

Research LLM: Using Large Language Models as a Research Companion

Guillaume Guerard*, Sonia Djebali, Maxime Hanus, Mark-Killian Zinenberg

Abstract: Large Language Models (LLM) are increasingly used as research companions, yet guidance for effective and responsible integration into scholarly workflows remains under-specified. This paper presents a three-stage, human-in-the-loop framework supporting (i) proposal development, (ii) research design, and (iii) manuscript preparation. Each stage operationalizes a toolbox of prompting and verification techniques—Context-Grounded Prompting, Graph-of-Thought, Self-Consistency, Contrastive Chain-of-Thought, and Chain-of-Verification—paired with explicit human checkpoints. The framework is instantiated in a case study on mitigating intersectional bias in hiring models, with detailed datasets, objectives, and evaluation metrics, and an exploratory pilot in which Master's students apply the pipeline to thesis proposals. Results indicate that structured prompting improves traceability and broadens the set of considered alternatives, while verification steps curb overconfident errors. Prompts, artifacts, and rubrics are made available to support replication. The paper concludes with guidance on ethical deployment and limitations, emphasizing the primacy of human judgment in validating claims and shaping scholarly contributions.

Keywords: Artificial Intelligence; Large Language Model; Research Methodology

1 INTRODUCTION

Artificial Intelligence (AI) is fundamentally transforming academic research by enhancing knowledge discovery, data synthesis, and methodological structuring. The exponential growth of scholarly literature presents a significant challenge, making it increasingly difficult for researchers to filter, analyze, and contextualize relevant information [16]. AI-powered models and dedicated research companions, such as <https://answerthis.io> and LitLLM, address this challenge by providing real-time assistance in idea generation, literature summarization, and the structuring of research methodologies, thereby improving the efficiency of systematic reviews and reducing cognitive load [2].

Traditional research workflows remain time-intensive, requiring manual effort to formulate research questions, synthesize literature, and identify novel research gaps—a difficulty exacerbated in multidisciplinary contexts where integrating diverse perspectives can be overwhelming [10, 14]. AI mitigates these inefficiencies by contributing to study design, methodology validation, and bias detection [15]. Through retrieval-augmented insights and automated comparisons, these tools enable more data-driven and logically sound methodological choices [6].

Despite these advantages, the utility of AI is contingent upon rigorous human oversight. AI-generated recommendations are primarily pattern-based and can lack the contextual adaptation, ethical reasoning, and domain-specific accuracy essential for high-quality scholarship [1]. An over-reliance on AI outputs without critical evaluation risks the propagation and amplification of biases and inaccuracies. AI systems learn from human-generated data and can reproduce and even amplify the biases contained within it, which can in turn create a feedback loop that increases bias in the humans interacting with them [9]. This dependence on AI can also lead to cognitive offloading, where the delegation of mental tasks to technology may result in the deterioration of critical thinking and analytical skills over time [8].

To address this, a human-in-the-loop approach is critical, positioning AI as a collaborative augmentation tool rather than an autonomous decision-maker. This model leverages the complementary strengths of humans and AI, allowing machines to handle data-intensive work while humans provide oversight, interpretation, and final judgment [21]. By integrating human expertise and accountability, this collaborative process helps to mitigate risks and ensures that AI-driven processes align with human values, leading to more robust and trustworthy outcomes.

This model ensures that human researchers retain ultimate control, validating AI-generated outputs, refining nuanced arguments, and preserving the intellectual rigor required to align research with scholarly standards.

This paper introduces a systematic framework for AI-assisted research that formalizes the collaborative process between scholar and machine, grounding the approach in the long-standing scholarly tradition of using methodological frameworks to ensure rigor, transparency, and replicability. Established methodologies such as Systematic Literature Reviews (SLRs), often guided by protocols like PRISMA [11], and Framework Analysis for qualitative data [13], demonstrate the value of a structured approach to managing and interpreting complex information. Extending these principles, the proposed framework presents a structured methodology where an iterative feedback loop is used to discover, organize, and refine academic work. This process is designed to create a synergy between AI's efficiency and the researcher's indispensable domain expertise, in line with emerging models that aim to augment, not replace, scholarly judgment. This approach addresses the need for structure emphasized in recent work, including conceptual frameworks for AI-assisted literature reviews [14] and broader models that delineate human and AI roles to ensure quality and ethical oversight. By building on these established synthesis and analysis methodologies, the framework provides a structured pathway for developing a symbiotic relationship between researchers and AI, leveraging the strengths of each to advance scholarly inquiry.

The following sections detail this framework, beginning with the role of the AI as an iterative research companion, followed by specific workflows for research proposal development, literature synthesis, and manuscript preparation. This paper is concluded by summarizing our findings and emphasizing the indispensable role of human-driven validation in the age of AI-assisted research.

