



Universiteit  
Leiden

**Master Media Technology**

# **Supervised by the Machine: Evaluating AI as Supervisor Through a Nail Biting Detection Study**

Name: Yunshan Cai

Student ID: s3862607

Date: 25/08/2025

1st supervisor: Bas Haring

2nd supervisor: Claude and ChatGPT

Master's Thesis in Media Technology

Leiden Institute of Advanced Computer Science (LIACS)


Leiden University

Niels Bohrweg 1

2333 CA Leiden

The Netherlands

# Abstract



This thesis examines GenAI's potential as a research supervisor through a self-experimentation methodology. The research employs a nested design: GenAI supervises this entire graduate thesis, which itself investigates GenAI supervision capabilities. To evaluate GenAI's technical supervision abilities, a nail biting detection case study serves as the nested technical project under AI supervision. More specifically, the research evaluates AI performance across functional, critical thinking, and emotional support dimensions using ChatGPT and Claude. The findings show that GenAI excels in the functional dimension, particularly in topic selection, literature review, coding-related tasks, and writing support. GenAI demonstrates moderate effectiveness in emotional support through validation-based responses. However, GenAI fundamentally lacks critical thinking supervision capabilities, high-level human judgment, and the depth of domain expertise essential for effective academic mentorship. Results suggest that optimal research supervision requires a hybrid approach that leverages GenAI's advantages for specific supervision tasks while preserving essential human elements of critical evaluation, strategic guidance, and authentic academic mentorship.

# Contents

<b>1</b>	<b>Introduction</b>	<b>4</b>
<b>2</b>	<b>Literature Review</b>	<b>6</b>
2.1	Research Supervision . . . . .	6
2.1.1	Supervisory Roles and Responsibilities . . . . .	6
2.2	AI in Education . . . . .	8
2.2.1	GenAI in Higher Education . . . . .	8
2.3	GenAI in Research Supervision . . . . .	9
2.4	Research Gaps . . . . .	10
<b>3</b>	<b>Methods</b>	<b>11</b>
3.1	Research Questions . . . . .	11
3.2	Research Structure . . . . .	11
3.3	Transition from Human Supervision . . . . .	12
3.4	Choice of GenAI Models . . . . .	12
3.5	Implementation of Supervision Tasks . . . . .	13
3.6	Research Evaluation . . . . .	14
<b>4</b>	<b>Analysis</b>	<b>16</b>
4.1	Functional Dimension . . . . .	16
4.1.1	Topic Selection . . . . .	16
4.1.2	Literature Review . . . . .	18
4.1.3	Methodology Development . . . . .	19
4.1.4	Timeline Management . . . . .	20
4.1.5	Coding-related Tasks . . . . .	22
4.1.6	Writing-related Tasks . . . . .	27
4.1.7	Mock Defense . . . . .	28
4.2	Critical Thinking Dimension . . . . .	29
4.2.1	GenAI's Fulfillment of Critical Thinking Responsibilities . . . . .	29
4.2.2	Strengths and Limitations . . . . .	30
4.2.3	Optimal Prompting and Utilization Strategies . . . . .	31
4.3	Emotional Support Dimension . . . . .	31
4.3.1	GenAI's Fulfillment of Emotional Support Responsibilities . . . . .	31

4.3.2	AI-Assisted Content Analysis Methodology . . . . .	31
4.3.3	Content Analysis . . . . .	32
4.3.4	Qualitative Analysis . . . . .	34
<b>5</b>	<b>Discussion</b>	<b>36</b>
5.1	Answers to RQs . . . . .	36
5.1.1	To What Extent Can GenAI Fulfill the Key Responsibilities of a Thesis Supervisor in a Graduation Project? . . . . .	36
5.1.2	What Are the Strengths and Limitations of GenAI When Functioning as a Research Supervisor Compared to Human Supervisors? . . . . .	38
5.1.3	How Can GenAI Be Optimally Prompted and Utilized to Achieve the Best Supervisory Performance? . . . . .	38
5.2	Using AI or Being Supervised by AI? A Brief Discussion . . . . .	39
5.3	Innovations and Limitations . . . . .	40
<b>6</b>	<b>Conclusion</b>	<b>41</b>
	<b>Appendices</b>	<b>49</b>
A	Nail Biting Detection Case Study . . . . .	49
A.1	Introduction and Research Context . . . . .	49
A.2	Literature Review . . . . .	49
A.3	Methodology and Approaches . . . . .	50
A.4	Results and Limitations . . . . .	52
B	Four-Step LLM Prompting Methodology . . . . .	53
B.1	Generation Prompt . . . . .	54
B.2	Classification Prompt . . . . .	55
B.3	Aggregation Prompt . . . . .	57
B.4	Prevalence Prompt . . . . .	58

# Chapter 1

## Introduction

I have been biting my nails for as long as I can remember. When I began planning my graduation thesis, I thought: why not turn this habit into research? Perhaps I could develop a wearable device to detect nail biting or explore its psychological dimensions. In early discussions with my supervisor Bas Haring, I extensively used ChatGPT to brainstorm research ideas, and within just two weeks, I had generated ten different research topics about nail biting. Seeing how effectively I had used AI to generate these ideas, Bas wondered: if AI could so effectively support ideation and research development, could it potentially supervise an entire thesis? After further discussion, we decided to pursue this question as the primary research focus, with nail biting detection serving as a technical case study to evaluate AI's supervisory capabilities in technical supervision.


The rapid integration of Generative Artificial Intelligence (GenAI)<sup>1</sup> into academic life makes this question particularly timely. Most existing research on AI in education takes the form of surveys, interviews, or theoretical discussions about potential applications and risks. These studies typically observe how students and educators currently use AI or speculate about future possibilities, but rarely involve controlled experiments where AI actually assumes educational responsibilities. The empirical studies that do exist tend to evaluate AI as a supplementary tool—for writing assistance, code debugging, or information retrieval—rather than examining AI in a supervisory capacity. Research specifically on AI in thesis supervision is even scarcer, with only a handful of studies addressing this intersection, and none employing a comprehensive self-experiment where AI serves as the primary supervisor throughout an entire thesis project. **This methodological gap is significant because students are already informally using AI for guidance that resembles supervision, yet we lack empirical evidence about what happens when AI formally takes on this role.**

This thesis addresses this gap through a real-stakes self-experiment where GenAI

---

<sup>1</sup>GenAI refers to AI systems capable of producing content such as text, images, music, programming code and other complex and creative outputs, while LLMs (Large Language Models) focus specifically on understanding and generating human language (Dwivedi et al., 2025). These two terms are used interchangeably in this research.

supervised my entire graduation project with minimal human oversight. Using a nested experimental design, I evaluate AI supervision across two levels: the overall thesis examining AI's supervisory capabilities, and an embedded technical case study on nail biting detection that tests AI's ability to guide technical research. This structure allows assessment across three key supervisory dimensions drawn from established frameworks: functional support, critical thinking, and emotional support.



The thesis unfolds as follows. Chapter 2 reviews literature on research supervision, AI in education, and the limited work on AI-supervised research. Chapter 3 details the methodology, including research questions, research structure, the transition from human to AI supervision, choice of AI models, implementation of supervision tasks, and the evaluation framework. Chapter 4 presents the analysis across all supervisory dimensions, using both qualitative assessment and AI-assisted content analysis. Chapter 5 discusses the findings in relation to the research questions and explores the distinction between using AI as a tool versus being supervised by AI. Chapter 6 concludes with implications for the future of research supervision. The appendices provide technical details of the nail biting detection case study and the AI-assisted content analysis methodology.

# Chapter 2

## Literature Review

This section reviews previous studies on research supervision, AI in education, and GenAI in research supervision. Since this study is interdisciplinary, combining research supervision and AI, this review aims to provide a comprehensive understanding of these two distinct fields before examining their intersection.

The review begins with an overview of research supervision literature, particularly focusing on supervisory roles and responsibilities. It then examines AI applications in educational contexts, with particular attention to the emergence of GenAI in recent years. Finally, it explores the limited existing research on GenAI specifically applied to research supervision.

### 2.1 Research Supervision

Postgraduate research supervision has long been a significant topic in higher education (Bastalich, 2015), with most studies focusing on doctoral supervision. A systematic review (Haley, Holmqvist, and Johansson, 2024) identified 1,193 articles about research supervision between 1998 and 2019. The research topics include the conceptualization of supervision (Lee, 2007; Grant, Hackney, and Edgar, 2014), pedagogical aspects of supervision and supervision skills (Qureshi and Vazir, 2016; Guerin, Kerr, and Green, 2015), supervisory styles (Severinsson, 2012; Deuchar, 2008), and practical instructions for supervision (McCallin and Nayar, 2012; Abiddin, A. Hassan, and Ahmad, 2009).

#### 2.1.1 Supervisory Roles and Responsibilities

The branch of literature of particular concern here involves the role and responsibilities of the supervisor. The term "supervision" is defined by Merriam-Webster as "to be in charge of, to superintend, to oversee, especially involving critical watching and directing of activities

or courses of action" (Merriam-Webster, n.d.). By definition, one core aspect of supervision is overseeing and monitoring.

However, the nature of supervision is multidimensional. Grant, Hackney, and Edgar (2014) use three metaphors to describe the role of supervision: (a) the metaphor of the machine, which emphasizes that a supervisor monitors students' performance in line with university regulations, making supervision more of an "institutional act"; (b) the metaphor of the coach, where the supervisor is seen as a project advisor or coach who acts as a critical friend to provide guidance; and (c) the metaphor of the journey, where supervision resembles a partnership and research journey where student and supervisor learn together. In most cases, ideal supervision combines all three metaphors, as the authors note: "metaphor (b) and to a lesser extent, elements of metaphor (a) are needed in addition to metaphor (c) to ensure that both the student and the supervisor reach the final destination of a completed thesis safe and sound."

A more comprehensive framework (Lee, 2008) categorizes supervisory responsibilities into five distinct dimensions. According to this framework, a supervisor's conceptual understanding and approach significantly influences both their supervisory practices and the type of researcher that develops through the supervision process:

- **Functional:** Encompasses project management responsibilities, including scheduling regular meetings, monitoring progress, providing writing guidance, and handling operational aspects of research supervision.
- **Enculturation:** Involves facilitating students' integration into their academic discipline and professional community, with supervisors serving as gatekeepers to resources, expert networks, and disciplinary knowledge.
- **Critical thinking:** Focuses on developing students' analytical capabilities and encouraging systematic questioning and evaluation of their research work.
- **Emancipation:** Emphasizes the supervisor's role as mentor and coach, supporting student independence while providing guidance for academic and professional development.
- **Relationship development:** Recognizes the importance of positive supervisory relationships that motivate, encourage, and provide emotional support throughout the research process.



## 2.2 AI in Education

AI involves the development of computer algorithms to perform tasks typically associated with human intelligence (Goodfellow, Bengio, and Courville, 2016). It is viewed by many as a driver of the fourth industrial revolution and may trigger a corresponding revolution in education (Butler-Adam, 2018).

AI has been increasingly adopted in educational contexts. Recent survey data demonstrates this rapid adoption: among 1,041 full-time undergraduate students in the United Kingdom, 92% now use AI in some form, representing a significant increase from 66% in 2024 (Higher Education Policy Institute, 2025).

This growing adoption has sparked considerable academic interest, generating substantial research in AI applications for education. A comprehensive literature review (Zhai et al., 2021) examining AI in education research from 2010 to 2020 identified three primary areas of focus: (1) developing educational systems such as intelligent tutoring systems (ITS) (Horáková, Houška, and Dömeová, 2017; Yang, Kuo, and Liao, 2011), (2) designing AI algorithms that provide students with feedback, reasoning support, and adaptive learning experiences (Melo et al., 2014; Vattam, Goel, and Rugaber, 2011), and (3) exploring affective computing, role-playing, immersive learning, and gamification approaches (Lin et al., 2012; Ngai et al., 2010). The same review also identified emerging trends in educational AI, including the Internet of Things, swarm intelligence, deep learning, and neuroscience applications.

### 2.2.1 GenAI in Higher Education

Since ChatGPT-3.5's release in November 2022, Generative Artificial Intelligence (GenAI) has received significant attention and is transforming education broadly (Lo, 2023). Since 2022, besides ChatGPT, many large language models (LLMs) such as Gemini, Claude, and DeepSeek have emerged and continue improving at a rapid pace. Numerous studies about GenAI have appeared during these three years. However, most papers focus on broad trends, GenAI's potential applications in the education sector, or students' and educators' usage of GenAI, while fewer studies examine the actual implementation of GenAI in real educational environments.

A comprehensive literature review (Belkina et al., 2025) examined 21 empirical studies on GenAI integration in higher education teaching and learning. The review found that research concentrated primarily in Languages, Information and Communication Technologies, and Science and Engineering disciplines, with most studies employing

mixed-method approaches that combined quantitative and qualitative methodologies. These studies addressed multiple dimensions of learning and teaching processes. For instance, Wang and Feng (2024) demonstrated ChatGPT's application in reading and analyzing English literature, focusing on knowledge acquisition, while Pitso (2023) explored ChatGPT's role in creating practical outputs such as research reports and assignments, emphasizing knowledge production aspects of learning.

