/jdis-2024-0013

Thomas, R. M. (2003). *Blending Qualitative and Quantitative Research Methods in Theses and Dissertations*. Corwin.

Turabian, K. L., Booth, W. C., Colomb, G. G., Williams, J. M., Bizup, J., & FitzGerald, W. T. (2018). *A Manual for Writers of Research Papers, Theses, and Dissertations: Chicago Style for Students and Researchers* (9th ed.). University of Chicago Press.

Wang, Q., & Gayed, J. M. (2024). Effectiveness of large language models in automated evaluation of argumentative essays: Finetuning vs. zero-shot prompting. *Computer Assisted Language Learning*, 1–29. https://doi.org/10.1080/09588221.2024.2371395

Biodata

Dr. Tran The Phi currently works as Dean of Faculty of Foreign Languages, Saigon University, Vice President of South Vietnam Teaching English to Speakers of Other Languages (STESOL) under Association of Vietnam Universities and Colleges (AVU&C). He has been teaching English for over 20 years. His research interests are in cognitive linguistics, curriculum development and English teaching methods.

Ms. Nguyen Thi Hoang Lan is presently working as a lecturer in the Faculty of Foreign Languages at Saigon University, Vietnam. She is responsible for delivering a range of courses in English linguistics and language education. Her research interests primarily focus on the pedagogy and acquisition of research writing skills, as well as the integration of artificial intelligence in educational practice and assessment.

APPENDIX A

RUBRIC

Sections to evaluate	Evaluating questions	Evaluation scale
Research Title	**Question T1:** Is the title sufficiently specific?	1. Very Unsatisfactory (1 point): The title is extremely vague and does not provide any specific details about the research topic. 2. Unsatisfactory (2 points): The title is somewhat vague, providing a general idea of the topic but lacking precise details or key concepts. 3. Neutral (3 points): The title provides some specificity but could be more detailed in describing the specific focus or variables of the research. 4. Satisfactory (4 points): The title is highly specific, clearly defining the research focus with precise details such as key variables or methodologies. 5. Very Satisfactory (5 points): The title provides a clear, precise, and engaging description of the research focus without any ambiguity.
	Question T2: Is the title reasonably brief?	1. Very Unsatisfactory (1 point): The title is overly wordy and confusing or has incorrect spelling or grammar structure, making it difficult to grasp the main focus of the research. 2. Unsatisfactory (2 points): The title is too long and vague, making it challenging to understand the specific area of focus for the research. 3. Neutral (3 points): The title is neither overly long nor very brief, offering a general sense of the research subject without being overly detailed. 4. Satisfactory (4 points): The title is concise and clear, effectively conveying the primary topic of the research without being overly wordy. 5. Very Satisfactory (5 points): The title is extremely concise while still conveying the core focus of the research in a clear and understandable manner.
	Question T3: Has the author avoided using a "yes–no" question as a title?	1. Very unsatisfactory (1 point): The title is a straightforward "yes-no" question. 2. Unsatisfactory (2 points): The title is predominantly a "yes-no" question but may include additional context or elements. 3. Neutral (3 points): The title could potentially be interpreted as a "yes-no" question, but it also suggests complexity or additional layers that go beyond a simple answer. 4. Satisfactory (4 points): The title avoids being a direct "yes-no" question and instead implies complexity, multiple factors, or a range of possible answers. 5. Very satisfactory (5 points): The title clearly avoids any "yes-no" question and effectively captures the complexity or scope of the research without oversimplification.
	Question T4: Is the title free of jargon and acronyms that might be unknown to the audience for the research report?	1. Very unsatisfactory (1 point): The title contains numerous technical terms, acronyms, or specialized jargon that are likely unfamiliar in the field of English teaching and linguistics. 2. Unsatisfactory (2 points): The title includes some technical terms or acronyms that could confuse a general audience, though it might still be understandable with effort. 3. Neutral (3 points): The title might have a few technical terms or acronyms, but they are explained or can be easily inferred from the context provided. 4. Satisfactory (4 points): The title avoids unnecessary technical terms and acronyms, using clear and accessible language suitable for a broad audience. 5. Very satisfactory (5 points): The title is completely free of jargon and