2 CORE CONCEPTS: A GLOSSARY OF ADVANCED PROMPTING TECHNIQUES

Effective AI-assisted research hinges on the strategic application of advanced prompting techniques. These methods provide a structured interface for interacting with AI models, guiding their reasoning processes to enhance output quality and ensure methodological rigor. Each technique serves a specific function, from grounding outputs in established literature to validating claims and mitigating cognitive biases. The core techniques integral to the research framework proposed in this paper are defined in Tab. 1.

Table 1 A Glossary of AI Prompting Techniques for Academic Research

Technique	Description	Primary Use Case
Context-Grounded Prompting (CGP)	Grounds AI outputs in user-provided documents, reducing hallucinations and ensuring factual accuracy.	Literature synthesis, content validation, ensuring recommendations are based on seminal or relevant papers.
Graph-of-Thought (GoT) Prompting [20]	Visually structures connections between concepts, mapping relationships between variables, theories, and methodologies.	Structuring research designs, conceptual mapping, identifying literature gaps.
Self-Consistency Prompting [17]	Generates multiple reasoning paths or outputs and selects the most coherent and logical one.	Refining research questions, validating research designs, generating robust paper outlines
Contrastive Chain-of-Thought (CCoT) [5]	Enables structured, side-by-side comparisons of multiple methodologies, theories, or viewpoints.	Methodological selection (e.g., comparing quantitative vs. qualitative approaches).
Chain-of-Verification (CoVe) Prompting [7]	Improves reliability by systematically cross-checking facts, questioning assumptions, and generating counterexamples.	Bias detection, robustness analysis, validation of claims and research gaps.
Automatic Reasoning and Tool Usage (ART) [12]	Applies step-by-step logic and can integrate external tools to perform feasibility checks or complex reasoning tasks.	Evaluating methodological feasibility, assessing ethical considerations, checking data reliability.
Chain-of-Thought (CoT) Reasoning [19]	Guides the AI through a step-by-step reasoning process to break down complex problems into logical, sequential parts.	Drafting methodologically sound arguments, ensuring transparency in the reasoning process.
Zero-Shot & Few-Shot Prompting [18]	Generates insights without prior examples (Zero-Shot) or with a few structured examples (Few-Shot) to guide its response.	Initial topic exploration (Zero-Shot), refining tone and style in writing (Few-Shot).

Beyond their generative capabilities, these techniques are crucial for implementing the "human-in-the-loop" model, which acts as a safeguard against common pitfalls in AI-assisted research. Cognitive biases such as confirmation bias (favoring supportive information) and anchoring bias (over-relying on initial suggestions) can be actively countered. For instance, a researcher can employ CoVe to compel the AI to seek counterexamples to its own claims, use CCoT to force an impartial comparison between multiple perspectives, or leverage Self-Consistency Prompting to evaluate a range of possible solutions instead of a single output. By integrating these validation techniques into the workflow, researchers can augment their expertise with AI, ensuring the final work is critically sound and meets the highest scholarly standards.

3 THE THREE-STAGE AI-ASSISTED RESEARCH PIPELINE

AI-powered research operates not as a linear process, but as a continuous feedback loop that enhances conceptual clarity and methodological rigor. By structuring the research process into three distinct stages—proposal development, methodological design, and manuscript preparation—the AI research companion facilitates systematic knowledge development. The following sections detail these structured pipelines, emphasizing the synergy between advanced AI prompting techniques and essential human oversight at critical checkpoints.

3.1 Stage 1: Research Proposal Development

The initial stage focuses on transforming a broad area of interest into a focused, well-grounded, and defensible research proposal. This workflow ensures that research questions are not only innovative but also firmly rooted in the existing body of academic literature (see Fig. 1).

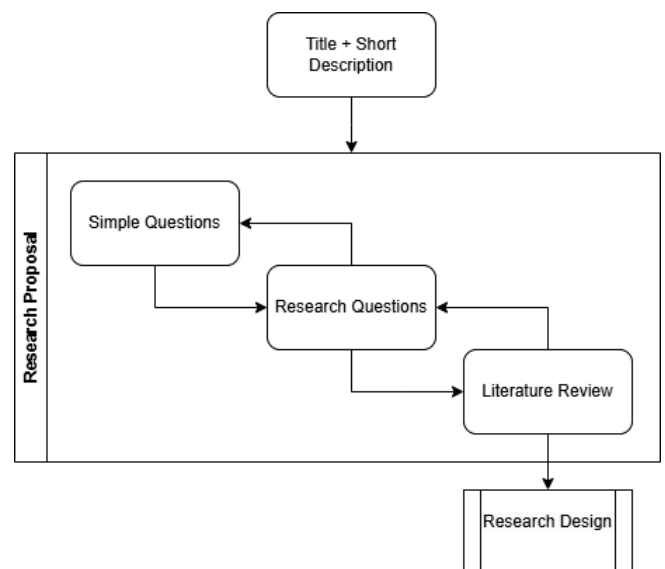


Figure 1 Overview of the iterative research proposal workflow, moving from broad literature synthesis to a refined research question

Conceptualization and Grounding (CGP): The process begins with establishing a broad but accurate understanding of the research topic. The researcher uses CGP

by providing the AI with a curated set of seminal papers, keywords, or preliminary notes. The AI retrieves and synthesizes this literature to identify major themes, foundational theories, key findings, and established definitions. This grounds the entire project in verified academic sources, ensuring relevance and preventing early-stage misconceptions.