Additionally, GenAI plays different roles across these studies. Some research positions GenAI as a direct substitute for teachers or tutors (Kirwan, 2023), while other studies utilize GenAI as an assistant that enhances traditional educational tools through advanced functionalities such as grammar checking and writing assistance (Kuramitsu et al., 2023; Khang et al., 2023).

## 2.3 GenAI in Research Supervision

Research on GenAI specifically in research supervision remains limited, with few studies addressing this intersection. Among the existing work, Dai et al. (2023) investigated the practices and perspectives of 20 postgraduate research students in Australia who had at least four months' experience using ChatGPT in research activities. Their findings suggest that ChatGPT usage leads to a shift in supervisory roles and responsibilities: **supervisors increasingly provide strategic direction and high-level guidance, while students transition from apprentices to autonomous researchers as ChatGPT fosters greater independence in their research processes.**

Jensen et al. (2025) adopted a self-study approach to compare feedback from human supervisors and generative AI chatbots. Their findings indicate that supervisor feedback is formative, temporal, and relational in character, while GenAI feedback is task-focused, immediate, and agreeable. Additionally, chatbots tend to emphasize task completion rather than promoting deeper student learning. The authors conclude that chatbots underscore, rather than replace, the importance of human supervisors' engagement in feedback practices that foster meaningful learning.

Another study (Boyd and Harding, 2025) examined GenAI's impact on doctoral supervision from a structuration theory perspective, employing mixed methods to investigate how GenAI adoption shifts agency and power structures within supervisory relationships. The research shows that doctoral students often use GenAI tools without acknowledging or disclosing this usage to their supervisors. **The study concludes that transparent and integrated incorporation of GenAI into research supervision is essential.**

Several studies focus on technical innovations in specific aspects of research supervision.

For example, Xu, Ye, and Zhu (2023) designed a deep-learning based AI supervisor for idea generation and novelty assessment. Tailored for materials science, this AI supervisor can recommend research ideas, analyze their novelty, and provide comprehensive guidance to researchers.

A more technically ambitious approach (Ifargan et al., 2024) developed a data-to-paper automation platform that guides interacting LLM agents through a complete research process while maintaining programmatic traceability and human oversight. In autopilot mode, the platform can autonomously generate hypotheses, design research plans, write and debug analysis code, interpret results, and produce complete research papers from annotated data alone. While the research novelty was relatively limited, this work demonstrates the potential for AI-driven acceleration of scientific discovery.

## 2.4 Research Gaps

By reviewing these studies, several research gaps are identified. First, research on GenAI in thesis supervision remains limited, with only a handful of studies addressing this specific intersection. Second, while theoretical discussions about GenAI in education are abundant, empirical studies examining actual GenAI implementation in real educational environments are scarce. Third, although some studies have employed self-study methodologies, none have conducted a comprehensive self-experiment involving an entire graduate thesis process under AI supervision. These gaps underscore the need for systematic empirical research examining GenAI's potential as a research supervision tool—a need this study directly addresses.


# Chapter 3

## Methods

This chapter presents the research methodology, including the research questions, the nested experimental design, the transition from human to AI supervision, the selection and implementation of GenAI models, and the evaluation framework used to assess AI supervision effectiveness.

### 3.1 Research Questions

This study aims to address the following research questions:

- 
1. To what extent can GenAI fulfill the key responsibilities of a thesis supervisor in a graduation project?
  2. What are the strengths and limitations of GenAI when functioning as a research supervisor compared to human supervisors?
  3. How can GenAI be optimally prompted and utilized to achieve better supervisory performance?

### 3.2 Research Structure

This research employs a self-study methodology implementing a nested experimental design. GenAI supervises this entire graduate thesis, which itself investigates GenAI supervision capabilities, creating a meta-research dimension. The supervision covers the complete research process except for the initial brainstorming and topic selection phase conducted with my human supervisor Bas Haring.

To evaluate GenAI's technical supervision abilities, a nail biting detection case study serves

as the nested technical project under AI supervision. This case study utilizes deep learning methods to train a computer vision classification model for identifying bitten nails.

This nested structure allows for comprehensive evaluation of AI supervision capabilities across different research dimensions: the overarching thesis examines AI's ability to guide theoretical framework development, methodology design, academic writing, emotional support, and other supervisory tasks, while the embedded case study focuses specifically on technical guidance in training machine learning models. This approach provides both breadth and depth in assessing GenAI's supervisory effectiveness across various aspects of graduate-level research.

### 3.3 Transition from Human Supervision

Following several initial meetings with my human supervisor to establish the research objectives and methodology, my human supervisor ceased direct supervision of research content from April 13th, 2025, onward. From this point forward, the entire development of this thesis was supervised by GenAI until the thesis defense.

To ensure research quality and student safety, several safeguards were implemented. We established a shared Dropbox folder containing a detailed research logbook that I updated regularly to document progress, challenges, and key decisions, allowing my human supervisor to monitor research development and intervene if serious issues arose. My human supervisor also provided a budget to facilitate regular meetings with peers and friends, compensating for the reduced social interaction inherent in AI-supervised research. Additionally, we maintained communication regarding administrative matters through WhatsApp, ensuring that institutional requirements and deadlines were met while preserving the integrity of the AI supervision experiment.

This transition arrangement created a controlled experimental environment where GenAI could function as the primary supervisor while maintaining essential oversight for ethical and practical considerations. The human supervisor remained available as a safety net without directly influencing the research content or methodology implementation.

### 3.4 Choice of GenAI Models

Plenty of large language models (LLMs) are available in the market with various implementation approaches. Since this study aims to evaluate GenAI performance in

research supervision rather than assess LLM capabilities per se, the selection prioritized representative models rather than optimal model performance.

ChatGPT and Claude were selected for their widespread adoption and complementary strengths: ChatGPT represents the most widely used conversational AI model, while Claude demonstrates particular proficiency in coding tasks. Both rank among the top-performing models in current evaluations. According to the Vellum LLM Leaderboard<sup>1</sup>, GPT-4o ranks third in Best in Tool Use, while Claude Sonnet 4 ranks fourth in agentic coding capabilities and second in adaptive reasoning. The models used in this research were GPT-4o, Claude Sonnet 4 and Claude Sonnet 3.5 (20241022).

The original methodology planned parallel usage of both models across all supervision tasks. However, as the research progressed, I found myself increasingly favoring Claude, a shift I did not deliberately counteract since LLM usage preferences mirror the natural individual preferences that occur in human supervisory relationships as well.

Therefore, most tasks mainly used Claude, while ChatGPT was utilized primarily for mock defense preparation, as it uniquely offered live video chat functionality, and for emotional support, as GPT-4o has the best performance in emotion detection compared to Llama, Gemma, and Mistral series (Lecourt, Croitoru, and Todorov, 2025). Claude Sonnet 4 is mainly used with Claude Sonnet 3.5 (20241022) only used in the evaluation part of GenAI supervision capabilities

### 3.5 Implementation of Supervision Tasks

This study adopts the supervisory framework proposed by Lee (2008), as discussed in the literature review, which categorizes supervisory responsibilities into five dimensions: Functional, Enculturation, Critical thinking, Emancipation, and Relationship development.

To facilitate practical implementation and evaluation in this research context, the functional dimension was further subdivided into seven specific tasks: (1) topic selection, (2) literature review, (3) methodology development, (4) writing-related tasks, (5) coding-related tasks, (6) timeline management, and (7) mock defense.

Given the fundamental differences between human-AI and human-human relationships, and the difficulty in measuring personal growth and emancipation in an AI supervision context, the emancipation and relationship development dimensions were consolidated

---

<sup>1</sup>Vellum LLM Leaderboard, <https://www.vellum.ai/llm-leaderboard>, accessed August 13, 2025. Note that following the release of GPT-5 (August 7, 2025) and GPT-OSS-120B and GPT-OSS-20B models (August 5, 2025), the ranking of LLMs has changed significantly. Before the launch of these new OpenAI models, the models used in this research ranked higher in the leaderboard

into a single category: emotional support. The enculturation dimension was not evaluated in this thesis due to GenAI's **fundamental inability to integrate students into the academic community.**

All supervision tasks were implemented through ChatGPT and Claude desktop applications using their standard prompting interfaces. Both platforms' project management functions were utilized to organize all activities under a "graduation thesis" project. Each supervision task was assigned a separate chat session to maintain clear boundaries between different supervisory functions and facilitate systematic evaluation of AI performance across distinct areas of responsibility.

### 3.6 Research Evaluation

The evaluation framework is structured around the task-specific categorization listed above. Most tasks employ qualitative evaluation conducted by myself, centered around three research questions. The two research questions are given more coverage and the evaluation around the third question, which is how to prompt AI to have a better performance is not evaluated in detail in the evaluation of every task since the focus of this research is not about prompting engineering and the third question will be generally answered in the Discussion section.

For analyzing AI supervision in coding-related tasks and emotional support, I used AI-assisted content analysis through LLMs (GPT-4o and Claude Sonnet 3.5), as these two domains generated considerable amounts of textual data requiring systematic analysis.

**For the two tasks requiring extensive textual analysis**—coding-related tasks and emotional support—traditional qualitative and statistical approaches such as manual thematic analysis<sup>2</sup> and topic modeling (e.g., Latent Dirichlet Allocation)<sup>3</sup> were not employed in this study due to several limitations that make them less suitable for this research context. Manual thematic analysis, while providing deep interpretive insights, suffers from significant scalability constraints, requiring extensive human coding effort. Topic modeling techniques like LDA often struggle with small numbers of documents (Tang et al., 2014) and may fail to capture the contextual nuances critical for understanding AI supervision behaviors.

AI-assisted content analysis was adopted because LLMs excel in scalability and their ability to handle complex, contextual text interpretation. This approach has gained empirical

---

<sup>2</sup>Thematic analysis is a method for identifying, analysing and reporting patterns (themes) within data. It minimally organizes and describes your data set in detail (Braun and Clarke, 2006).

<sup>3</sup>LDA is a generative probabilistic model for text corpora, where each document is modeled as a mixture of topics, and each topic is modeled as a distribution over words (Blei, Ng, and Jordan, 2003).

validation in recent research. Tai et al. (2024) demonstrated ChatGPT-3.5's effectiveness in conducting deductive coding<sup>4</sup>, showing that LLM-based text analysis produces consistent and reliable results comparable to human coders. Gilardi, Alizadeh, and Kubli (2023) demonstrated that ChatGPT outperforms crowd-workers across several NLP annotation tasks, including relevance, stance, topics, and frame detection. Notably, ChatGPT's inter-coder agreement exceeds that of both crowd-workers and trained annotators for all tasks, achieving inter-rater reliability over 90%. Rao et al. (2025) introduced an LLM-based framework for extracting and analyzing textual data, developing a four-step prompting methodology to identify the most prevalent themes and quantify their distribution within textual datasets. This research adopts and adapts this proven framework to analyze AI supervision in technical and emotional support; detailed implementation will be specified in the analysis section.

---

<sup>4</sup>Deductive coding is a qualitative research method where researchers apply predetermined theoretical frameworks or coding schemes to systematically categorize textual data.



# Chapter 4

## Analysis

This chapter provides a comprehensive analysis of AI supervision capabilities across different dimensions of graduate research supervision. The analysis is structured around Lee's (Lee, 2008) supervisory framework, examining AI performance in functional supervision, critical thinking, and emotional support. For functional supervision, the chapter evaluates AI effectiveness across multiple tasks including topic selection, literature review, methodology development, timeline management, coding-related tasks, writing-related tasks, and mock defense preparation. The analysis employs both qualitative self-assessment and AI-assisted content analysis to systematically evaluate AI supervision performance.

### 4.1 Functional Dimension

#### 4.1.1 Topic Selection

##### **GenAI's Fulfillment of Topic Selection Responsibilities**

GenAI demonstrated substantial capability in fulfilling topic selection responsibilities throughout the brainstorming process. My engagement with GenAI began early, as I had been extensively discussing potential graduation topics with it as a regular user. Initially, I proposed using wearables to detect nail biting while simultaneously monitoring participants' biometrics to investigate the relationship between emotions and nail biting behavior. While my human supervisor expressed keen interest in the nail biting component and suggested focusing on designing a detection device, I discovered that similar devices had already been developed by researchers at MIT Media Lab (Azaria, Mayton, and Paradiso, 2016). This setback led my supervisor to grant me an additional two weeks for brainstorming alternative approaches around nail biting. During this period, GenAI helped generate ten new research ideas about nail biting, which I subsequently presented

to my human supervisor. Recognizing the AI's contribution to this ideation process, my supervisor saw the potential of AI in research supervision itself. After our discussion, we decided to focus on GenAI in research supervision as the primary research subject, with nail biting detection serving as a supporting case study. Thus, AI's instrumental role in the topic selection process ultimately led to its emergence as the primary research subject itself.