		acronyms, ensuring it is easily understandable and accessible to any reader.
	Question T5: Has the author avoided describing results in the title?	1. Very unsatisfactory (1 point): The title explicitly describes specific results or findings of the research. 2. Unsatisfactory (2 points): The title strongly hints at specific results or findings without directly stating them. 3. Neutral (3 points): The title may imply results or findings, but it does not explicitly describe them, leaving some ambiguity. 4. Satisfactory (4 points): The title avoids describing specific results or findings, focusing instead on the research topic or question. 5. Very satisfactory (5 points): The title is completely free from any hint or description of specific results or findings, maintaining a focus on the research topic or objective.
	Question T6: Are the key variables mentioned in the title?	1. Very unsatisfactory (1 point): The title does not mention any key variables relevant to the research. 2. Unsatisfactory (2 points): The title mentions some variables, but they are not clearly identified as key variables in the research. 3. Neutral (3 points): The title mentions key variables, but it may not clearly specify them as central to the research. 4. Satisfactory (4 points): The title mentions key variables clearly and indicates their importance to the research topic. 5. Very satisfactory (5 points): The title not only mentions key variables clearly but also indicates their relationship or significance within the research context.
	Question T7: Does the title identify the types of individuals who participated in the study or the types of aggregate units in the sample?	1. Very unsatisfactory (1 point): The title does not mention any information about the participants or population in the sample. 2. Unsatisfactory (2 points): The title vaguely hints at the participants or population but lacks specificity or clarity. 3. Neutral (3 points): The title mentions some information about the participants or population, but it is not clearly specified. 4. Satisfactory (4 points): The title identifies the types of individuals who participated in the study or the types of population in the sample, providing clarity about the study's focus. 5. Very satisfactory (5 points): The title clearly identifies and specifies the types of individuals who participated in the study or the types of population, giving a precise description of the study's target group.
	Question T8: Are any highly unique or very important characteristics of the study referred to in the title or subtitle?	1. Very unsatisfactory (1 point): The title lacks specificity or uniqueness about the study's focus or contribution. 2. Unsatisfactory (2 points): The title suggests some unique aspects but does not clearly articulate what makes the study noteworthy compared to others. 3. Neutral (3 points): The title outlines certain unique features or contributions of the study without standing out significantly among other research titles. 4. Satisfactory (4 points): The title clearly highlights distinctive features or contributions of the study, making it stand out in terms of its focus or importance. 5. Very satisfactory (5 points): The title effectively conveys unique and important aspects of the study, clearly indicating its significant contributions or innovative elements.

	Question T9: If there are a main title and a subtitle, do both provide important information about the research?	1. Very unsatisfactory (1 point): Neither the main title nor the subtitle provides meaningful information about the research. 2. Unsatisfactory (2 points): One of either the main title or subtitle is vague or unclear about the research. 3. Neutral (3 points): Both the main title and subtitle provide some information, but could be more explicit. 4. Satisfactory (4 points): Both the main title and subtitle clearly convey important aspects of the research. 5. Very satisfactory (5 points): The main title and subtitle together effectively communicate the research's main focus and contributions.
	Question T10: Does the research title clearly reflect its focus on an aspect of English language learning, teaching, or usage?	1. Very unsatisfactory (1 point): The title does not indicate any connection to English language learning, teaching, or usage. It is unclear how the research relates to these aspects. 2. Unsatisfactory (2 points): The title vaguely suggests a connection to English language learning, teaching, or usage, but it is not clearly defined or specified. 3. Neutral (3 points): The research title mentions English language learning, teaching, or usage, but the connection is moderate or not prominently highlighted. 4. Satisfactory (4 points): The title clearly indicates a focus on an aspect of English language learning, teaching, or usage. It explicitly states its relevance and connection to these areas. 5. Very satisfactory (5 points): The title exceptionally and clearly reflects its focus on an aspect of English language learning, teaching, or usage. It effectively communicates the specific area of study within these domains.
Research questions	**Question Q1:** Can the research question be answered scientifically using the existing research methodology?	1. Very unsatisfactory (1 point): Existing methodologies are inappropriate or completely insufficient to scientifically address the research question. 2. Unsatisfactory (2 points): Existing methodologies have significant limitations, making it difficult to scientifically address the research question. 3. Neutral (3 points): Existing methodologies are somewhat adequate but may require considerable adaptation or supplementation to scientifically address the research question. 4. Satisfactory (4 points): Existing methodologies are suitable and can be effectively used to scientifically address the research question with proper application. 5. Very satisfactory (5 points): Existing methodologies are highly suitable and can be used with ease to scientifically address the research question.
	Question Q2: Can the research question be answered given the following available resources: less than 2 years of time, 5 people, and a budget under 100 million VND?	1. Very unsatisfactory (1 point): The resources are insufficient, and it is almost certain that the research question cannot be answered. 2. Unsatisfactory (2 points): The resources may be insufficient, making it challenging to answer the research question. 3. Neutral (3 points): The resources are barely adequate, and answering the research question will require careful management and possibly some compromises. 4. Satisfactory (4 points): The resources are adequate, and the research question can be answered with proper planning and effort. 5. Very satisfactory (5 points): The resources are more than sufficient, and answering the research question should be straightforward.
	Question Q3: Does the research question specify	1. Very unsatisfactory (1 point): Participants are not mentioned at all, making it impossible to understand who would be involved in the study. 2. Unsatisfactory (2 points): Participants are vaguely mentioned or