Knowledge Mapping and Gap Identification (GoT):

With a grounded understanding of the field, the next step is to identify a clear, actionable research gap. Using GoT prompting, the AI constructs a visual map of the relationships between concepts, theories, and methodologies identified in the synthesized literature. This conceptual map allows the researcher to visualize the current state of knowledge, identify underexplored connections, note conflicting findings, and articulate a preliminary research gap with precision.

Research Question Refinement (Self-Consistency & CoVe): Based on the identified gap, the researcher prompts the AI to generate multiple versions of a potential research question using Self-Consistency Prompting. This technique produces a diverse set of alternatives. Each candidate question is then subjected to rigorous validation using CoVe. In this step, the researcher prompts the AI to systematically challenge the assumptions, scope, and feasibility of each question. This adversarial process pressure-tests the questions for robustness, resulting in a final research question that is precise, impactful, and methodologically answerable.

Human Checkpoint: At the conclusion of this stage, the researcher's intervention is paramount. The human expert must: (1) Critically evaluate the AI-synthesized literature to ensure the sources are appropriate and the interpretation is accurate; (2) Validate that the identified research gap is relevant and significant within the discipline; and (3) Make the final selection of the research question, ensuring it aligns with the project's long-term goals and their own domain expertise. This checkpoint prevents over-reliance on AI suggestions and ensures the proposal is built on a solid intellectual foundation.

3.2 Stage 2: Methodological Design and Validation

This stage translates the validated research question into a robust, transparent, and defensible research design. It provides a systematic framework for constructing, evaluating, and refining the study's methodology to ensure it is sound, aligned with disciplinary standards, and resilient to bias (see Fig. 2).

Methodological Analysis (CGP & GoT): The researcher initiates this phase by using CGP to provide the AI with existing studies that have similar research aims. The AI deconstructs and synthesizes the methodologies from these papers, identifying common research designs, data collection techniques, and analytical frameworks. To deepen this analysis, GoT is used to map the components of these designs (e.g., variables, controls, statistical tests), clarifying the intricate relationships and dependencies within established methodological precedents.

Comparative Selection (CCoT & ART): With a clear understanding of existing methods, the AI is prompted to generate and compare multiple viable research design

alternatives. Using CCoT, the AI creates a structured, side-by-side comparison of different approaches (e.g., experimental vs. observational, quantitative vs. qualitative). For each option, ART is applied to systematically assess its feasibility, ethical implications, resource requirements, and methodological trade-offs (e.g., internal vs. external validity). This produces a ranked recommendation, empowering the researcher to make an informed choice.

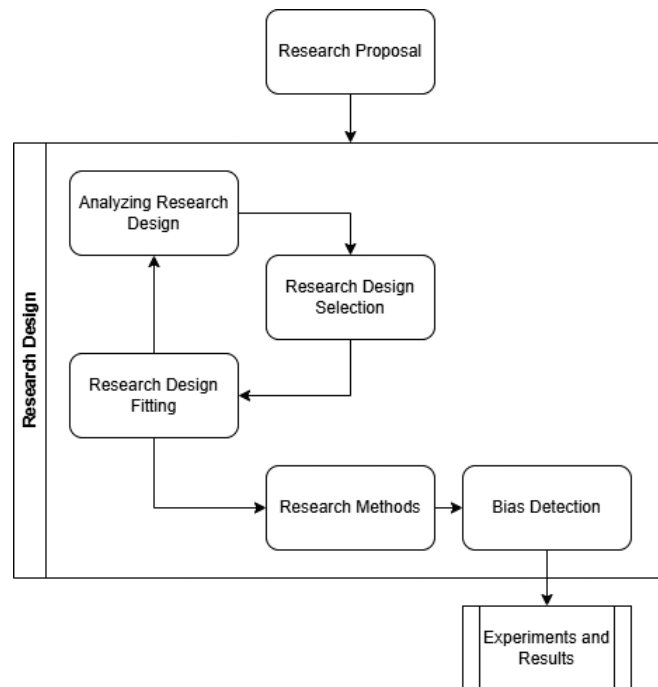


Figure 2 The iterative research design workflow, focused on selecting and validating a methodologically sound approach.

Bias Detection and Robustness Check (CoVe & ART): Before finalizing the design, a critical stress test is performed. The researcher uses CoVe to systematically challenge the chosen design's underlying assumptions and proactively identify potential sources of bias in sampling, data collection, or interpretation. ART can be used in parallel to apply formal logic to probe for hidden methodological weaknesses or inconsistencies. This critical validation step ensures the final design is as rigorous and impartial as possible.