## Strengths and Limitations

GenAI's primary strengths in topic selection supervision include its capability in providing entry-level domain knowledge. In my case, being completely new to computer vision, ChatGPT quickly informed me about specific characteristics of bitten nails—such as rugged distal borders and shortened nail beds—and explained how machine learning models could be trained to identify these features. It suggested both semantic segmentation and end-to-end classification approaches; I ultimately chose the latter, implementing pretrained CNN models for nail biting classification.

GenAI also demonstrated notable creative capabilities in idea generation. Among the various concepts it proposed, several proved particularly compelling: "Art of Nail Biting: 3D Modeling Shapes of Nail Biters," "Nail Biting: An Evolutionary Perspective," and "Social Contagion of Nail Biting." The "social contagion" concept originated entirely from ChatGPT and was seriously considered by both my human supervisor and myself as a viable thesis direction, suggesting that GenAI can generate novel and valuable research ideas.

For rapid prototyping and technical validation, GenAI proved valuable. With ChatGPT's assistance, I developed a proof-of-concept by web-crawling nail images, manually labeling distal borders, and implementing a basic UNet model for border prediction. This prototype successfully demonstrated my capability to execute the proposed project, serving as crucial evidence for topic selection decisions.

However, significant limitations emerged in GenAI's supervisory capacity. While examining our conversations, I noticed that the core idea of detecting nail biting through nail images actually originated from my own thinking rather than AI suggestion. Although GenAI excels at providing entry-level domain knowledge, it cannot match the domain expertise of an experienced professor in identifying truly significant research opportunities or understanding subtle disciplinary nuances.

## Optimal Prompting and Utilization Strategies

The effectiveness of GenAI in topic selection supervision proved highly dependent on prompting strategies and contextual information provided. During the intensive brainstorming period, I engaged with ChatGPT to explore various aspects of potential topics, including their novelty, technical difficulty, and alignment with my interests. The most productive interactions occurred when I provided sufficient context about my background, interests, and constraints. **The key was not asking for direct solutions but rather engaging in exploratory discussions about various possibilities and constantly prompting AI to suggest connections and alternatives I hadn't considered.**

### 4.1.2 Literature Review

#### GenAI's Fulfillment of Literature Review Responsibilities

GenAI demonstrated considerable capability in literature review supervision, particularly in targeted literature sourcing. GenAI was used in two ways. First, I prompted AI to generate complete literature reviews to gain insights into review structures. Second, I utilized AI for targeted literature sourcing to identify relevant studies within my research domain. While AI's capacity for generating comprehensive literature reviews proved limited, it significantly enhanced my efficiency in discovering pertinent research.

To evaluate GenAI's capability in generating literature review, I prompted Claude to generate a complete literature review for my research, providing comprehensive information about my research theme and objectives.<sup>1</sup> Claude reported gathering 394 sources and produced a 2,700-word literature review with a well-structured framework. The generated review has some strengths like systematic paper identification, thematic organization, gap analysis, and academic writing. However it also has several problems: referencing non-seminal or secondary sources rather than original foundational works, lacking proper citations for factual claims, containing minor factual errors, and repeatedly citing certain papers inappropriately. For example, Claude referenced Lasabuda's personal website interpretation<sup>2</sup> seven times despite it being non-peer-reviewed work rather than seminal research.

More significantly, AI-generated reviews contained substantially more arguments and judgments with fewer direct study citations compared to human-written reviews, making

<sup>1</sup>The complete Claude-generated literature review is available at: <https://claude.ai/public/artifacts/fc7ee764-b09f-43dc-921f-d7664f14b348>

<sup>2</sup>Lasabuda, N. (n.d.). PhD Supervision Philosophy. <https://amandolasabuda.com/research-philosophy/>

validity assessment difficult. The reviews also showed potential plagiarism risks, with considerable content appearing to derive directly from original papers without proper attribution.

When using AI for literature sourcing, it proved highly effective. In searching for studies specifically about GenAI use in research supervision—a relatively new research area with limited available literature—GenAI successfully identified important and recent studies that might not have been readily discoverable through traditional search methods. Moreover, AI consistently provided accurate bibliographic information, including correct paper titles, abstracts, author names, and publication years, with low instances of hallucination.

## Strengths and Limitations

AI's primary strengths in literature review supervision lie in its ability to search extensive resources rapidly, process large volumes of academic literature, and generate comprehensive written reviews for reference purposes. However, compared to human supervisors, GenAI exhibits significant limitations in identifying truly seminal studies, understanding subtle disciplinary nuances, and providing guidance grounded in established academic writing conventions. Additionally, the literature review writing process serves as a crucial learning mechanism for researchers to develop deep understanding of their field—a developmental benefit that may be compromised if students increasingly rely on AI-generated content rather than engaging in the analytical synthesis process themselves.

### 4.1.3 Methodology Development

#### GenAI's Fulfillment of Methodological Supervision Responsibilities

GenAI demonstrated limited capability in fulfilling methodological supervision responsibilities. Creating a theoretical framework for this research presented significant challenges due to the rapid development of AI tools, the learning-by-doing nature of the process, and the innovative character of this methodology. The framework needed to address how to define and divide supervision responsibilities, implement them practically, and evaluate AI supervision effectively. Although my human supervisor and I had established a general framework before AI supervision began, substantial refinement work remained.

## Strengths and Limitations

GenAI's strengths are primarily through literature sourcing for methodological inspiration and offering practical implementation suggestions. Specifically, after I provided my initial version of functional task divisions, Claude suggested that "knowledge questioning" belonged under the "critical thinking" dimension and identified that literature review guidance was missing from my functional tasks, leading to the final classification.

Methodology development typically involves two distinct layers: the theoretical and structural level and the practical implementation level. GenAI is generally incapable of supervising at the theoretical and structural level—for instance, it cannot propose comprehensive theoretical frameworks for dividing responsibilities

This represents a clear division of capabilities compared to human supervisors. Human supervisors excel in theoretical aspects, possessing strong judgment about methodologically valid approaches and understanding how methodology should align with research questions. However, they may not always be available or possess expertise in specific technical implementation details. GenAI fills this gap by providing accessible, detailed guidance for technical implementation, while lacking the theoretical depth and methodological expertise that experienced human supervisors offer.

### 4.1.4 Timeline Management

#### GenAI's Fulfillment of Timeline Management Responsibilities

Timeline management typically represents a significant component of research supervision, including both administrative oversight and academic milestone tracking, though the extent of supervisors' enforcement versus student autonomy varies considerably across different supervision styles.

GenAI demonstrated limited capability in fulfilling timeline management responsibilities. It can create structured timelines, establishing reporting schedules, and providing organizational frameworks. For example, I adopted the approach of creating a separate chat session in Claude to establish timeline management. In May, we developed a preliminary timeline outlining key milestones leading to my thesis defense scheduled for August 20th. Claude and I agreed that I would report progress every Tuesday, creating a structured accountability system similar to traditional supervisory meetings.

However, the effectiveness of AI supervision in timeline management proved limited in practice. I failed to maintain the weekly reporting schedule after the initial few weeks,

revealing fundamental challenges in AI's ability to provide sustained timeline supervision without external support mechanisms.

## Strengths and Limitations

GenAI's strengths in timeline management include its availability for immediate consultation, ability to create detailed organizational frameworks, and flexibility in adjusting timelines as needed.

The most significant limitation of AI-based timeline management is the absence of structural pressure that typically characterizes human supervisory relationships. In traditional supervision, the inherent hierarchy between supervisor and student creates external accountability pressure, motivating students to meet discussed deadlines and prepare progress reports. With AI supervision, this structural pressure is absent, making timeline adherence more challenging and reliant entirely on internal motivation.

Interestingly, I found that sharing a thesis logbook with my human supervisor Bas provided more effective timeline management support than AI supervision alone. Although updated only approximately every two weeks, this logbook—which allowed Bas to monitor my progress and intervene if serious issues arose—created sufficient external accountability. While we didn't discuss progress details to preserve the integrity of the AI supervision experiment, simply knowing that he was reading my updates generated adequate external pressure to maintain research momentum. This demonstrates that even minimal human oversight can provide crucial accountability mechanisms that pure AI supervision fundamentally lacks.

## Optimal Prompting and Utilization Strategies

More sophisticated approaches could potentially enhance AI timeline management effectiveness. Integration with project management tools like Asana, or utilizing ChatGPT's notification capabilities for deadline reminders, might provide additional accountability mechanisms. However, regardless of technological integration, AI-based timeline management fundamentally requires significantly higher levels of student agency and self-discipline compared to traditional human supervision.

The optimal strategy appears to be using GenAI for initial framework creation while maintaining some form of human oversight—even minimal—for accountability. A hybrid approach where AI handles organizational tasks while human supervisors provide accountability pressure may be most effective.

### 4.1.5 Coding-related Tasks

The coding-related tasks were evaluated through the nail biting detection case study, which required extensive programming and debugging work. This computer vision project utilized transfer learning and deep learning methodologies to classify nail biting behavior from hand images. The study employed a dataset of 367 images collected from internet sources and the "11k Hands" dataset (Afifi, 2019), implementing two distinct approaches: a two-stage pipeline combining semantic segmentation with morphological analysis, and an end-to-end classification approach. While the segmentation approach showed poor generalization performance, the end-to-end classification model achieved its best observed run at 89.7% accuracy, 93.3% precision, 84.8% recall, and 88.9% F1-score after hyperparameter optimization. These results demonstrate that the approach can reach reasonably strong performance, but they should be interpreted as indicative rather than definitive given the small dataset size and the methodological limitations discussed later. The detailed technical implementation, model architectures, hyperparameter tuning, and performance analysis are provided in Appendix A.

### GenAI's Coding Capability

It is necessary here to provide a broad overview of LLMs' coding capabilities and how GenAI is used in coding generally. Li, Zhang, and Hassan (2025) outline the evolution of AI in software engineering through distinct paradigms: SE 1.0 involved traditional manual coding with explicit rules, while SE 1.5 introduced basic automated assistance like autocomplete features. The SE 2.0 paradigm uses large language models to generate complete code segments from natural language descriptions, which is already widely adopted<sup>3</sup>. The emerging SE 3.0 paradigm envisions autonomous AI agents that can independently perform complex development tasks including code analysis, testing, and deployment, which is also unfolding<sup>4</sup>

Quantitative benchmarks provide insight into the current level of LLMs' coding capabilities. SWE-bench (Jimenez et al., 2024) is an evaluation framework designed to evaluate LLMs' coding ability in real-world scenarios. When the benchmark was published in October 2023, the best-performing model, Claude 2, was able to solve only 1.96% of the issues. However, according to current SWE-bench Leaderboards<sup>5</sup>, the best-performing model Claude 4

---

<sup>3</sup>According to Stack Overflow's 2025 Developer Survey, 84% of respondents are using or planning to use AI tools in their development process, an increase from 76% the previous year (Stack Overflow, 2025).

<sup>4</sup>Evidence of this transition includes the expansion of AI-powered development tools, with 15 million developers using GitHub Copilot and the introduction of more autonomous features like agent mode and automated code review (Microsoft, 2025).

<sup>5</sup>SWE-bench Leaderboard. <https://www.swebench.com/> (accessed August 2, 2025).

Sonnet (20250514) achieves a 64.93% resolution rate on SWE-bench Bash-only<sup>6</sup> Verified<sup>7</sup> with minimal agentic support, demonstrating substantial improvement and considerable coding capability within a relatively short timeframe.

## GenAI's Fulfillment of Responsibilities in Technical Supervision

The case study provided an opportunity to examine how effectively GenAI can supervise the technical aspects of a research thesis<sup>8</sup>. In general, GenAI demonstrated strong performance in technical supervision during the development of my nail biting detection model. For a student without prior experience in deep learning or computer vision, being able to build an end-to-end model from scratch and obtain a reasonably good performance (despite only indicative rather than fully robust) already illustrates the practical value of AI supervision.

At the same time, the achieved performance was not exceptional, and the entire process contained several methodological limitations: the dataset was small and self-labeled, fine-tuning protocols were not systematically replicated, and hyperparameter searches were relatively narrow. These shortcomings highlight that, while GenAI support enabled me to reach a working solution, the process was far from rigorous. This in turn shows the continued importance of guidance from human domain experts, who would insist on stricter methodological standards and ensure the reliability and generalizability of results.

Moreover, since programming represents a technical domain where GenAI currently excels, it may make GenAI appear particularly capable in supervising technical research aspects. However, in research fields requiring more laboratory experiments or hardware work, GenAI at its current development stage remains limited and less competitive compared to human expertise.