	the participants involved in the study?	described, lacking clarity on who exactly would be involved in the study. 3. Neutral (3 points): Participants are partially specified, but details are insufficient to identify the participants group. 4. Satisfactory (4 points): Participants are clearly specified, allowing for a well-defined scope and easy identification of the participants. 5. Very satisfactory (5 points): Participants are not only clearly specified and described but also justified in detail, ensuring alignment with the study's objectives and comprehensive planning.
	Question Q4: Does the research question specify the research site where the study will be conducted?	1. Very unsatisfactory (1 point): No specific research site is mentioned, leaving it unclear where the study will be conducted. 2. Unsatisfactory (2 points): The research question vaguely mentions a research site without providing clear details on its location (country, city, district) 3. Neutral (3 points): The research question mentions a research site but lacks sufficient detail for effective identification (country, city, district). 4. Satisfactory (4 points): The research question clearly specifies the research site, allowing for well-defined planning in a particular location. 5. Very satisfactory (5 points): The research question specifies the research site in complete detail, leaving no ambiguity. All relevant specifics are included, making the site unmistakable.
	Question Q5: Does the research question clearly focus on a single phenomenon or concept?	1. Very Unsatisfactory (1 point): The research question lacks any clear focus on a single phenomenon or concept. It is vague or tries to address multiple unrelated phenomena. 2. Unsatisfactory (2 points): The research question identifies a broad area but still encompasses multiple phenomena or concepts, making it unclear what the main focus of the research is. 3. Neutral (3 points): The research question narrows down to a more specific phenomenon or concept or lacks clarity or precision, making it somewhat ambiguous or broad. 4. Satisfactory (4 points): The research question focuses on a single phenomenon or concept with specification, allowing for identification of the aspects under research. 5. Very Satisfactory (5 points): The research question is entirely focused on a single, well-defined phenomenon or concept, leaving no ambiguity. It is specific, concise and includes all variables.
	Question Q6: Is the research question clear, involving only keywords and terms that are understandable to readers?	1. Very unsatisfactory (1 point): Confusing or overly technical terms make the research question hard to understand. 2. Unsatisfactory (2 points): Some unclear terms or jargon confuse the meaning of the research question. 3. Neutral (3 points): Mostly understandable language, but some terms could be clearer. 4. Satisfactory (4 points): Clear language with terms generally understandable to readers. 5. Very satisfactory (5 points): Clear language with terms generally understandable to readers.
	Question Q7: Can the keywords and terms used in the research question be defined precisely for measurement or	1. Very unsatisfactory (1 point): The research question uses abstract words and terms that are difficult to be defined or measured. The terms lack specificity and are open to multiple interpretations. 2. Unsatisfactory (2 points): The research question includes some keywords and terms that are slightly more specific but still lack precise definitions. There is some ambiguity in how the terms can be measured or identified. 3. Neutral (3 points): The research question includes keywords and terms

	identification?	that are fairly specific and can be defined with some effort. However, some terms might still be open to interpretation and require additional clarification. 4. Satisfactory (4 points): The research question uses keywords and terms that are well-defined and can be precisely measured or identified. There is minimal ambiguity, but a detailed explanation of the terms might still be necessary. 5. Very satisfactory (5 points): The research question uses keywords and terms that are exceptionally clear and precise, with no ambiguity. The terms can be easily defined, measured, or identified, ensuring accurate and reliable research.
	Question Q8: Is the research question worth investigating and contributes valuable knowledge to the field of teaching, learning or using English language?	1. Very unsatisfactory (1 point): The research question is not relevant or significant to the field. The topic might be outdated, trivial, or already well-established. 2. Unsatisfactory (2 points): The research question has limited relevance and contributes minimal new knowledge or insights. It might address a minor or niche aspect of the field but lacks broader significance or impact. 3. Neutral (3 points): The research question is somewhat relevant and contributes a moderate amount of new knowledge or insights. It addresses a specific aspect of the field and adds value but may not have a wide-reaching impact. 4. Satisfactory (4 points): The research question is highly relevant and contributes significant new knowledge or insights. It addresses an important aspect of the field and has the potential to influence practice, policy, or further research. 5. Very satisfactory (5 points): The research question is extremely relevant and contributes highly valuable new knowledge or insights. It addresses a critical gap in the field, has the potential for transformative impact, and can significantly influence practice, policy, or further research
	Question Q9: Is the research question ethical, meaning it does not involve potential threat or harm to the subjects involved?	1. Very unsatisfactory (1 point): The research question poses severe ethical concerns and potential harm to subjects involved, lacking any ethical considerations 2. Unsatisfactory (2 points): Ethical considerations in the research question are insufficient, with notable potential risks or harm to subjects. 3. Neutral (3 points): Ethical considerations are partially addressed, but some potential risks or harm to subjects are not adequately mitigated 4. Satisfactory (4 points): The research question demonstrates adequate ethical considerations, with minimal risk or harm to subjects involved. 5. Very satisfactory (5 points): The research question is highly ethical, ensuring no potential threat or harm to subjects and demonstrating exemplary ethical standards.
	Question Q10: Does the research question align closely with the Research Title?	1. Very unsatisfactory (1 point): The research question and the research title are not aligned at all. They address entirely different topics or phenomena. 2. Unsatisfactory (2 points): The research question and the research title have minimal alignment. They touch on related areas but do not clearly connect or focus on the same specific topic, participants or context. 3. Neutral (3 points): The research question and the research title are moderately aligned. They address similar topics but lack a direct and clear connection or focus. 4. Satisfactory (4 points): The research question and the research title are well aligned. They address the same topic, participants and context with a clear and direct connection, but some minor discrepancies or broader focus might exist.