Human Checkpoint: Human oversight is indispensable for methodological validation. The researcher must: (1) Scrutinize the AI's feasibility analysis, bringing practical, real-world knowledge of costs, timelines, and data accessibility; (2) Make the final decision on the research design, balancing the AI's logical recommendations with their own disciplinary expertise and intuition; and (3) Assume full responsibility for the ethical integrity of the design, ensuring it meets institutional and scholarly standards.

3.3 Stage 3: AI-Assisted Manuscript Preparation

The final stage streamlines the transformation of research findings into a polished, publication-ready paper. This workflow integrates AI for efficiency in structuring,

drafting, and refining, while preserving the researcher's intellectual ownership and narrative control (see Fig. 3).

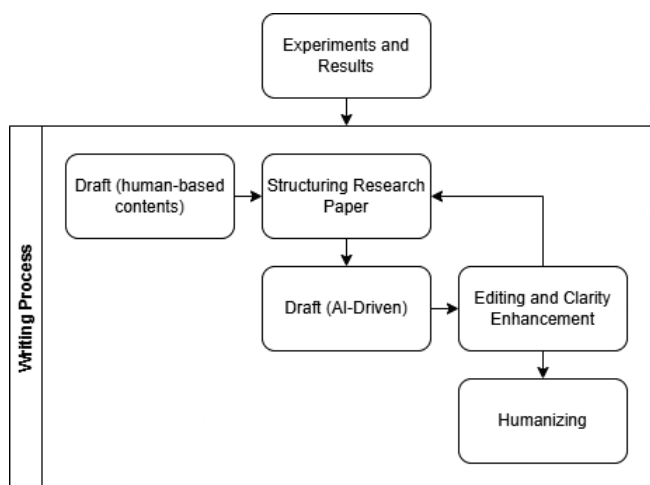


Figure 3 The iterative research writing workflow, balancing AI-driven drafting with essential human refinement and intellectual contribution.

Structural Outlining (CGP, GoT & Self-Consistency): The process begins with creating a comprehensive and logical outline. The AI uses CGP to retrieve standard paper structures aligned with target journal guidelines (e.g., IMRaD). GoT is then used to map the narrative flow, ensuring the core arguments connect logically from the introduction to the conclusion. Finally, Self-Consistency Prompting can generate several complete outline variations, allowing the researcher to select and refine the structure that best highlights the study's contributions.

AI-Powered Drafting (CoT & Few-Shot): With an approved outline, the AI assists in generating initial draft content. For complex sections requiring logical transparency, such as the methodology or discussion, CoT reasoning is used to produce well-structured, step-by-step paragraphs. To align the text with the desired academic voice, Few-Shot prompting is then applied, where the researcher provides high-quality writing samples to guide the AI's tone, style, and terminology.

Iterative Refinement and Humanization (Human Review): This final sub-step is fundamentally human-driven and iterative. The researcher directs the AI to perform specific revisions, such as improving clarity, checking for inconsistencies, and enhancing readability by removing repetitive or formulaic phrasing. However, the core intellectual work rests with the researcher, who is responsible for weaving the AI-generated components into a cohesive and compelling narrative, integrating domain-specific nuance, and ensuring the final arguments are persuasive and insightful.

Human Checkpoint: The final checkpoint is the most critical, as it determines the manuscript's intellectual integrity. The researcher must: (1) Conduct a thorough review to verify the technical accuracy of all claims and data interpretations; (2) Assume full ownership of the narrative, ensuring the paper's voice is authentic and the arguments are novel; and (3) Infuse the manuscript with the high-level

critical insights and scholarly perspective that only a human expert can provide. This final humanization phase is what elevates an AI-assisted draft into a genuine contribution to academic knowledge.

4 ILLUSTRATIVE CASE STUDY: MITIGATING BIAS IN AI HIRING ALGORITHMS

To demonstrate the practical application of the three-stage pipeline, this section provides a concrete case study of a machine learning research project. The project's goal is to design and validate a novel technique for mitigating intersectional bias in AI-powered hiring algorithms. Each stage includes examples of user prompts and illustrative AI responses, showcasing the collaborative workflow.

4.1 Stage 1: Research Proposal Development

Conceptualization and Grounding (CGP): "Based on references Buolamwini & Gebru (2018) [22] and Kearns & Roth (2019) [23], synthesize the primary themes and established definitions related to algorithmic bias."

Illustrative AI Response: "The provided literature establishes two primary themes: (1) Algorithmic systems often amplify societal biases present in training data, and (2) 'Fairness' is not a monolithic concept, with definitions like demographic parity and equalized odds often being mutually exclusive. Key definitions include 'intersectional bias,' where performance disparities are significantly worse for individuals in subgroups defined by multiple attributes (e.g., women of color)."