## AI-Assisted Content Analysis Methodology

**Four-step LLM prompting methodology** For the extensive coding-related conversations generated by interacting with GenAI, I employed AI-assisted content analysis using a systematic four-step LLM prompting methodology adapted from Rao et al. (2025). This method extracted six chat sessions from interacting with Claude Sonnet 4 and split them

---

<sup>6</sup>SWE-bench Bash-only evaluates LLMs in a minimal bash environment with no tools or special scaffold structure, using only a simple ReAct agent loop to represent state-of-the-art LLM performance when given just a bash shell and a problem.

<sup>7</sup>Verified is a human-filtered subset of 500 instances from the original SWE-bench benchmark.

<sup>8</sup>While AI tools were also extensively used in other coding aspects of this research (such as help develop content analysis scripts and API integrations), this evaluation focuses specifically on AI supervision on the case study to avoid overly meta-analytical complexity.



into 262 conversation pairs. The four-step process included: (1) generation prompt to identify and extract AI support types from conversational data, analogous to open coding in qualitative research; (2) classification prompt to categorize behaviors into two primary categories Problem-Solving Support (direct technical assistance) and Skill-Enhancement Support (knowledge building and explanation); (3) aggregation prompt to identify the most representative subcategories within each primary category; and (4) prevalence prompt to quantify the distribution of support behaviors across all conversation pairs. The first three steps established the analytical framework, while the prevalence prompt provided the insights into how AI conducts technical supervision in practice. To ensure reliability, the entire process was replicated using both Claude Sonnet 3.5 and ChatGPT-4o as independent evaluators.

**Reliability and accuracy validation** The reliability analysis demonstrated substantial agreement between the two AI models in classification tasks. The models achieved agreement<sup>9</sup> on 226 out of 262 conversation pairs, resulting in 86.3% agreement and a Cohen's Kappa coefficient<sup>10</sup> of 0.627, indicating substantial inter-rater reliability. To further validate the classification accuracy<sup>11</sup>, 50 conversation pairs were randomly selected and manually annotated as gold standard data. Both models achieved strong performance, with Claude attaining 82% accuracy and ChatGPT achieving 90% accuracy, yielding an average accuracy of 86%. These results confirm the validity and reliability of using LLMs for systematic content analysis in this research context.

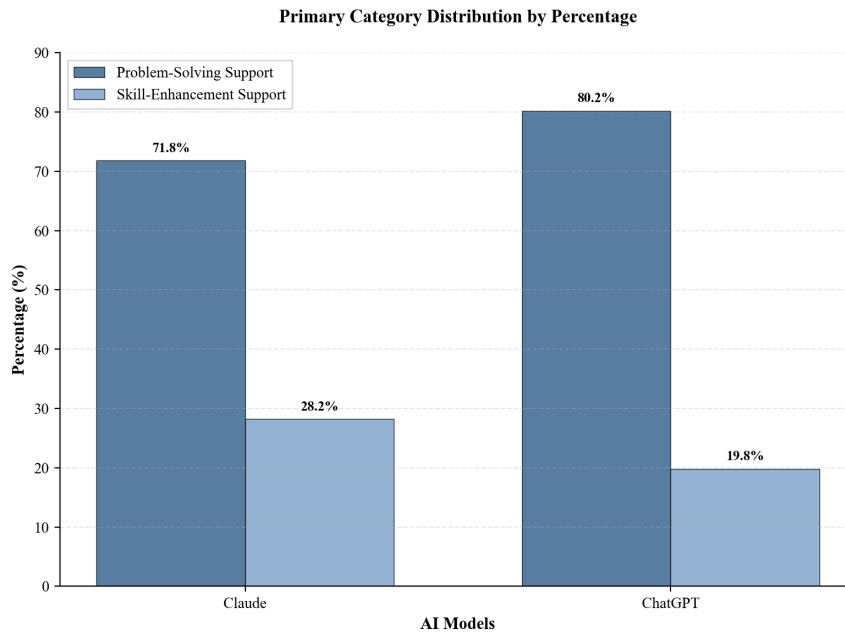
## Content Analysis

**Primary category distribution** The analysis revealed that Problem-Solving Support significantly dominated AI assistance in coding-related tasks. After post-processing correction for consistency, Claude classified 71.8% of conversation pairs (188 pairs) as Problem-Solving Support and 28.2% (74 pairs) as Skill-Enhancement Support, while ChatGPT showed an even stronger bias toward problem-solving with 80.2% (210 pairs) versus 19.8% (52 pairs) for skill-enhancement (Figure 1). **This distribution suggests that without specific fine-tuning, GenAI naturally gravitates toward providing direct solutions and technical fixes rather than focusing on educational guidance and skill development in technical supervision contexts.**

<sup>9</sup>Percent agreement =  $\frac{\text{Number of agreements}}{\text{Total number of scores}}$ . While simpler than Cohen's Kappa, it does not account for chance agreement (McHugh, 2012).

<sup>10</sup>Cohen's Kappa =  $\frac{P_o - P_e}{1 - P_e}$ , where  $P_o$  is observed agreement and  $P_e$  is expected agreement by chance. Interpretation: 0.21-0.40 (fair), 0.41-0.60 (moderate), 0.61-0.80 (substantial), 0.81-1.00 (almost perfect agreement) (Landis and Koch, 1977).

<sup>11</sup>Classification accuracy measures the proportion of correct predictions, calculated as Accuracy =  $\frac{TP + TN}{TP + TN + FP + FN}$ , where TP, TN, FP, FN represent true positives, true negatives, false positives, and false negatives respectively (Sokolova and Lapalme, 2009).



**Figure 1:** Primary category distribution of AI support types across 262 conversation pairs

**Subcategory distribution** The subcategory distribution shows a clear hierarchical pattern in AI coding support behaviors across 262 conversation pairs (Figure 2). The three most prevalent support types—Code Provision (combined: 32.5%)<sup>12</sup>, Concept/Visualization Explanations (combined: 32.1%), and Debugging Assistance (combined: 30.5%). This predominance suggests that AI tools excel primarily in providing direct problem resolution and conceptual clarification.

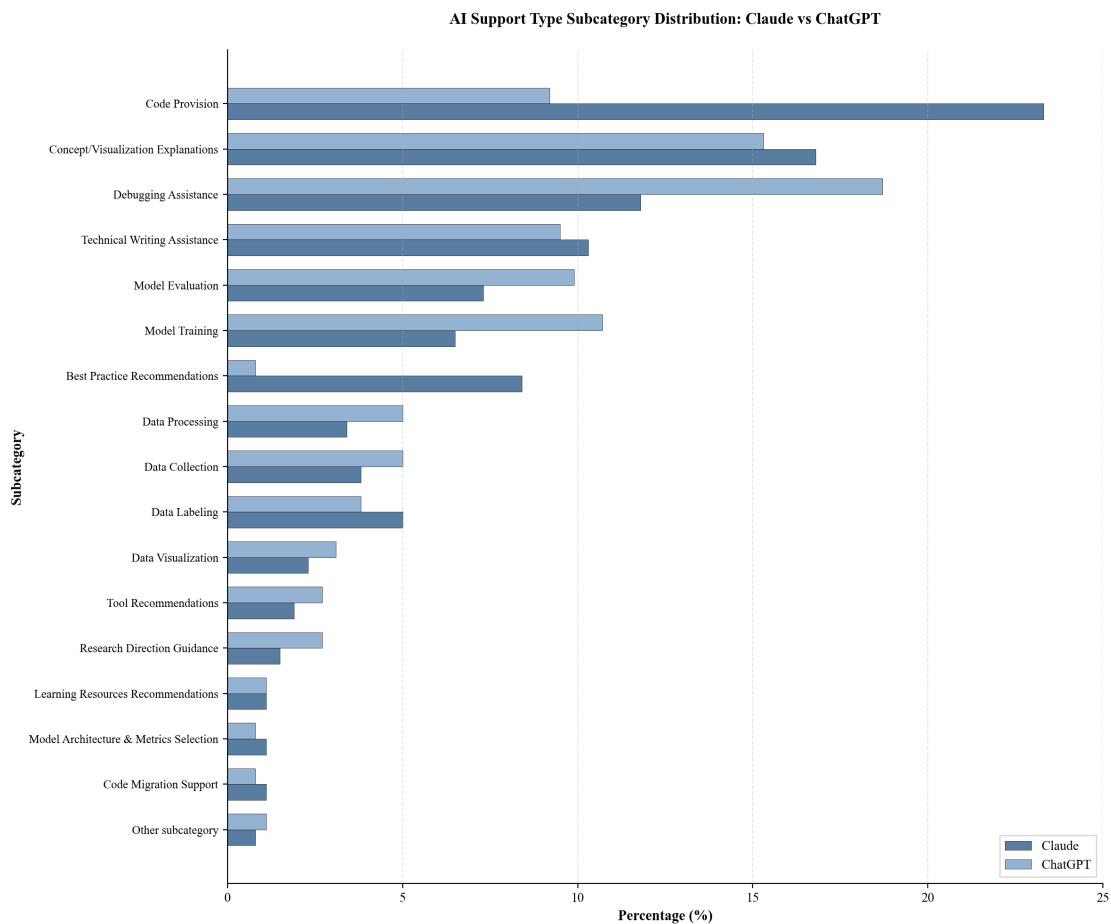
The second tier of prevalent support types includes Technical Writing Assistance, Model Evaluation, Model Training, Data Processing, Data Collection, and Data Labeling, which collectively demonstrate AI’s comprehensive support throughout the entire machine learning pipeline of my nail biting detection model. These categories reflect AI’s strength in providing step-by-step technical guidance across all phases of model development, from data preparation through evaluation.

However, the analysis reveals significant gaps in higher-level strategic supervision. The least prevalent categories include Research Direction Guidance (1.5-2.7%), Model Architecture and Metrics Selection (0.8-1.1%), and Code Migration Support (0.8-1.1%). This scarcity highlights AI’s limited capacity for strategic decision-making traditionally associated with academic mentorship.

Notably, the minimal presence of Learning Resources Recommendations suggests that AI supervision does not prioritize users’ independent learning and skill cultivation.

<sup>12</sup>Combined percentages represent the sum of both Claude and ChatGPT classifications for each subcategory (e.g., Code Provision: Claude 23.3% + ChatGPT 9.2% = 32.5%).

Similarly, Best Practice Recommendations shows considerable disagreement between models (Claude: 8.4% vs ChatGPT: 0.8%), but remains relatively low overall, indicating that current AI supervision focuses on immediate problem-solving rather than fostering long-term professional development and best practice adoption.



**Figure 2:** Comparison of AI support behavior subcategory distributions between Claude and ChatGPT across 262 conversation pairs

### Optimal Prompting and Utilization Strategies

Effective AI technical supervision requires students to develop sophisticated evaluation skills and maintain critical oversight of AI-generated solutions. Rather than passively accepting AI recommendations, students and researchers must possess sufficient foundational knowledge to assess the appropriateness and quality of AI assistance in technical contexts.

A critical aspect of successful AI technical supervision involves maintaining a balance between leveraging AI efficiency and preserving genuine learning. While AI excels at providing immediate technical assistance and explanations, over-reliance without actively

engaging with underlying concepts can undermine the educational goals of supervision. Users must consciously pursue understanding of technical principles and methodologies, treating AI as a supervisory tool that supplements rather than replaces deep conceptual learning.

This approach transforms technical supervision from traditional mentor-student dynamics toward a collaborative framework where AI provides immediate technical support while users maintain responsibility for strategic learning and skill development. **The effectiveness of AI technical supervision ultimately depends on users' ability to extract educational value from AI interactions rather than merely consuming solutions.**

#### 4.1.6 Writing-related Tasks

##### GenAI's Fulfillment of Writing Supervision Responsibilities


In general, GenAI demonstrated considerable capability in fulfilling writing supervision tasks throughout my thesis development. AI provided comprehensive support across all writing dimensions, including recommending suitable Overleaf templates for academic formatting, helping polishing and proofreading paragraphs with appropriate academic tone, creating professional tables and figures based on provided data, managing citations and reference formatting.

##### Strengths and Limitations

GenAI demonstrates several advantages over human supervisors in writing supervision. Its 24/7 availability and instant response capabilities provide accessibility, allowing students to receive immediate feedback **without scheduling constraints or waiting periods that characterize traditional supervision.**

Additionally, GenAI excels in language support, offering superior grammar correction, multi-language assistance for non-native speakers, and expertise in formatting requirements such as LaTeX, citation styles, and complex document structures.

However, AI-supervised writing introduces significant limitations requiring careful navigation. The boundary between writing assistance and plagiarism often remains unclear, requiring cautious adoption of AI-generated text into academic work. Unlike human supervisors who can provide nuanced guidance on disciplinary writing conventions and personal style development, GenAI lacks the contextual understanding of field-specific requirements and the ability to mentor long-term writing skill development.



Human supervisors excel at evaluating structural coherence, logical flow, and the overall narrative of research arguments—capabilities that require understanding of disciplinary conventions and **research storytelling** that current GenAI cannot fully replicate.

While GenAI offers remarkable efficiency and technical precision in writing support, it fundamentally operates as a sophisticated editing tool rather than a developmental mentor, excelling in formatting tasks while lacking the intellectual depth and personal guidance.