		5. Very satisfactory (5 points): The research question and the research title are perfectly aligned. They address the exact same topic, participants and context with precise and clear connection, leaving no ambiguity.
Hypotheses	**Question H1:** Is the hypothesis formulated in a way that it can be empirically tested through existing research methodologies?	1. Very unsatisfactory (1 point): Existing methodologies are inappropriate or completely insufficient to scientifically address the research hypothesis. 2. Unsatisfactory (2 points): Existing methodologies have significant limitations, making it difficult to scientifically address the research hypothesis. 3. Neutral (3 points): Existing methodologies are somewhat adequate but may require considerable adaptation or supplementation to scientifically address the research hypothesis. 4. Satisfactory (4 points): Existing methodologies are suitable and can be effectively used to scientifically address the research hypothesis with proper application. 5. Very satisfactory (5 points): Existing methodologies are highly suitable and can be used with ease to scientifically address the research hypothesis.
	Question H2: Can the hypothesis be tested given the following available resources: less than 2 years of time, 5 people, and a budget under 100 million VND?	1. Very unsatisfactory (1 point): The resources are insufficient, and it is almost certain that the research hypothesis cannot be tested. 2. Unsatisfactory (2 points): The resources may be insufficient, making it challenging to test the research hypothesis. 3. Neutral (3 points): The resources are barely adequate, and testing the research hypothesis will require careful management and possibly some compromises. 4. Satisfactory (4 points): The resources are adequate, and the research hypothesis can be tested with proper planning and effort. 5. Very satisfactory (5 points): The resources are more than sufficient, and testing the research hypothesis should be straightforward.
	Question H3: Is the hypothesis sufficiently precise and specific to generate its methods?	1. Very unsatisfactory (1 point): Hypothesis lacks specificity and cannot generate clear methods. 2. Unsatisfactory (2 points): Hypothesis is somewhat specific but needs more precision in defining variables and methods. 3. Neutral (3 points): Hypothesis is moderately specific but requires further detail for method generation. 4. Satisfactory (4 points): Hypothesis is sufficiently specific to generate clear and appropriate research methods. 5. Very satisfactory (5 points): Hypothesis is exceptionally specific and precise, ensuring accurate and effective method generation.
	Question H4: Does the hypothesis specify the participants involved in the study?	1. Very unsatisfactory (1 point): Participants are not mentioned at all, making it impossible to understand who would be involved in the study. 2. Unsatisfactory (2 points): Participants are vaguely mentioned or described, lacking clarity on who exactly would be involved in the study. 3. Neutral (3 points): Participants are partially specified, but details are insufficient for to identify the participants group. 4. Satisfactory (4 points): Participants are clearly specified, allowing for a well-defined scope and easy identification of the participants. 5. Very satisfactory (5 points): Participants are not only clearly specified and described but also justified in detail, ensuring alignment with the study's objectives and comprehensive planning
	Question H5: Does the hypothesis specify the	1. Very unsatisfactory (1 point): No specific research site is mentioned, leaving it unclear where the study will be conducted. 2. Unsatisfactory (2 points): The research hypothesis vaguely mentions a research site without providing clear details on its location (country, city,