Knowledge Mapping and Gap Identification (GoT): "Create a conceptual map showing the relationships between different fairness mitigation techniques (pre-processing, in-processing, post-processing) and the fairness definitions they typically address. Highlight areas with weaker connections."

Illustrative AI Response: "A knowledge map shows strong connections between pre-processing techniques (e.g., re-weighting) and demographic parity. Similarly, in-processing techniques are often linked to equalized odds. However, the map reveals a weak link between most established techniques and the explicit mitigation of 'intersectional bias,' suggesting a significant research gap."

Research Question Refinement (Self-Consistency & CoVe): "Generate three potential research questions to address the gap in intersectional bias mitigation. For each, list its primary challenge or assumption."

Illustrative AI Response: "1) Question: How can adversarial debiasing be adapted to mitigate intersectional bias in hiring models? Challenge: Adversarial training can be unstable and may harm model accuracy. 2) Question: What is the impact of synthetic data generation for rebalancing intersectional subgroups on fairness? Challenge: Synthetic data may not capture the true characteristics of underrepresented groups. 3) Question: To what extent can multi-objective optimization balance predictive accuracy with intersectional fairness metrics? Challenge: Defining a robust intersectional fairness metric for optimization is non-trivial."

4.2 Stage 2: Methodological Design and Validation

Methodological Analysis (CGP & GoT): "Analyze the methodologies of the attached papers on fairness in hiring. Map their experimental components, including datasets, model architectures, and evaluation metrics."

Illustrative AI Response: "The analysis shows a common pattern: most studies use the 'Adult Census' or 'COMPAS' datasets with models like Logistic Regression or Gradient Boosting. Evaluation relies on metrics like Disparate Impact and Statistical Parity Difference. A conceptual map indicates that the 'Folktables' dataset, which is better suited for intersectional analysis, is underutilized in this context."

Comparative Selection (CCoT & ART): "Compare an experimental design using the public Folktables dataset against a case study design using proprietary data from a tech company. Evaluate each based on internal validity, external validity, and feasibility."

Illustrative AI Response: "Experimental Design: High internal validity and reproducibility due to the controlled public dataset. Lower external validity as Folktables may not represent a specific company's hiring pipeline. Case Study Design: High external validity and real-world impact. Low internal validity due to confounding variables and lack of reproducibility. Feasibility is lower due to data privacy and access constraints."

Bias Detection and Robustness Check (CoVe & ART): "We have selected the experimental design. Challenge this choice. Identify the three most significant potential sources of bias when using the Folktables dataset to study hiring bias."

Illustrative AI Response: "1. Measurement Bias: The target variable ('income >\$50K') is a poor proxy for a successful hire. 2. Sampling Bias: The dataset is not representative of all industries, limiting generalizability. 3. Construct Validity: The available features may not include key qualifications relevant to hiring, forcing the model to rely on demographic proxies."

4.3 Stage 3: AI-Assisted Manuscript Preparation

Structural Outlining (CGP, GoT, & Self-Consistency): "Generate a paper outline based on the IEEE conference format for our study on mitigating intersectional bias with multi-objective optimization."

Illustrative AI Response: "The AI generates a standard IMRaD outline formatted for IEEE: I. Introduction; II. Related Work (A. Fairness Metrics, B. Mitigation Techniques); III. Methodology (A. Dataset and Preprocessing, B. Model Architecture, C. Proposed Optimization Algorithm); IV. Experiments and Results; V. Discussion; VI. Conclusion."

AI-Powered Drafting (CoT & Few-Shot): "Draft the 'Methodology' section. Step 1: Describe the Folktables dataset. Step 2: Detail the XGBoost model architecture. Step 3: Explain our multi-objective loss function. Here is an example methodology section from another paper to guide the academic tone: [paste example]."

Illustrative AI Response: "The AI generates a multi-paragraph draft. 'III. METHODOLOGY. A. Dataset and Preprocessing: We use the 2018 Folktables dataset from the American Community Survey... B. Model Architecture: Our classification model is an XGBoost classifier with... C. Proposed Fairness Intervention: We introduce a multi-objective loss function defined as $L = \alpha L_{\{CE\}} + (1 - \alpha) L_{\{IF\}}$..."

Iterative Refinement and Humanization (Human Review): "Review the following draft of the 'Discussion' section. Improve the flow and remove repetitive phrases like 'It is important to note that...'. Ensure the claims are directly supported by the results mentioned."

Illustrative AI Response (Revised Text): "It is important to note that our model showed a 2% drop in accuracy. This finding is significant. It is also important to note that fairness for Black women improved by 15%."

The revised text become "While our intervention led to a modest 2% decrease in overall accuracy, it achieved a substantial 15% improvement in fairness for Black women, demonstrating a clear and impactful trade-off that prioritizes equity for the most vulnerable intersectional subgroup."