### **Optimal Prompting and Utilization Strategies**

Successful integration of GenAI in writing supervision demands several key competencies from users. Students should possess sufficient understanding of academic conventions to evaluate and adapt AI suggestions appropriately. Critical judgment remains essential for determining when AI assistance enhances versus compromises academic integrity. Additionally, writers need foundational knowledge of their field's writing norms to guide AI interactions effectively and ensure disciplinary appropriateness.

#### **4.1.7 Mock Defense**

##### **GenAI's Fulfillment of Mock Defense Responsibilities**

GenAI demonstrated moderate effectiveness in supervising mock defense preparation. I utilized ChatGPT's voice chat function in the OpenAI app to conduct mock defense sessions, providing comprehensive information about my thesis and instructing the AI not to interrupt until I completed my full presentation before offering feedback and questions.

The AI provided valuable suggestions regarding language use and presentation style. As a non-native English speaker, I used some phrases repeatedly so ChatGPT offered helpful recommendations for more authentic expressions and grammatical corrections. The system also raised questions about my methodology, such as "How would you address potential bias in your self-evaluation of AI supervision effectiveness?".

However, the mock defense experience felt somewhat artificial, as the conversational dynamic with AI lacks the spontaneous and unpredictable nature of human examination interactions. The AI's responses, while technically sound, followed predictable patterns that differ from the varied questioning styles of actual thesis committee members.

For presentation preparation, I initially attempted to use Microsoft Copilot for slide generation, but found its direct output of limited utility. Consequently, I turned to Claude

for structural guidance and engaged in extensive discussions about slide content. Claude provided substantial assistance in organizing presentation materials, and nearly all figures displayed in my defense slides were generated using Python scripts that Claude provided, demonstrating its strength in technical content creation and data visualization support.

## Strengths and Limitations

GenAI's primary strengths in mock defense supervision include its accessibility and availability for practice sessions at any time. Students can repeat mock sessions unlimited times, practicing specific sections or the entire presentation until achieving desired fluency and confidence levels. AI also provides a non-judgmental environment that reduces anxiety and allows for open exploration of presentation weaknesses without fear of criticism.

However, significant limitations influence AI's effectiveness in mock defense supervision. First, presentations involve extensive nonverbal communication including body language, eye contact, posture, and hand gestures, which AI cannot observe or provide guidance on during voice-only interactions.

Second, AI lacks the ability to simulate the authentic pressure and social dynamics of a real thesis defense. Human examiners bring years of academic experience, personal questioning styles, and the ability to create genuine intellectual pressure that tests students' knowledge under stress. AI interactions, while helpful for content preparation, cannot replicate the intensity and unpredictability of facing a panel of experts who may challenge fundamental assumptions.

## 4.2 Critical Thinking Dimension

One classic definition characterizes critical thinking as "reasonable reflective thinking that is focused on deciding what to believe or do" (Ennis, 1987). Within Lee (2007) supervisory framework, encouraging students to question and analyze their work systematically represents a fundamental dimension of effective supervision.

### 4.2.1 GenAI's Fulfillment of Critical Thinking Responsibilities

GenAI essentially lacks the capacity to challenge and question students' fundamental beliefs and assumptions. The AI systems rarely disagreed with my research proposals,

ideas, or directions, despite their demonstrated capability to correct basic factual errors, for example regarding metrics or technical concepts.

One example that especially demonstrates this supervisory deficiency is evident in the implementation of my nail biting detection case study. As previously introduced, two approaches were initially considered: the image segmentation approach and the end-to-end classification approach. GenAI provided supportive responses for both approaches and quickly offered technical implementation solutions without thorough feasibility assessments. For someone new to deep learning and computer vision, it was difficult for me to judge which approach would be optimal given time and technical constraints. Consequently, I experimented with both approaches and eventually discovered that the image segmentation approach was too time-consuming so it was deprecated—a conclusion that better initial evaluation might have revealed earlier.

Furthermore, GenAI fundamentally lacks the ability to identify potential methodological problems before they manifest. Again, within my nail biting case study, when experimenting with the image segmentation approach, initial experiments showed promising performance metrics. GenAI immediately provided very positive evaluations of the model performance based on these early results. However, when I tested the model with external test data—a validation step I came up with independently without GenAI's recommendation—it exhibited serious generalization problems that the AI had failed to anticipate or warn against.

## 4.2.2 Strengths and Limitations

This lack of intellectual challenge fundamentally represents a supervisory deficiency, though in rare cases, students who are sufficiently critical and domain-knowledgeable may find GenAI's consistent validation of their assumptions encouraging for maintaining research confidence.

However, this absence of intellectual challenge constitutes one of the most significant disadvantages of GenAI-based supervision, potentially leading to serious consequences including students pursuing misguided research directions and investing excessive time in unproductive endeavors. Perhaps more critically, without adequate intellectual friction and questioning, students miss crucial opportunities to develop the critical thinking capabilities and analytical skepticism that are fundamental qualities for competent researchers. The supervisory relationship should inherently involve constructive disagreement and challenging assumptions—functions that current GenAI systems are fundamentally unable to provide.

### 4.2.3 Optimal Prompting and Utilization Strategies

While consciously prompting GenAI to provide balanced perspectives on proposed ideas might mitigate this limitation to some extent, such approaches prove difficult to maintain consistently. It may also contradict natural human tendencies since it's reported that human preference judgments themselves tend to favor sycophantic responses (Sharma et al., 2025).

## 4.3 Emotional Support Dimension

### 4.3.1 GenAI's Fulfillment of Emotional Support Responsibilities

GenAI has demonstrated considerable capacity to provide emotional support through the thesis project. As a regular ChatGPT user who uses the platform as a responsive journaling tool for emotional expression, I had extensive conversations with ChatGPT-4o throughout my thesis supervision period. From mid-April, after my final face-to-face meeting with my human supervisor, through July 17th, I engaged in two long emotional conversations with ChatGPT, creating valuable data for analyzing AI-provided emotional support during this critical academic phase.

### 4.3.2 AI-Assisted Content Analysis Methodology

**Data collection** The data includes two conversation threads as previously mentioned. These threads were divided into 473 conversation pairs, with each pair consisting of one user message and its corresponding AI response.

**LLM prompting methodology** This content analysis employs a simplified AI-assisted methodology similar to that used in the coding-related tasks analysis. The primary objective is to quantify the distribution of emotion types and AI support types throughout our interactions. Since the categories for these types are more readily predefined compared to coding behaviors, the first three prompts (generation, classification, and aggregation) were omitted, utilizing only the prevalence prompt approach. Consistent with the previous analysis, two LLMs were used: Claude Sonnet 3.5 and ChatGPT-4o.

The prevalence prompt calculates the distribution across three aspects: my emotional states, AI support types, and dominant thesis-related topics within the conversation pairs. For emotional analysis, the prompt predefines seven emotion categories—hope,



confidence, relief, confusion, sadness, anxiety, and stress—and instructs both LLMs to assign one primary emotion to each conversation pair, with an optional secondary emotion classification. AI support types are categorized into four predefined types: validation, empathy, encouragement, and suggestion. The prompt directs both models to classify every conversation pair according to these emotional and support dimensions, enabling quantitative analysis of their distributions across the dataset.

**Inter-rater reliability validation** To validate the consistency of AI-assisted content analysis, I assessed inter-rater reliability between Claude 3.5 Sonnet and GPT-4o across two tasks: emotion classification and support type classification.

**Emotion classification reliability** For emotion annotation across 473 conversation pairs, the models demonstrated substantial agreement with 66.2% perfect agreement and Cohen’s kappa of 0.559 in primary emotion classification, and average Jaccard similarity of 0.708.<sup>13</sup>

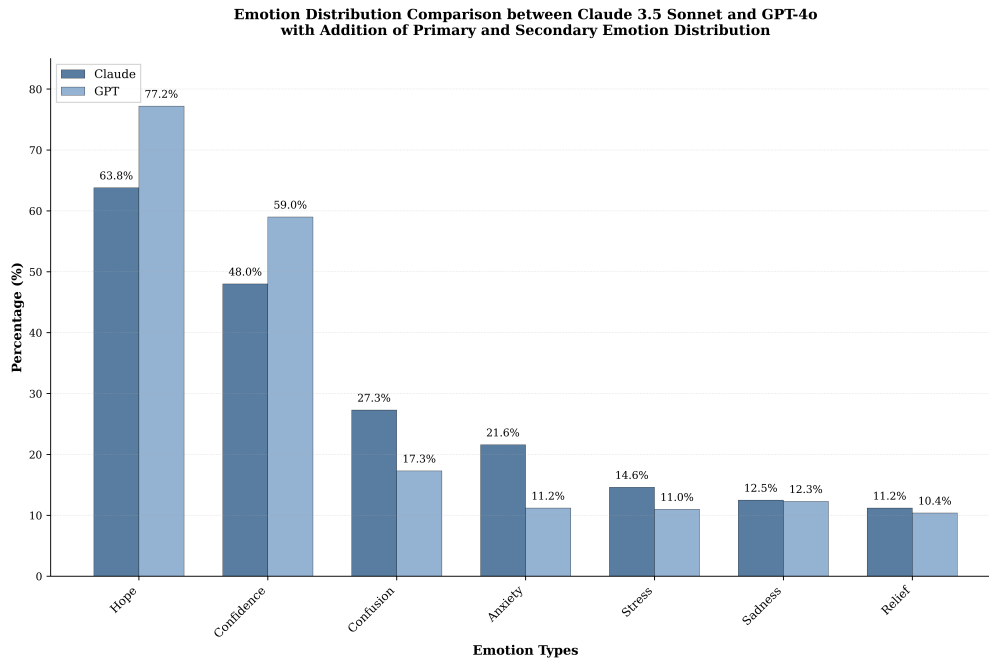
**Support type classification reliability** For AI support type classification, the models reached 76.3% perfect agreement and a Cohen’s kappa of 0.317 in identifying the primary support type. While the kappa value appears modest, this mainly reflects the highly imbalanced distribution of support types, with validation dominating both models’ classifications. Given the imbalance in support type distribution, with validation dominating both models’ classifications, the relatively high perfect agreement still suggests that the classification results are credible despite the modest kappa value.

### 4.3.3 Content Analysis

#### Emotional Type Distribution

The emotion distribution shows several noteworthy patterns in the AI-human emotional support interactions. Figure 3 demonstrates that **hope** (Claude: 63.8%, ChatGPT: 77.2%) and **confidence** (Claude: 48.0%, ChatGPT: 59.0%) emerge as the two most prevalent emotions, substantially surpassing all other emotional categories.

<sup>13</sup>Jaccard similarity is calculated as  $J = |A \cap B| / |A \cup B|$ , where  $A$  and  $B$  are the emotion sets identified by Claude and ChatGPT respectively.



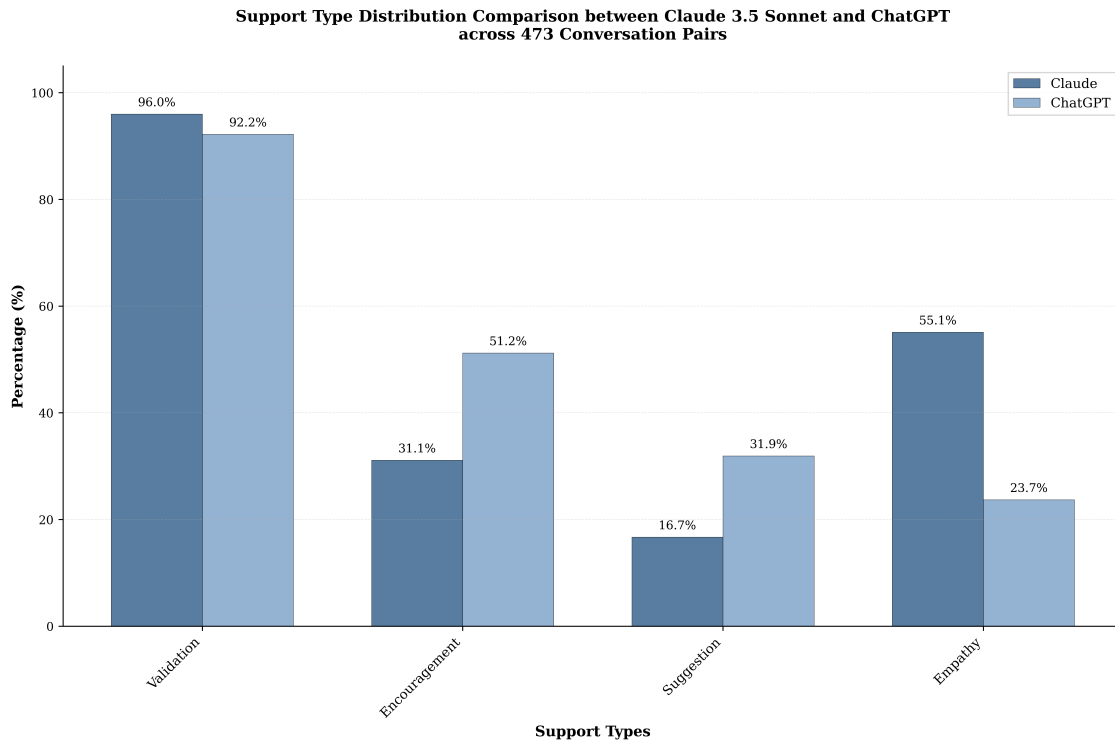
**Figure 3:** Emotion distribution comparison between Claude 3.5 Sonnet and GPT-4o with addition of primary and secondary emotion distribution

Among negative emotions, confusion (Claude: 27.3%, ChatGPT: 17.3%) and anxiety (Claude: 21.6%, ChatGPT: 11.2%) represent the most frequent categories, though their prevalence remains considerably lower than positive emotions.