	research site where the study will be conducted?	district) 3. Neutral (3 points): The research hypothesis mentions a research site but lacks sufficient detail for effective identification (country, city, district). 4. Satisfactory (4 points): The research hypothesis clearly specifies the research site, allowing for well-defined planning in a particular location. 5. Very satisfactory (5 points): The research hypothesis specifies the research site in complete detail, leaving no ambiguity. All relevant specifics are included, making the site unmistakable.
	Question H6: Does the hypothesis focus on a single phenomenon or concept?	1. Very unsatisfactory (1 point): The research hypothesis lacks any clear focus on a single phenomenon or concept. It is vague or tries to address multiple unrelated phenomena. 2. Unsatisfactory (2 points): The research hypothesis identifies a broad area but still encompasses multiple phenomena or concepts, making it unclear what the main focus of the research is. 3. Neutral (3 points): The research hypothesis narrows down to a more specific phenomenon or concept or lack lacks clarity or precision, making it somewhat ambiguous or broad. 4. Satisfactory (4 points): The research hypothesis focuses on a single phenomenon or concept with specification, allowing for identification of the aspects under research. 5. Very satisfactory (5 points): The research hypothesis is entirely focused on a single, well-defined phenomenon or concept, leaving no ambiguity. It is specific, concise and includes all variables.
	Question H7: Is the research hypothesis clear, involving only keywords and terms that are understandable to readers?	1. Very unsatisfactory (1 point): Confusing or overly technical terms make the research hypothesis hard to understand. 2. Unsatisfactory (2 points): Some unclear terms or jargon confuse the meaning of the research hypothesis. 3. Neutral (3 points): Mostly understandable language, but some terms could be clearer. 4. Satisfactory (4 points): Clear language with terms generally understandable to readers. 5. Very satisfactory (5 points): Clear language with terms generally understandable to readers.
	Question H8: Can the keywords and terms used in the hypothesis be defined precisely for measurement or identification?	1. Very unsatisfactory (1 point): The research hypothesis uses abstract words and terms that are difficult to be defined or measured. The terms lack specificity and are open to multiple interpretations. 2. Unsatisfactory (2 points): The research hypothesis includes some keywords and terms that are slightly more specific but still lack precise definitions. There is some ambiguity in how the terms can be measured or identified. 3. Neutral (3 points): The research hypothesis includes keywords and terms that are fairly specific and can be defined with some effort. However, some terms might still be open to interpretation and require additional clarification. 4. Satisfactory (4 points): The research hypothesis uses keywords and terms that are well-defined and can be precisely measured or identified. There is minimal ambiguity, but a detailed explanation of the terms might still be necessary. 5. Very satisfactory (5 points): The research hypothesis uses keywords and terms that are exceptionally clear and precise, with no ambiguity. The terms can be easily defined, measured, or identified, ensuring accurate and reliable.
	Question H9: Is	1. Very unsatisfactory (1 point): The research hypothesis poses severe

	the hypothesis ethical, meaning it does not involve potential threat or harm to the subjects involved?	ethical concerns and potential harm to subjects involved, lacking any ethical considerations 2. Unsatisfactory (2 points): Ethical considerations in the research hypothesis are insufficient, with notable potential risks or harm to subjects. 3. Neutral (3 points): Ethical considerations are partially addressed, but some potential risks or harm to subjects are not adequately mitigated 4. Satisfactory (4 points): The research hypothesis demonstrates adequate ethical considerations, with minimal risk or harm to subjects involved. 5. Very satisfactory (5 points): The research hypothesis is highly ethical, ensuring no potential threat or harm to subjects and demonstrating exemplary ethical standards.
	Question H10: Does the hypothesis align closely with the Research Title?	1. Very unsatisfactory (1 point): The research hypothesis and the research title are not aligned at all. They address entirely different topics or phenomena. 2. Unsatisfactory (2 points): The research hypothesis and the research title have minimal alignment. They touch on related areas but do not clearly connect or focus on the same specific topic, participants or context. 3. Neutral (3 points): The research hypothesis and the research title are moderately aligned. They address similar topics but lack a direct and clear connection or focus. 4. Satisfactory (4 points): The research hypothesis and the research title are well aligned. They address the same topic, participants and context with a clear and direct connection, but some minor discrepancies or broader focus might exist. 5. Very satisfactory (5 points): The research hypothesis and the research title are perfectly aligned. They address the exact same topic, participants and context with precise and clear connection, leaving no ambiguity.
Paradigm	**Question P1:** Does the Methodology section clearly specify research paradigms (i.e. quantitative, qualitative or mixed methods)?	1. Very unsatisfactory (1 point): The methodology section does not specify any research paradigms or is unclear about the paradigm used. 2. Unsatisfactory (2 points): The methodology vaguely mentions research paradigms without clearly identifying whether it is quantitative, qualitative, or mixed. 3. Neutral (3 points): The methodology section mentions research paradigms but lacks clarity or specificity in defining whether it is quantitative, qualitative, or mixed. 4. Satisfactory (4 points): The methodology clearly specifies whether it is quantitative, qualitative, or mixed, providing a clear understanding of the research paradigm. 5. Very satisfactory (5 points): The methodology precisely and explicitly states whether it is quantitative, qualitative, or mixed, ensuring complete clarity on the research paradigm.
	Question P2: Based solely on explicitly stated information, does the Methodology section clearly justify how the research paradigms chosen (i.e. quantitative, qualitative or	1. Very unsatisfactory (1 point): The methodology does not explicitly state how the chosen paradigm will gather data relevant to the research questions or hypotheses of the proposal. 2. Unsatisfactory (2 points): The methodology provides vague text on the how the chosen paradigm will solicit data, lacking clarity or relevance to research questions/hypotheses of the proposal. 3. Neutral (3 points): The methodology somewhat how the chosen paradigm will solicit data and what kind of data it will provide, but the explanation lacks depth or direct relevance to research questions/hypotheses of the proposal. 4. Satisfactory (4 points): The methodology clearly justifies how the chosen paradigm will solicit data that can directly answer the research