5 LIMITATIONS AND ETHICAL CONSIDERATIONS

While the AI-assisted research pipeline offers powerful tools for enhancing scholarly productivity, its responsible implementation demands a critical awareness of its limitations and a firm commitment to ethical academic practice. The framework presented in this paper is designed to leverage pre-trained LLMs as assistants, not as autonomous creators of knowledge. The efficacy and integrity of the entire process, therefore, depend on the researcher's ability to maintain rigorous oversight over tools they did not create.

Factual Accuracy and the Risk of "Hallucination". A primary limitation of LLMs is their lack of genuine comprehension; they operate on sophisticated pattern recognition, not true reasoning [17]. This can lead to outputs that are grammatically correct and stylistically plausible but factually inaccurate, logically flawed, or entirely nonsensical—a phenomenon known as "hallucination." A significant risk for academic writing is the generation of fabricated citations, references, or data points that appear authentic but are completely fictitious [3]. This makes human verification a non-negotiable step. Every claim, citation, and piece of data generated by an AI must be meticulously fact-checked against reliable sources to ensure the manuscript's scholarly integrity.

Bias Amplification in AI Outputs. LLMs are fundamentally shaped by the vast, static datasets on which they were trained, and these datasets inevitably reflect existing societal and historical biases [4]. Consequently, AI models can inadvertently replicate and even amplify these biases in their outputs. For example, an AI's synthesis of literature may overemphasize dominant theoretical perspectives while underrepresenting or mischaracterizing marginalized viewpoints. The researcher bears the ultimate responsibility to critically evaluate all AI-generated content

for hidden biases, actively anticipate how they might manifest, and correct for them to ensure a balanced and equitable scholarly narrative.

Authorship, Accountability, and Plagiarism. A firm consensus has emerged in the academic community: an LLM cannot be credited as an author. Authorship requires accountability and responsibility for the work's integrity, which a machine cannot assume. The human researcher who employs AI tools is fully and solely responsible for the final manuscript's content, originality, and accuracy. To maintain transparency, many journals and institutions now mandate explicit disclosure of AI tool usage. Furthermore, researchers must diligently check AI-generated text for unintentional plagiarism, as models may reproduce passages from their training data verbatim without proper attribution. The boundary between using AI as a tool and presenting its output as one's own intellectual work must be scrupulously managed to uphold academic integrity.

Over-Reliance and the Risk of Stifled Creativity. While AI can accelerate structured tasks, over-reliance on it poses a risk to the core of scholarly inquiry: critical thinking and intellectual innovation. Using a handful of dominant LLMs for fundamental tasks like identifying research gaps or structuring arguments could lead to a homogenization of ideas and a convergence of methodologies across the academic landscape. This may stifle the creativity, serendipity, and diversity of thought that fuel groundbreaking research. The primary role of AI should be to handle rote or organizational tasks, thereby freeing up the human researcher to focus on novel insights, creative problem-solving, and the development of a unique intellectual voice. AI should augment, not replace, the critical and creative faculties of the scholar.

6 PILOT STUDY: FRAMEWORK APPLICATION WITH MASTER'S-LEVEL RESEARCHERS

To assess the real-world feasibility and practical utility of the three-stage framework, an exploratory pilot study was conducted with a cohort of Master's students (both M1 and M2 levels) in a research-focused track. The students were tasked with using the AI-assisted pipeline to develop an initial proposal for their master's thesis. The objective of this study was not to derive quantitative performance metrics but to adopt a qualitative and observational approach. The study aimed to identify the framework's pedagogical value, its practical challenges, and the operational reality of the "Human Checkpoint" when implemented by novice researchers. The following sections synthesize the key observations from this pilot, organized thematically.

6.1 Theme 1: Scaffolding the Research Process for Novice Researchers

Conceptualization and Grounding (Stage 1): Students reported that the structured process of using CGP with a small set of foundational papers provided a tangible entry point into a dense body of literature. The AI-driven synthesis

of key themes and definitions furnished them with a foundational vocabulary and a conceptual map.

Methodological Literacy (Stage 2): Stage 2 proved particularly valuable for students struggling to connect a research question to a concrete research design. The use of CCoT to generate side-by-side comparisons of different methodologies (e.g., a qualitative case study vs. a quantitative survey) transformed an abstract decision into a structured analysis of trade-offs. This process appeared to enhance students' understanding of concepts like internal vs. external validity by forcing consideration of the practical implications of each choice.

6.2 Theme 2: The "Human Checkpoint" as a Pedagogical Tool

Observations from the study strongly validated the necessity of the "Human Checkpoint", revealing its dual function as both a quality control mechanism and a critical pedagogical tool. A consistent pattern of anchoring bias was observed, wherein students tended to over-rely on the initial AI-generated outputs, particularly in the early stages of research.

Overcoming Anchoring Bias: Without explicit guidance, students frequently demonstrated a reluctance to challenge or significantly refine the AI's first suggestions. For example, several students initially accepted an AI-generated research question without critically evaluating its scope or feasibility.