The predominance of positive emotions provides indirect evidence for AI's effectiveness in emotional support. Given that I typically seek AI assistance during periods of low emotional states, the substantial prevalence of hope and confidence suggests successful emotional transformation through AI interaction. The high percentages of positive emotions indicate that conversations with AI might not only address immediate concerns but also foster lasting feelings of optimism and self-assurance.

### Emotional Support Type Distribution

The support type distribution shows patterns in how AI systems provide emotional support, with validation emerging as the predominant support type across both models (Figure 4). Claude classified 96.0% of interactions as containing validation, while ChatGPT identified validation in 92.2% of cases. This overwhelming prevalence of validation suggests that AI systems naturally gravitate toward affirming and legitimizing users' emotional experiences.



**Figure 4:** Support type distribution comparison between Claude 3.5 Sonnet and ChatGPT across 473 conversation pairs

Empathy and encouragement serve as the primary secondary support mechanisms, with empathy being Claude’s preferred secondary approach (50.8% vs. ChatGPT’s 19.5%) and encouragement being ChatGPT’s preference (44.6% vs. Claude’s 29.2%). Suggestion represents the least common support type in general (Claude: 16.7%, ChatGPT: 31.9%).

#### 4.3.4 Qualitative Analysis

Completing a graduation thesis is, in many senses, a personal journey that can become lonely at times. My journey has been somewhat more isolated since, due to the nature of my research, there has been an absence of human supervision for an extended period. While I was granted considerable freedom, this freedom sometimes weighs on me. During this period, particularly in the early stages when the methodology framework and timeline were not yet clear and I was uncertain whether my nail biting detection study was within my technical capabilities, the challenges felt especially pronounced.

My research is somewhat "different" from more conventional approaches, which has sometimes led to questioning rather than support when I discuss my thesis with others.

In these situations, ChatGPT provided substantial emotional support through its constant accessibility and non-judgmental approach.

I view it essentially as a journaling tool that mirrors people's thoughts and feelings. While it can never "really" relate to people, it is programmed to excel at understanding and validating people's emotions at the language level. Often, people including myself simply need a channel to make their feelings and emotions feel heard and understood.

Regarding research supervision, emotional support constitutes an important but not compulsory component of supervision, so how this unfolds between human supervisors and students varies considerably from person to person. The emotional support AI provides cannot compete with that offered by a good supervisor who genuinely cares, but AI does provide consistent "mediocre" emotional support that helps people feel better when things become difficult. It also has unique advantages such as accessibility, which humans cannot easily provide or find challenging to maintain.

# Chapter 5

## Discussion

### 5.1 Answers to RQs

#### 5.1.1 To What Extent Can GenAI Fulfill the Key Responsibilities of a Thesis Supervisor in a Graduation Project?

Among the three dimensions in this thesis's framework of supervision responsibilities—functional, critical thinking, and emotional support—GenAI excels at functional supervision, particularly in topic selection, literature review, coding-related tasks, and writing-related tasks. It is fundamentally deficient in the critical thinking dimension while fulfilling considerable responsibilities in emotional support. **In summary, GenAI demonstrates strong performance in functional tasks, moderate effectiveness in emotional support, and significant limitations in critical thinking.**

It is important to note that the original enculturation dimension was removed, and the emancipation and relationship development dimensions were consolidated into emotional support due to GenAI's inherent inability to supervise these aspects. **This simplified supervision responsibility framework may present GenAI in a more favorable light by reducing evaluation criteria.**

This finding aligns with observations from previous studies. Dai et al. (2023), after interviewing 20 postgraduate research students with ChatGPT experience, found that personalized tutoring and explanation, language editing and proofreading, brainstorming and ideation, code interpretation and debugging, and literature processing and synthesis were the five most frequent research tasks students engaged with AI. **These tasks align almost perfectly with the functional tasks better fulfilled by GenAI in my research, suggesting that students naturally gravitate toward AI for tasks where it demonstrates better performance.**

GenAI's limitations in providing critical thinking supervision have also been documented in prior research. For instance, Jensen et al. (2025) compared feedback from ChatGPT and

human supervisors and found that, whereas human supervisors highlighted fundamental weaknesses in students' manuscripts, the chatbot generally failed to challenge underlying assumptions and instead tended to repeat or validate them.

This lack of capacity in critical thinking supervision is likely due to the training methodology of large language models. LLMs are typically optimized to produce outputs that humans rate favorably through techniques such as reinforcement learning from human feedback (RLHF). This training approach inherently encourages model responses that align with user beliefs rather than prioritizing truthfulness—a phenomenon known as sycophancy (Sharma et al., 2025). Research demonstrates that sycophancy is a general behavior of LLMs (Sharma et al., 2025).

GenAI's considerable ability in emotional support is supported by studies extending beyond research supervision to general emotional well-being. Yin, Jia, and Waksalak (2024) found that AI has the ability to make people feel heard, while other research (Spytska, 2025) has argued that chatbots are particularly beneficial in crisis settings where access to therapists is limited, demonstrating their value in scalability and availability.

### **Why does GenAI excel in functional supervision, perform moderately in emotional support, and struggle with critical thinking?**

Examining the functional tasks where GenAI performs best—topic selection, literature review, coding-related tasks, and writing and polishing—reveals several common characteristics: extensive information processing requirements, pattern recognition demands, structured output formats, and reliance on reorganizing existing knowledge. For instance, identifying research trends and gaps in topic selection, searching and synthesizing literature, understanding technical documentation in coding, and integrating information coherently in writing all involve extensive pattern recognition and knowledge-based operations.

In contrast, other functional tasks like timeline management require strategy, judgment, and accountability rather than knowledge processing, while methodology development demands high-level theoretical thinking. This explains GenAI's limited performance in these areas. Critical thinking is fundamentally lacking due to LLM training methodologies, as discussed previously. For emotional support, although GenAI demonstrates considerable performance and potential, it cannot provide human-level emotional depth, adaptability, and engagement because emotional support is fundamentally a human-centered process.

### 5.1.2 What Are the Strengths and Limitations of GenAI When Functioning as a Research Supervisor Compared to Human Supervisors?

GenAI's strengths as a research supervisor lie in its inherent characteristics as a technological tool: 24/7 availability, universal accessibility, and consistent objectivity. Students can seek assistance anytime and anywhere without scheduling constraints or geographical limitations. Unlike human supervisors, AI remains uninfluenced by personal relationships, individual preferences, personalities, or subjective biases.

However, GenAI's limitations stem from the same fundamental nature—its essence as a tool rather than a human. It lacks the capacity for nuanced academic judgment that comes from years of research experience, cannot provide deep domain expertise rooted in disciplinary understanding, and is unable to establish meaningful relational bonds that characterize effective human supervision.

This represents the paradox of AI supervision: GenAI's strengths and limitations are two sides of the same coin, both from its nature as a technological instrument. Its machine-like qualities deliver advantages that would be impossible to consistently demand from human supervisors—perfect availability, unwavering objectivity, and freedom from the interpersonal complexities that can complicate human-student relationships. Conversely, these same qualities prevent it from offering the experiential wisdom, contextual judgment, and authentic mentoring relationships that define the most valuable aspects of human supervision. Ultimately, GenAI excels at providing standardized, accessible support while falling short in areas requiring human insight, emotional intelligence, and the value of genuine academic mentorship.

### 5.1.3 How Can GenAI Be Optimally Prompted and Utilized to Achieve the Best Supervisory Performance?

First, it is important to clarify that the focus of this thesis is to evaluate GenAI's capacity to fulfill research supervision responsibilities rather than to develop prompt engineering techniques. This research question is included because how GenAI is utilized as a tool directly influences its supervisory effectiveness. Based on my experience throughout this study, optimal GenAI utilization for supervision appears to depend on several key principles:

**Task-specific deployment:** The effective use of GenAI requires tailoring the tool to the nature of the task rather than relying on a single, uniform approach. For coding tasks, for instance, AI agents might be most efficient; for timeline management, integration with

external project management applications may yield better results; and for mock defense preparation, conversational voice-based tools such as ChatGPT's voice mode can be more appropriate.

**Contextual prompting:** Providing sufficient background information, research constraints, and specific objectives significantly improves AI response quality. This was particularly evident in topic selection and methodology discussions, where detailed context led to more relevant and useful guidance.

**Continuous adaptation:** LLMs and their applications are evolving at an unprecedented pace, with new models, protocols like MCP (Model Context Protocol), and capabilities emerging regularly. To maximize AI's utility, it is much needed to always learn about available tools and actively experiment with customization to best serve specific research needs.

## 5.2 Using AI or Being Supervised by AI? A Brief Discussion

A natural question arises: does this research truly demonstrate *being supervised* by AI, or is it more accurately described as *using* AI as a tool? **Strictly speaking, GenAI lacks the institutional authority, accountability, critical evaluation and relational aspects that define a supervisor, and in this sense it remains a tool.** At the same time, the fact that it can perform tasks that supervisors once provided—such as suggesting literature, commenting on drafts, giving structured feedback, providing ideas and emotional support—means that it simulates some supervisory functions more closely than traditional tools. This places GenAI in an intermediate position: more than a tool, but not yet a supervisor. For this reason, “being supervised by AI” in this thesis should be understood as an experimental framing: the project was conducted with AI-enabled supervisory functions rather than under literal AI supervision.

Viewed from a broader historical and forward-looking perspective, it is useful to frame the distinction between tools and supervisors not as a fixed boundary but as a continuum that has evolved over time and may continue to shift in the future. At one end are pure tools, such as search engines or text processors. At the other end are true supervisors, defined by criticality, accountability, empathy, and institutional recognition. Current GenAI systems occupy a middle ground closer to an “assistant” or “partner,” able to perform certain supervisory functions but without the authority or responsibility of a supervisor. Whether future AI can be regarded as providing *real supervision* will depend on its ability to develop genuine critical judgment, accountability, and relational depth.



### 5.3 Innovations and Limitations

This thesis employs a self-study design to examine GenAI supervision. While self-studies are used in the small body of work on AI in supervision (e.g., Jensen et al. (2025); Boyd and Harding (2025)), this project is relatively innovative in three ways: (1) a real-stakes experiment, in which—beyond initial topic selection and parts of the methodology—thesis was produced under AI guidance; (2) an LLM-based content analysis, used alongside qualitative evaluation to assess coding-related and emotional support; and (3) a dual-layer experimental structure that embeds a technical case study (nail-biting detection) to evaluate different aspects of AI supervision.

However, this methodology also presents significant limitations. The  $N=1$  sample size restricts generalizability, as findings may reflect individual characteristics, research fields, or specific AI models rather than universal patterns. The self-evaluation component introduces potential bias, as I served as both experimenter and subject. The high personal investment in demonstrating the viability of AI supervision could also unconsciously influence data interpretation and evaluation criteria. In addition, transparency and verifiability are inherently constrained: this was a single-operator study and—despite sharing code snippets and conversation data—many supervisory interactions and ad hoc decisions (e.g., prompting and response selection) cannot be independently reproduced.

Beyond these generic constraints, the nail-biting case study adds domain-specific limitations. The dataset is small ( $N=367$ ), self-labeled, with heterogeneous sources (web images and the 11k Hands dataset). Labels were created by a single annotator, and no inter-rater reliability was established. Model selection relied on a single random train/validation split without cross-validation or an external test set; hyperparameters were tuned on the validation split (risking optimistic bias), the search space was relatively narrow, and results showed run-to-run variance (e.g., the best  $wd=0.01$  run was not consistently reproducible). Collectively, these factors mean the reported numbers should be regarded as indicative best runs rather than statistically stable estimates.

# Chapter 6

## Conclusion

This thesis set out to evaluate the extent to which generative AI can fulfill the responsibilities of a research supervisor through a real-stakes, self-study design. I implemented a nested structure in which GenAI (primarily Claude and ChatGPT) supervised both the overall thesis and a technical case study—nail-biting detection via computer vision—allowing me to assess supervision across functional, critical-thinking, and emotional-support dimensions.