	mixed methodologies) provide data that will address the research questions or hypotheses of the proposal?	questions or hypotheses of the proposal. 5. Very satisfactory (5 points): The methodology extensively and precisely justifies how the chosen paradigm will solicit data to answer the research questions or test the hypotheses of the proposal
	Question P3: Based on the theories of Research paradigms from the book "How to Research", can the research paradigms (i.e. quantitative, qualitative or mixed methodologies) chosen solicit data sufficient for answering the research questions or testing the hypotheses?	1. Very unsatisfactory (1 point): There is a significant mismatch between the chosen research paradigm and the theories from the book "How to Research", making it unlikely to solicit sufficient data for answering research questions or testing hypotheses of the proposal. 2. Unsatisfactory (2 points): The alignment between the chosen research paradigms and the theories from The book "How to Research" is weak, limiting the effectiveness of data solicitation for research questions or hypotheses of the proposal. 3. Neutral (3 points): The alignment between the chosen research paradigms and the theories from The book "How to Research" is moderate, requiring some adjustment or clarification to ensure sufficient data solicitation to answer the research questions of the proposal. 4. Satisfactory (4 points): The chosen research paradigms align reasonably well with the theories from the book "How to Research", supporting adequate data solicitation for research questions or hypotheses of the proposal. 5. Very satisfactory (5 points): The chosen research paradigms closely align with and effectively utilize the theories from The book "How to Research", ensuring highly sufficient data solicitation for comprehensive addressing of research questions or hypotheses of the proposal.
Design	**Question D1:** Does the Methodology section clearly specify research designs (i.e. action research, case study, experiment or survey)?	1. Very unsatisfactory (1 point): The methodology section does not specify any research design or is unclear about the design used. 2. Unsatisfactory (2 points): The methodology vaguely mentions research design without clearly identifying whether it is action research, case study, experiment, or survey. 3. Neutral (3 points): The methodology section mentions research design but lacks clarity or specificity in defining whether it is action research, case study, experiment, or survey. 4. Satisfactory (4 points): The methodology clearly specifies whether it is action research, case study, experiment, or survey, providing a clear understanding of the research design. 5. Very satisfactory (5 points): The methodology precisely and explicitly states whether it is action research, case study, experiment, or survey, ensuring complete clarity on the research design.
	Question D2: Based solely on explicitly stated information, Does the Methodology section clearly justify how the research designs (i.e. action research, case	1. Very unsatisfactory (1 point): The methodology does not explain how the chosen design will gather data relevant to the research questions or hypotheses of the proposal. 2. Unsatisfactory (2 points): The methodology provides vague justification for how the chosen design will solicit data, lacking clarity or relevance to research questions/hypotheses of the proposal. 3. Neutral (3 points): The methodology somewhat justifies how the chosen design will solicit data, but the explanation lacks depth or direct relevance to research questions/hypotheses of the proposal. 4. Satisfactory (4 points): The methodology clearly justifies how the chosen design will solicit data that directly addresses the research questions

	study, experiment or survey) solicit data that will address the research questions or hypotheses?	or hypotheses of the proposal. 5. Very satisfactory (5 points): The methodology extensively and precisely justifies how the chosen design will solicit data, ensuring strong alignment with research questions or hypotheses of the proposal.
	Question D3: Based on the theories of Research designs in Chapter 3 of the book "How to Research", can the research designs (i.e. action research, case study, experiment or survey) chosen solicit data sufficient for answering the research questions or testing the hypotheses?	1. Very unsatisfactory (1 point): There is a significant mismatch between the chosen research design and the theories from The book "How to Research", making it unlikely to solicit sufficient data for answering research questions or testing hypotheses of the proposal. 2. Unsatisfactory (2 points): The alignment between the chosen research design and the theories from the book "How to Research" is weak, limiting the effectiveness of data solicitation for research questions or hypotheses of the proposal. 3. Neutral (3 points): The alignment between the chosen research design and the theories from the book "How to Research" is moderate, requiring some adjustment or clarification to ensure sufficient data solicitation to answer the research questions of the proposal. 4. Satisfactory (4 points): The chosen research design aligns reasonably well with the theories from the book "How to Research", supporting adequate data solicitation for research questions or hypotheses of the proposal. 5. Very satisfactory (5 points): The chosen research design closely aligns with and effectively utilize the theories from the book "How to Research", ensuring highly sufficient data solicitation for comprehensive addressing of research questions or hypotheses of the proposal.
Technique	**Question Te1:** Does the Methodology section clearly specify research technique (i.e. documents, interviews, observations or questionnaires)?	1. Very Unsatisfactory (1 point): The Methodology section does not specify any research techniques. It lacks clarity on the methods employed for data collection, including documents, interviews, observations, or questionnaires. 2. Unsatisfactory (2 points): The Methodology section mentions research techniques in a vague or ambiguous manner. It provides minimal detail or explanation regarding the specific methods used, leaving room for confusion. 3. Neutral (3 points): The Methodology section adequately lists some research techniques such as documents, interviews, observations, or questionnaires employed. However, it lacks thoroughness or specificity in describing how these techniques will be applied. 4. Satisfactory (4 points): The Methodology section clearly specifies research techniques such as documents, interviews, observations, or questionnaires. It provides sufficient detail on how each technique will be implemented, demonstrating a solid understanding of research techniques. 5. Very Satisfactory (5 points): The Methodology section is exceptionally clear and detailed in specifying research techniques. It not only lists documents, interviews, observations, or questionnaires but also provides a comprehensive explanation of how each technique will be utilized, ensuring transparency and understanding.
	Question Te2: Based solely on explicitly stated information, does	1. Very Unsatisfactory (1 point): The Methodology lacks clear statements on how the chosen techniques (documents, interviews, etc.) collect data relevant to research questions or hypotheses. 2. Unsatisfactory (2 points): The Methodology vaguely connects the