The Supervisor as the Human Checkpoint: The intervention of a human supervisor proved paramount. One student, following an AI suggestion, drafted a proposal to study hiring bias across all Fortune 500 companies---a project far too broad for a master's thesis. The subsequent human checkpoint with their supervisor was essential. The supervisor guided the student to use the AI's initial output as a starting point, but then to apply critical constraints (e.g., focusing on a single industry or a specific type of algorithm). This transformed the checkpoint from a simple "go/no-go" gate into a teachable moment about the practical realities of research scoping. In this context, the framework facilitated a more focused and effective conversation between the student and supervisor.

6.3 Theme 3: Practical Challenges and Directions for Future Work

The application of the framework by novice researchers also highlighted several practical challenges that inform the need for future research.

Prompt Engineering as a Skill Barrier: Students often struggled to write effective prompts. Their initial prompts were frequently too vague, leading to generic or irrelevant AI outputs. This suggests that the successful application of this framework depends on a secondary skill---prompt engineering---which itself requires training and practice.

The "Credibility Illusion": The highly fluent and authoritative tone of AI-generated text often created a "credibility illusion". Students found it difficult to distinguish

between a well-written but methodologically flawed suggestion and a genuinely sound one. This underscores the cognitive burden that remains on the human researcher to possess enough domain knowledge to critically evaluate AI outputs.

These observations lead directly to avenues for future study. While this pilot confirmed the framework's feasibility, the next stage of research must focus on developing and measuring the effectiveness of pedagogical interventions. Future studies should aim to create structured training materials for students on prompt engineering and critical AI output evaluation. Furthermore, developing robust metrics to assess the quality of a research proposal (e.g., clarity, feasibility, novelty) will be essential for any future quantitative comparison between this framework and traditional research methods.

7 DISCUSSION

The proposed framework positions AI-assisted research as a structured collaboration, balancing technological affordances with established principles of scholarly rigor. By formalizing a HITL model, the framework defines AI's role as one of augmentation, not autonomous execution. This distinction is critical for mitigating risks inherent in automation, such as the uncritical adoption of AI-generated outputs or the amplification of algorithmic biases. The model's emphasis on active intellectual oversight directly counters the risk of cognitive offloading, a phenomenon where reliance on technology can diminish a researcher's critical skills.

A primary implication of this approach is the promotion of transparency and replicability in a domain where processes can become opaque. In a manner analogous to the PRISMA guidelines for systematic reviews, the framework provides a clear methodology that can be documented, scrutinized, and reproduced. This structured process strengthens the credibility of the research findings and serves as a vital pedagogical model for training the next generation of researchers in the responsible and effective use of AI tools.

The framework's most significant limitation, however, is the methodological and evaluative challenge of defining its success. There are currently no standardized Key Performance Indicators (KPIs) to objectively compare the quality of its outputs against that of traditional, human-only research. While metrics of efficiency like time saved are easily measured, they fail to capture the core dimensions of scholarly value: the novelty of insights, the depth of critical analysis, or the serendipity of discovery. This challenge is amplified by the framework's intrinsic dependence on researcher expertise. The quality of the final output is inextricably linked to the user's ability to craft precise prompts and critically vet AI-generated content. Consequently, assessing the framework's definitive advantage remains a complex, qualitative endeavour, moving beyond the question of "faster" to the much harder-to-define "better."

Addressing these limitations requires further research, particularly empirical validation through cross-disciplinary case studies. Such work would help refine the framework's stages and identify discipline-specific adaptations. Future

efforts could also focus on developing integrated software tools that guide researchers through the framework, facilitating the systematic documentation required for transparent reporting and, eventually, comparative analysis.

8 CONCLUSION

This paper introduced a systematic, three-stage framework that operationalizes the collaboration between researchers and large language models. By structuring the research process into discrete phases—Proposal Development, Data Analysis and Interpretation, and Manuscript Drafting—the framework provides a replicable workflow for leveraging AI to enhance scholarly output. Its principal contribution is a structured methodology that integrates AI as a specialized assistant for knowledge discovery and content generation while ensuring that logical coherence and intellectual integrity are maintained at every step.

The central thesis of this work is that the efficacy of AI in research hinges upon the non-negotiable role of human intellect. The framework is intentionally designed with critical "Human Checkpoints" where the researcher's domain expertise, critical judgment, and ethical oversight are not just beneficial, but indispensable. AI-generated content, while often structurally coherent, lacks genuine comprehension and remains susceptible to factual inaccuracies and embedded biases. The AI can provide the scaffolding, but it is the human scholar who must act as the architect—validating claims, challenging assumptions, and infusing the work with the novel interpretation that constitutes a true contribution to knowledge.