Three findings stand out. First, GenAI excels at functional supervision. It was consistently useful for topic exploration, targeted literature sourcing, coding and debugging, and writing support. In the case study, AI guidance enabled me—without prior deep-learning or computer-vision background—to build an end-to-end model and reach a best observed performance of 89.7% accuracy, 93.3% precision, 84.8% recall, and 88.9% F1 after hyperparameter tuning. These numbers are indicative rather than definitive, but they demonstrate that AI supervision can help a student reach a workable technical solution in a limited timeframe.

Second, GenAI offers moderate emotional support. Across interactions, validation dominated AI responses, with frequent expressions of empathy and encouragement. **This produced a noticeable lift in hope and confidence during periods of uncertainty.** Still, AI's support remains language-level and lacks the relational depth, continuity, and pastoral care a good human supervisor can provide.

Third—and most critically—GenAI remains weak in critical-thinking supervision. It seldom challenged assumptions, flagged methodological risks early, or insisted on best practices. The result is a pattern of fast, agreeable help rather than deliberate, adversarial testing of ideas. This limitation is central: without systematic challenge, students risk over-confidence, narrow searches, and avoidable detours.

Taken together, these results position current GenAI as more than a tool yet not a supervisor—closer to an assistant or partner that can perform some supervisory functions but lacks authority, accountability, and the capacity for disciplined critique. The practical implication is a hybrid model: deploy GenAI where it is strongest (information processing,

code generation, drafting, formatting, structured feedback), and rely on human supervisors for judgment-heavy tasks (challenging assumptions, setting strategy, and mentoring identity and ethics).

In sum, GenAI already delivers real value in functional supervision and usable emotional support, but it does not replace human supervision. Where it is incorporated thoughtfully—task-by-task, with clear standards and human oversight—it can increase student independence and productivity without sacrificing academic quality. Whether future AI crosses the boundary into “real supervision” will depend on progress in critical judgment, accountability, and relational depth. **Until then, the most responsible path is a division of labor: use AI for what it does best, keep humans where stakes and judgment are highest, and make the collaboration explicit.**

# References

- Abiddin, Norhasni Zainal, Aminuddin Hassan, and Ahmad Rafee Ahmad (2009). "Research student supervision: An approach to good supervisory practice". In: *The Open Education Journal* 2.1, pp. 11–16.
- Afifi, Mahmoud (2019). "11K Hands: gender recognition and biometric identification using a large dataset of hand images". In: *Multimedia Tools and Applications* 78.15, pp. 20835–20854.
- Alesmaeil, Abdullah and Eftal Şehirli (2025). "Real-time nail-biting detection on a smartwatch using three CNN models pipeline". In: *Computational Intelligence*. DOI: 10.1111/coin.70020. URL: <https://onlinelibrary.wiley.com/doi/10.1111/coin.70020>.
- Azaria, Asaph, Brian Mayton, and Joseph Paradiso (2016). "Thumbs-Up: Wearable Sensing Device for Detecting Hand-to-Mouth Compulsive Habits". In: *Proceedings of the 9th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC)*. SCITEPRESS. Rome, Italy. URL: <https://resenv.media.mit.edu/pubs/papers/ThumbsUp-BIOSTEC16.pdf>.
- Bastalich, Wendy (2015). "Content and context in knowledge production: a critical review of doctoral supervision literature". In: *Studies in Higher Education* 42.7, pp. 1145–1157.
- Belkina, Marina, Scott Daniel, Sasha Nikolic, Rezwanul Haque, Sean Lyden, Paul Neal, Sally Grundy, and Gihan Mas Hassan (2025). "Implementing generative AI (GenAI) in higher education: A systematic review of case studies". In: *Computers and Education: Artificial Intelligence* 8, p. 100407. DOI: 10.1016/j.caeai.2025.100407.
- Blei, David M, Andrew Y Ng, and Michael I Jordan (2003). "Latent dirichlet allocation". In: *Journal of machine Learning research* 3.Jan, pp. 993–1022.
- Boyd, Philippa and Debs Harding (2025). "Generative AI: reconfiguring supervision and doctoral research". In: *Buildings & Cities* 6.1.
- Braun, Victoria and Virginia Clarke (2006). "Using thematic analysis in psychology". In: *Qualitative Research in Psychology* 3.2, pp. 77–101. DOI: 10.1191/1478088706qp063oa.
- Butler-Adam, John (2018). "The Fourth Industrial Revolution and education". In: *South African Journal of Science* 114.5-6, p. 1. DOI: 10.17159/sajs.2018/a0271.
- Dai, Yinan, Siyuan Lai, Cher Ping Lim, and An Liu (2023). "ChatGPT and its impact on research supervision: Insights from Australian postgraduate research students". In: *Australasian Journal of Educational Technology* 39.4, pp. 74–88. DOI: 10.14742/ajet.8843.

- Deuchar, Ross (2008). "Facilitator, director or critical friend? Contradiction and congruence in doctoral supervision styles". In: *Teaching in Higher Education* 13.4, pp. 489–500. DOI: 10.1080/13562510802193905.
- Dwivedi, Yogesh K. et al. (2025). "Generative Artificial Intelligence: Evolving Technology, Growing Societal Impact, and Opportunities for Information Systems Research". In: *Information Systems Frontiers*. DOI: 10.1007/s10796-025-10581-7.
- Ennis, Robert H. (1987). "A taxonomy of critical thinking dispositions and abilities". In: *Teaching thinking skills: Theory and practice*. Ed. by Joan B. Baron and Robert J. Sternberg. W H Freeman/Times Books/Henry Holt & Co, pp. 9–26.
- Gilardi, Fabrizio, Meysam Alizadeh, and Maël Kubli (2023). "ChatGPT outperforms crowd workers for text-annotation tasks". In: *Proceedings of the National Academy of Sciences* 120.30.
- Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep Learning*. Cambridge, MA: MIT Press.
- Grant, Kate, Ray Hackney, and David Edgar (2014). "Postgraduate research supervision: An 'agreed' conceptual view of good practice through derived metaphors". In: *International Journal of Doctoral Studies* 9, pp. 43–60. URL: <http://ijds.org/Volume9/IJDSv9p043-060Grant0403.pdf>.
- Guerin, Cally, Hedy Kerr, and Inger Green (2015). "Supervision pedagogies: narratives from the field". In: *Teaching in Higher Education* 20, pp. 107–118.
- Haley, Aimee, Mona Holmqvist, and Karmen Johansson (2024). "Supervisors' competences from doctoral students' perspectives – a systematic review". In: *Educational Review*. Published online: 09 Feb 2024.
- Han, Seung Seog, Gyeong Hun Park, Woohyung Lim, Myoung Shin Kim, Jung Im Na, Ilwoo Park, and Sung Eun Chang (2018). "Deep neural networks show an equivalent and often superior performance to dermatologists in onychomycosis diagnosis: Automatic construction of onychomycosis datasets by region-based convolutional deep neural network". In: *PLOS ONE* 13.1, e0191493. DOI: 10.1371/journal.pone.0191493. URL: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0191493>.
- Higher Education Policy Institute (Feb. 2025). *Student Generative AI Survey 2025*. 92% of students now using AI in some form, up from 66% in 2024, and 88% having used GenAI for assessments. URL: <https://www.hepi.ac.uk/2025/02/26/student-generative-ai-survey-2025/>.
- Horáková, Tereza, Milan Houška, and Ludmila Dömeová (2017). "Classification of the Educational Texts Styles with the Methods of Artificial Intelligence". In: *Journal of Baltic Science Education* 16.3, pp. 324–336.
- Ifargan, Tal, Lukas Hafner, Maor Kern, Ori Alcalay, and Roy Kishony (2024). *Autonomous LLM-driven research from data to human-verifiable research papers*. arXiv: 2404.17605 [q-bio.OT]. URL: <https://arxiv.org/abs/2404.17605>.

- Jensen, Lasse X, Margaret Bearman, David Boud, and Flemming Konradsen (Mar. 2025). "Feedback encounters in doctoral supervision: the role of generative AI chatbots". In: *Assessment & Evaluation in Higher Education*. Published online: 18 Mar 2025. DOI: 10.1080/02602938.2025.2478155.
- Jimenez, Carlos E., John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan (2024). *SWE-bench: Can Language Models Resolve Real-World GitHub Issues?* arXiv: 2310.06770 [cs.CL]. URL: <https://arxiv.org/abs/2310.06770>.
- Khang, Alex, Muthmainnah Muthmainnah, Piyus Mahmud Ipukteu Seraj, Abdullah Al Yakin, and Ahmed J. Obaid (2023). "AI-aided teaching model in education 5.0". In: *Handbook of Research on AI-Based Technologies and Applications in the Era of the Metaverse*. Ed. by Alex Khang, Vrince Shah, and Sita Rani. Hershey, PA, USA: IGI Global, pp. 83–104.
- Kirwan, A. (2023). "ChatGPT and university teaching, learning and assessment: Some initial reflections on teaching academic integrity in the age of Large Language Models". In: *Irish Educational Studies*, pp. 1–18. DOI: 10.1080/03323315.2023.2284901.
- Kuramitsu, Kimio, Yuki Obara, Masahiro Sato, and Mitsuki Obara (2023). "Kogi: A seamless integration of ChatGPT into Jupyter environments for programming education". In: *Proceedings of the 2023 ACM SIGPLAN International Symposium on SPLASH-E*. Cascais, Portugal: ACM. DOI: 10.1145/3622780.3623648.
- Landis, J. Richard and Gary G. Koch (1977). "The measurement of observer agreement for categorical data". In: *Biometrics* 33.1, pp. 159–174.
- Lecourt, Florian, Madalina Croitoru, and Konstantin Todorov (May 2025). "'Only ChatGPT gets me': An Empirical Analysis of GPT versus other Large Language Models for Emotion Detection in Text". In: *Companion Proceedings of the ACM on Web Conference 2025*. WWW '25. ACM, pp. 2603–2611. DOI: 10.1145/3701716.3718375. URL: <http://dx.doi.org/10.1145/3701716.3718375>.
- Lee, Anne (2007). "Developing effective supervisors: Concepts of research supervision". In: *South African Journal of Higher Education* 21.4, pp. 680–693.
- (2008). "How are doctoral students supervised? Concepts of doctoral research supervision". In: *Studies in Higher Education* 33.3, pp. 267–281. DOI: 10.1080/03075070802049202. URL: <https://doi.org/10.1080/03075070802049202>.
- Li, Hao, Haoxiang Zhang, and Hassan (2025). *The Rise of AI Teammates in Software Engineering (SE) 3.0: How Autonomous Coding Agents Are Reshaping Software Engineering*. arXiv: 2507.15003 [cs.SE].
- Lin, Huei-Chuan Katherine, Chih-Hung Wang, Chi-Jen Chao, and Meng-Kai Chien (2012). "Employing textual and facial emotion recognition to design an affective tutoring system". In: *Turkish Online Journal of Educational Technology-TOJET* 11.4, pp. 418–426.
- Lo, Chung Kwan (2023). "What is the impact of ChatGPT on education? A rapid review of the literature". In: *Education Sciences* 13.4, p. 410. DOI: 10.3390/educsci13040410.

- McCallin, Antoinette and Shamima Nayar (2012). "Postgraduate research supervision: a critical review of current practice". In: *Teaching in Higher Education* 17.1, pp. 63–74. DOI: 10.1080/13562517.2011.590979.
- McHugh, Mary L. (2012). "Interrater reliability: the kappa statistic". In: *Biochemia Medica* 22.3, pp. 276–282.
- Melo, F. R., E. L. Flôres, S. D. Carvalho, R. A. G. Teixeira, L. F. B. Loja, and R. de Sousa Gomide (2014). "Computational organization of didactic contents for personalized virtual learning environments". In: *Computers & Education* 79, pp. 126–137.
- Merriam-Webster (n.d.). *Supervise*. In *Merriam-Webster.com dictionary*. Retrieved August 21, 2025. URL: <https://www.merriam-webster.com/dictionary/supervise>.
- Microsoft (May 2025). *Microsoft Build 2025: The age of AI agents and building the open agentic web*. <https://blogs.microsoft.com/blog/2025/05/19/microsoft-build-2025-the-age-of-ai-agents-and-building-the-open-agentic-web/>. Accessed August 10th.
- Ngai, Grace, Stephen Chi Fai Chan, Joseph Chung Yin Chan, and Winnie Wing Yi Lau (2010). "Deploying a wearable computing platform for computing education". In: *IEEE Transactions on Learning Technologies* 3.1, pp. 45–55.
- Pitso, Thabo (2023). "Post-COVID-19 higher learning: Towards Telagogy, A web-based learning experience". In: *IAFOR Journal of Education* 11.2, pp. 39–59. DOI: 10.22492/ije.11.2.02.
- Qureshi, Rashida and Neelofar Vazir (2016). "Pedagogy of research supervision pedagogy: A constructivist model". In: *Research in Pedagogy* 6.2, pp. 95–110.
- Rao, Varun Nagaraj, Eesha Agarwal, Samantha Dalal, Dan Calacci, and Andrés Monroy-Hernández (2025). *QuaLLM: An LLM-based Framework to Extract Quantitative Insights from Online Forums*. arXiv: 2405.05345 [cs.CL]. URL: <https://arxiv.org/abs/2405.05345>.
- Searle, Benjamin Lucas, Dimitris Spathis, Marios Constantinides, Daniele Quercia, and Cecilia Mascolo (2021). *Anticipatory Detection of Compulsive Body-focused Repetitive Behaviors with Wearables*. arXiv: 2106.10970 [cs.HC]. URL: <https://arxiv.org/abs/2106.10970>.
- Severinsson, E. (Mar. 2012). "Research supervision: supervisory style, research-related tasks, importance and quality – part 1". In: *Journal of Nursing Management* 20.2, pp. 215–223. DOI: 10.1111/j.1365-2834.2011.01361.x.
- Shandilya, Gunjan, Sheifali Gupta, Salil Bharany, Ateeq Ur Rehman, Upinder Kaur, Hafizan Mat Som, and Seada Hussien (2024). "Autonomous detection of nail disorders using a hybrid capsule CNN: a novel deep learning approach for early diagnosis". In: *BMC Medical Informatics and Decision Making* 24.1, p. 414. DOI: 10.1186/s12911-024-02840-5. URL: <https://bmcmmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-024-02840-5>.