	the Methodology section clearly justify how the research technique (i.e. documents, interviews, observations or questionnaires) can solicit data that will address the research questions or hypotheses?	chosen techniques to data collection for research questions or hypotheses, lacking specificity and clarity. 3. Neutral (3 points): The Methodology partially justifies how the chosen techniques (documents, interviews, etc.) solicit data for the research questions of the proposal, needing clearer rationale. 4. Satisfactory (4 points): The Methodology clearly justifies how the chosen techniques (documents, interviews, etc.) solicit data that directly addresses research questions or hypotheses. 5. Very Satisfactory (5 points): The Methodology thoroughly justifies how the chosen techniques (documents, interviews, etc.) solicit data aligned with research questions or hypotheses, demonstrating strong relevance, clarity, and thoroughness.
	Question Te3: Based on the theories of Research techniques in Chapter 3 of the book "How to Research", can the research techniques (i.e. documents, interviews, observations or questionnaires) chosen solicit data sufficient for answering the research questions or testing the hypotheses?	1. Very unsatisfactory (1 point): There is a significant mismatch between the chosen research techniques and the theories from the book "How to Research", making it unlikely to solicit sufficient data for answering research questions or testing hypotheses of the proposal. 2. Unsatisfactory (2 points): The alignment between the chosen research techniques and the theories from the book "How to Research" is weak, limiting the effectiveness of data solicitation for research questions or hypotheses of the proposal. 3. Neutral (3 points): The alignment between the chosen research techniques and the theories from the book "How to Research" is moderate, requiring some adjustment or clarification to ensure sufficient data solicitation to answer the research questions of the proposal. 4. Satisfactory (4 points): The chosen research techniques align reasonably well with the theories from the book "How to Research", supporting adequate data solicitation for research questions or hypotheses of the proposal. 5. Very satisfactory (5 points): The chosen research techniques closely align with and effectively utilize the theories from the book "How to Research", ensuring highly sufficient data solicitation for comprehensive addressing of research questions or hypotheses of the proposal.

APPENDIX B

ChatGPT prompt

Extract the Evaluating questions from the Rubric. DO NOT change them in any way.

You are a proposal reviewer and you must evaluate some section of a proposal based explicitly on the Rubric just provided. Your evaluation must be strictly based on evidence (text) from the proposal. You cannot make any assumption, implication or change. Try to scrutinize the text in the proposal to give accurate evaluation.

Follow the following steps to evaluate the Research title, Research questions, Research hypotheses, Research paradigm, Research design, and Research techniques of the proposal uploaded,

A. STEPS TO EVALUATE RESEARCH TITLE:

1. Extract the Research Title:

- Identify and extract the verbatim of research title from the provided file.

2. Evaluate the Research Title:

- Extract the verbatim of 10 questions from the 'Evaluating Questions' column of the Rubric file (from T1 to T10). Avoid changing the Evaluating questions.

- Use these questions to evaluate the title.

3. Use Evaluation Scale Descriptions:

- Assign scores using the descriptions in the 'Evaluation Scale' column of the Rubric file.

4. Calculate Average Score:

- Calculate the average score for the section.

5. Provide Comments for each Evaluating question

The comments should include:

- Parts of the research title analyzed

- Comments on these parts.

B. STEPS TO EVALUATE RESEARCH QUESTIONS:

1. Extract the verbatim of Research Questions:

- Identify and extract the verbatim of research questions.

2. Evaluate the Research Questions:

- Extract the verbatim of 10 questions from the 'Evaluating Questions' column from the Rubric file (from Q1 to Q10). Avoid changing the Evaluating questions.

- Use these questions to evaluate the Research questions.

3. Use Evaluation Scale Descriptions:

- Assign scores using the 'Evaluation Scale' column of the Rubric file .

4. Calculate Average Score:

- Calculate the average score for the section.

5. Provide Comments for each Evaluating question

The comments should include:

- Parts of the research question analyzed

- Comments on these parts.