Ultimately, the responsible integration of AI into research workflows presents a profound opportunity to deepen analysis, optimize efficiency, and expand the frontiers of inquiry. As these technologies evolve, the challenge for the academic community will be to co-evolve with them, refining collaborative frameworks that balance the automation of tasks with the amplification of intellect. This synergy does not replace the scholar; it redefines the scholar's role toward one of critical inquiry director, arbiter of validity, and architect of meaning in an increasingly complex information landscape.

9 REFERENCES

- [1] Albadarin, Y., et al. (2024). A systematic literature review of empirical research on ChatGPT in education. *Discover Education*, 3(1), 60. <https://doi.org/10.1007/s44217-024-00138-2>
- [2] Bolanos, F., et al. (2024). Artificial intelligence for literature reviews: Opportunities and challenges. *Artificial Intelligence Review*, 57(10), 259. <https://doi.org/10.1007/s10462-024-10902-3>
- [3] Chelli, M., et al. (2024). Hallucination rates and reference accuracy of ChatGPT and Bard for systematic reviews: Comparative analysis. *Journal of Medical Internet Research*, 26(1), e53164. <https://doi.org/10.2196/53164>
- [4] Chen, J., et al. (2024). *When large language models meet personalization: Perspectives of challenges and opportunities*. *World Wide Web*, 27(4), 42. <https://doi.org/10.1007/s11280-024-01276-1>
- [5] Chia, Y. K., et al. (2023). Contrastive chain-of-thought prompting. *arXiv preprint arXiv:2311.09277*.

- [6] Chubb, J., et al. (2022). Speeding up to keep up: Exploring the use of AI in the research process. *AI & Society*, 37(4), 1439–1457. <https://doi.org/10.1007/s00146-021-01259-0>
- [7] Dhuliawala, S., et al. (2023). Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*. <https://doi.org/10.18653/v1/2024.findings-acl.212>
- [8] Gerlich, M. (2025). *AI tools in society: Impacts on cognitive offloading and the future of critical thinking*. *Societies*, 15(1), 6. <https://doi.org/10.3390/soc15010006>
- [9] Glickman, M., & Sharot, T. (2025). How human–AI feedback loops alter human perceptual, emotional and social judgements. *Nature Human Behaviour*, 9(2), 345–359. <https://doi.org/10.1038/s41562-024-02077-2>
- [10] de la Torre-López, J., et al. (2023). Artificial intelligence to automate the systematic review of scientific literature. *Computing*, 105(10), 2171–2194. <https://doi.org/10.1007/s00607-023-01181-x>
- [11] Page, M. J., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372. <https://doi.org/10.1136/bmj.n71>
- [12] Paranjape, B., et al. (2023). Art: Automatic multi-step reasoning and tool-use for large language models. *arXiv preprint arXiv:2303.09014*.
- [13] Ritchie, J., et al. (2013). *Qualitative research practice: A guide for social science students and researchers*. SAGE.
- [14] Schryen, G., et al. (2025). Exploring the scope of generative AI in literature review development. *Electronic Markets*, 35(1), 1–26. <https://doi.org/10.1007/s12525-025-00754-2>
- [15] Scott-Kennel, J., et al. (2025). Artificial intelligence in academic research: Contributor, constructivist or cheat? *Journal of Marketing Theory and Practice*, 1–22. <https://doi.org/10.1080/10696679.2025.2457672>
- [16] Shahzad, T., et al. (2025). A comprehensive review of large language models: Issues and solutions in learning environments. *Discover Sustainability*, 6(1), 27. <https://doi.org/10.1007/s43621-025-00815-8>
- [17] Wang, X., et al. (2022). Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.
- [18] Wang, Y., et al. (2020). Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys (CSUR)*, 53(3), 1–34. <https://doi.org/10.1145/3386252>
- [19] Wei, J., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, 24824–24837.
- [20] Yao, Y., et al. (2024). GoT: Effective Graph-of-Thought Reasoning in Language Models. *Findings of the Association for Computational Linguistics: NAACL 2024*, 2901–2921. <https://doi.org/10.18653/v1/2024.findings-naacl.183>
- [21] Zhou, L., et al. (2021). Intelligence augmentation: Towards building human–machine symbiotic relationship. *AIS Transactions on Human-Computer Interaction*, 13(2), 243–264.
- [22] Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. *Proceedings of Machine Learning Research*, 81, 77–91. <https://proceedings.mlr.press/v81/buolamwini18a.html>
- [23] Kearns, M., & Roth, A. (2019). *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press.

Authors' contacts:**Guillaume Guerard**

(Corresponding author)
De Vinci Higher Education, De Vinci Research Center,
92916 Paris, France
guillaume.guerard@devinci.fr

Sonia Djebali

De Vinci Higher Education, De Vinci Research Center,
92916 Paris, France

Maxime Hanus

Ecole Supérieure d'Ingénieurs Léonard de Vinci,
92916 Paris, France

Mark-Killian Zinenberg

Ecole Supérieure d'Ingénieurs Léonard de Vinci,
92916 Paris, France