C. STEPS TO EVALUATE RESEARCH HYPOTHESES:

1. Extract the verbatim of Research Hypotheses:

- Identify and extract the verbatim of research hypotheses.

- If the Hypotheses are not explicitly stated under heading, marked this entire section as N/A

2. Evaluate the Research Hypotheses:

- Extract the verbatim of 10 questions from the 'Evaluating Questions' column of the Rubric file (from H1 to H10). Avoid changing the Evaluating questions.

- use these questions to evaluate the Hypotheses.

3. Use Evaluation Scale Descriptions:

- Assign scores using the 'Evaluation Scale' column of the Rubric file.

4. Calculate Average Score:

- Calculate the average score for the section.

5. Provide Comments for each Evaluating question

The comments should include:

- Parts of the research hypothesis analyzed

- Comments on these parts.

D. STEPS TO EVALUATE RESEARCH PARADIGM:

1. Extract the verbatim of Research Paradigm:

- Identify and extract the verbatim of whole text relating to research paradigm (quantitative, qualitative, or mixed methodologies).

2. Summarize Theories:

- Summarize theories on research paradigms from the book "How to Research".

3. Evaluate the Research Paradigm:

- Extract the verbatim of 3 questions from the 'Evaluating Questions' column of the Rubric file (from P1 to P3). Avoid changing the Evaluating questions.

- Use these questions to evaluate the Research paradigm.

- Compare with theories from "How to Research".

- Base your evaluation solely on the text in the proposal.

- Compare the language used to introduce and justify the paradigm with the examples from the Rubric to assign correct scores.

4. Use Evaluation Scale Descriptions:

- Assign scores using the 'Evaluation Scale' column of the Rubric file.

5. Calculate Average Score:

- Calculate the average score for the section.

- Bold the text in this row.

6. Provide Comments for each Evaluating question:

The Comments should include:

- Verbatim of the research paradigm.

- Summarized theory on the research paradigm from the book.

- How the research paradigm answers its research questions.

E. STEPS TO EVALUATE RESEARCH DESIGN:

1. Extract the verbatim of Research Design:

- Identify and extract the verbatim of whole text relating to the research design (action research, case study, experiment, or survey).

2. Summarize Theories:

- Summarize theories on research designs from the book "How to Research".

3. Evaluate the Research Design:

- Extract the verbatim of 3 questions from the 'Evaluating Questions' column of the Rubric file (from A1 to A3). Avoid changing the Evaluating questions.

- Use these questions to evaluate the Research Design.

- Compare with theories from "How to Research".

- Base your evaluation solely on the text in the proposal.

- Compare the language used to introduce and justify the research design with the examples from the Rubric to assign correct scores.

4. Use Evaluation Scale Descriptions:

- Assign scores using the 'Evaluation Scale' column of the Rubric file .

5. Calculate Average Score:

- Calculate the average score for the section.

- Bold the text in this row.

6. Provide Comments for each Evaluating question

The comment should include:

- Verbatim of the research design.

- Summarized theory on the research design from the book.

- How the research design answers its research questions.

F. STEPS TO EVALUATE RESEARCH TECHNIQUE:

1. Extract the verbatim of Research Technique:

- Identify and extract the verbatim of research technique (documents, interviews, observations, or questionnaires).

2. Summarize Theories:

- Summarize theories on research techniques from the book "How to Research".

3. Evaluate the Research Technique:

- Extract the verbatim of 3 questions from the 'Evaluating Questions' column of the Rubric file (from T1 to T3). Avoid changing the Evaluating questions.

- Use these questions to evaluate the Research Techniques.

- Compare with theories from "How to Research".
- Base your evaluation solely on the text in the proposal.
- Compare the language used to introduce and justify the research technique with the examples from the Rubric to assign correct scores.

4. Use Evaluation Scale Descriptions:

- Assign scores using the 'Evaluation Scale' column of the Rubric file.

5. Calculate Average Score:

- Calculate the average score for the section.
- Bold the text in this row.

6. Provide Comments for each evaluating question

The Comments should include:

- Verbatim of the research technique.
- Summarized theory on the research technique from the book.
- How the research technique answers its research questions.

G. RETURNING THE RESULTS:

After completing the evaluation:

Generate the Evaluation template (in the form of table) as described below:

Column 1: Section to evaluate: This column lists the sections of the research proposal that need to be evaluated. For example, "Research title" is one such section.

Column 2: Details: This column is intended for the extracted relevant text from those sections. Do not repeat the "Details" content in the subsequent rows for each section. Leave the "Details" column blank for all other rows in one section.

Column 3: Evaluating questions: This column contains specific questions to guide the evaluation of each section. These questions are extracted precisely from the Rubric file. For example, the information in this column should be "1. Is the title sufficiently specific?" …

Column 4: Score: This column is intended for entering the score for each Evaluating question

Column 5: Comments: This column is for the comments related to each Evaluating question. Provide the extract from the proposal to prove your evaluation.