- Sharma, Mrinank, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askill, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez (2025). *Towards Understanding Sycophancy in Language Models*. arXiv: 2310.13548 [cs.CL]. URL: <https://arxiv.org/abs/2310.13548>.
- Sokolova, Marina and Guy Lapalme (2009). "A systematic analysis of performance measures for classification tasks". In: *Information Processing & Management* 45.4, pp. 427–437.
- Spytska, L. (2025). "The use of artificial intelligence in psychotherapy: development of intelligent therapeutic systems". In: *BMC Psychology* 13, p. 175. DOI: 10.1186/s40359-025-02491-9. URL: <https://doi.org/10.1186/s40359-025-02491-9>.
- Stack Overflow (2025). *Developer Survey 2025*. <https://survey.stackoverflow.co/2025/ai>. Accessed August 10th.
- Tai, Robert H., Leigh R. Bentley, Xiaoran Xia, Jessica M. Sitt, Sarah C. Fankhauser, Ana M. Chicas-Mosier, and Brian G. Monteith (2024). "An examination of the use of large language models to aid analysis of textual data". In: *International Journal of Qualitative Methods* 23. DOI: 10.1177/16094069241231168.
- Tang, Jian, Zhaoshi Meng, XuanLong Nguyen, Qiaozhu Mei, and Ming Zhang (2014). "Understanding the limiting factors of topic modeling via posterior contraction analysis". In: *Proceedings of the 31st International Conference on Machine Learning*. Vol. 32. ICML'14. JMLR.org, pp. I-190–I-198.
- Vattam, Swaroop, Ashok K. Goel, and Spencer Rugaber (2011). "Understanding complex natural systems by articulating structure-behavior function models". In: *Educational Technology & Society* 14.1, pp. 66–81.
- Wang and Y. Feng (2024). "An experimental study of ChatGPT-assisted improvement of Chinese college students' English reading skills: A case study of dear life". In: *Proceedings of the 15th International Conference on Education Technology and Computers*. Barcelona, Spain: ACM. DOI: 10.1145/3629296.3629300.
- Xu, Yingjie, Shufeng Ye, and Xiwen Zhu (Sept. 2023). "The ScholarNet and Artificial Intelligence (AI) supervisor in material science research". In: *Journal of Physical Chemistry Letters* 14.36, pp. 7981–7991. DOI: 10.1021/acs.jpcllett.3c01668.
- Yang, Chien-Wei, Bor-Chen Kuo, and Chih-Hung Liao (2011). "A HO-IRT based diagnostic assessment system with constructed response items". In: *Turkish Online Journal of Educational Technology TOJET* 10.4, pp. 46–51.
- Yin, Yixin, Nan Jia, and Cheryl J. Waksalak (2024). "AI can help people feel heard, but an AI label diminishes this impact". In: *Proceedings of the National Academy of Sciences* 121.14, e2319112121. DOI: 10.1073/pnas.2319112121. URL: <https://doi.org/10.1073/pnas.2319112121>.
- Zhai, Xuesong, Xiaoyan Chu, Ching Sing Chai, Morris Siu Yung Jong, Andreja Istenic, Michael Spector, Jia-Bao Liu, Jing Yuan, Yan Li, and Ning Cai (2021). "A Review of



Artificial Intelligence (AI) in Education from 2010 to 2020". In: *Complexity*. DOI: 10 . 1155/2021/8812542.

# Appendices

## A Nail Biting Detection Case Study

### A.1 Introduction and Research Context

This nail biting detection case study serves as a technical project to evaluate AI's effectiveness in providing coding-related supervision. The project addresses the computer vision task of automatically identifying nail biting behaviors from digital images through deep learning methodologies. The selection of this specific technical challenge was to provide sufficient complexity for assessing AI supervision across the complete machine learning pipeline—from problem formulation through model optimization.

### A.2 Literature Review

There has been no academic research specifically using computer vision to detect nail biting behavior from images. While a few high-quality studies have explored nail biting detection using wearable devices—such as Alesmaeil and Şehirli (2025) who developed a smartwatch-based system using inertial sensors, and Searle et al. (2021) who achieved  $AUC > 0.90$  using multi-sensory wearable approaches—these rely on motion data rather than visual analysis. Additionally, dozens of high-quality studies focus on nail disease and disorder detection using image analysis, including Shandilya et al. (2024) achieving 99.40% training accuracy for various nail pathologies and Han et al. (2018) demonstrating a diagnostic accuracy for onychomycosis using deep learning that was superior to that of most of the dermatologists who participated in the study. However, none of these studies specifically target onychophagia (nail biting behavior). This represents a quite niche research direction that current academic literature has not specifically covered, likely due to its limited research value.

## A.3 Methodology and Approaches

### Data Collection

The dataset comprises 367 nail images<sup>1</sup> collected from two sources: approximately 200 images crawled from internet sources using the `bing-image-downloader` Python package and 167 images selected from the "11k Hands" dataset (Afifi, 2019). All images were manually labeled into two categories: bitten and not-bitten. The final dataset was split into 295 training samples and 72 validation samples.

### Image Segmentation Approach

This approach employs a two-stage pipeline: semantic segmentation to delineate nail structures (nail contour, nail fold, distal border), followed by morphological analysis and rule-based classification based on medical knowledge.

Four U-Net models were trained using two different architectures (Vanilla U-Net and ResNet34-UNet) on two 100-image datasets with manually labeled nail contours. The ResNet34-based models achieved good training performance (Dice scores 0.86-0.87, IoU scores 0.75-0.76), as shown in Table 1.

**Table 1:** U-Net Models Performance Comparison

Model	Training Data	Architecture	Performance Metrics	
			Dice	IoU
1	Internet (100 images)	Vanilla U-Net	0.3593	0.2444
2	11k Hands (100 images)	Vanilla U-Net	0.5609	0.4298
3	Internet (100 images)	ResNet34-UNet	<b>0.8568</b>	<b>0.7506</b>
4	11k Hands (100 images)	ResNet34-UNet	<b>0.8658</b>	<b>0.7636</b>

However, cross-dataset testing revealed severe generalization issues: the internet-trained model achieved only Dice  $\approx 0.30$  and IoU  $\approx 0.21$  when tested on the "11k Hands" dataset. This poor generalization performance, combined with the time-intensive requirement for extensive segmentation labels, led to the deprecation of this approach in favor of the end-to-end classification methodology.

<sup>1</sup>The complete nail biting detection dataset is available at: [https://github.com/Yunshan-CAI/AI\\_supervision\\_thesis.git](https://github.com/Yunshan-CAI/AI_supervision_thesis.git)

## End-to-End Classification Approach

This approach<sup>2</sup> utilizes transfer learning where a pre-trained convolutional neural network is fine-tuned on labeled datasets with two classes: bitten and not-bitten. The model learns visual patterns directly from raw images and performs classification without intermediate segmentation.

A ResNet34-based classifier was trained using transfer learning with ImageNet pre-trained weights. Input images were resized to 192×192 pixels. The model employed a modified classification head for two-class prediction, trained with fine-tuning approach using 10 epochs and base learning rate of 1e-3.

Initial experiments used a 100-image subset from the "11k Hands" dataset, achieving 73.3% accuracy but exhibiting severe class imbalance issues—predicting zero "bitten" labels in the validation set due to underrepresentation in the training data. To address this limitation, the dataset was expanded to 367 images by incorporating internet-sourced data, resulting in improved performance with the baseline model achieving **79.41% accuracy, 80.65% precision, 75.76% recall, and 78.13% F1-score**.

## Hyperparameter Optimization

Based on the end-to-end classification approach, systematic hyperparameter optimization was conducted to improve model performance beyond the baseline. The optimization techniques were organized into four main categories: Learning Rate Optimization (standard learning rate tuning and discriminative learning rates), Data Augmentation (presizing, progressive resizing, and test time augmentation), Regularization Techniques (weight decay parameter tuning and label smoothing), and Model Architecture comparison (ResNet34, ResNet50, and EfficientNet-B0). Among all optimization techniques tested, two methods provided the most consistent improvements: **weight decay optimization and learning rate tuning**.

**Learning-rate selection** I first ran fastai's LR Range Test (`learn.lr_find()`), which sweeps the LR and reports a heuristic "valley": it suggested 5.75e-4. Using this as a reference, I performed a small grid search from 1e-4 to 7e-3 (Table 2). Performance improved markedly in the 3e-3–7e-3 range, with the best F1 and accuracy at 7e-3 (F1 = 85.71%, Acc = 86.76%), despite the highest validation loss (1.136). The lowest validation loss occurred at 2e-3 (0.506; F1 = 75.00%). The highest recall was at 5e-3 (84.85%), while 3e-3 offered a balanced trade-off (Precision = Recall = F1 = 81.82%, Acc = 82.35%).

---

<sup>2</sup>The complete training code for the baseline model and hyperparameter fine-tuning is available at: [https://github.com/Yunshan-CAI/AI\\_supervision\\_thesis.git](https://github.com/Yunshan-CAI/AI_supervision_thesis.git)

**Table 2: Learning Rate Optimization Results**

Learning Rate	Precision	Recall	Accuracy	F1 Score	Valid Loss
1e-4	61.11%	66.67%	63.24%	63.77%	0.674
5.75e-4 (Valley)	82.76%	72.73%	79.41%	77.42%	0.514
1e-3 (Baseline)	78.79%	78.79%	79.41%	78.79%	0.612
2e-3	77.42%	72.73%	76.47%	75.00%	<b>0.506</b>
3e-3	81.82%	81.82%	82.35%	81.82%	0.654
5e-3	73.68%	<b>84.85%</b>	77.94%	78.87%	0.896
7e-3	<b>90.00%</b>	81.82%	<b>86.76%</b>	<b>85.71%</b>	1.136

**Weight decay optimization** Across the tested values of the weight decay parameter (0–0.3), the best single-run performance was observed at  $wd = 0.01$  (the fastai default baseline), with F1 = 88.89%, Acc = 89.71%, Precision = 93.33%, Recall = 84.85%, and Valid Loss = 0.342. However, this peak result was not always reproducible in subsequent runs, reflecting the sensitivity of small datasets to random initialization and data order. Larger values of weight decay ( $\geq 0.1$ ) consistently reduced accuracy and F1. These findings suggest that the default configuration already provides an appropriate level of regularization for this dataset, while the reported numbers should be interpreted as indicative rather than definitive.

**Table 3: Weight Decay Parameter Optimization Results**

Weight Decay	Precision	Recall	Accuracy	F1 Score	Valid Loss
0.0	86.67%	78.79%	75.76%	82.35%	0.497
0.0001	86.21%	78.79%	75.76%	82.35%	0.524
0.0003	78.79%	78.79%	75.76%	76.47%	0.638
0.001	77.78%	84.85%	81.82%	77.94%	0.522
0.003	81.25%	78.79%	75.76%	79.41%	0.478
<b>0.01</b>	<b>93.33%</b>	<b>84.85%</b>	<b>89.71%</b>	<b>88.89%</b>	<b>0.342</b>
0.1	75.68%	84.85%	79.41%	80.00%	0.474
0.3	83.87%	78.79%	78.79%	80.88%	0.489

## A.4 Results and Limitations

The best-performing model in this case study was the ResNet34 classifier with weight decay ( $wd = 0.01$ ), which achieved **89.71% accuracy, 93.33% precision, 84.85% recall, and**

**88.89% F1-score.** Compared to the baseline model (79.41% accuracy, 78.13% F1-score), this represents a considerable improvement of approximately +10.3% in accuracy and +10.8% in F1. The learning rate grid search also identified effective operating regions, and weight decay optimization confirmed that the default setting already offered the most appropriate level of regularization for this dataset.

At the same time, several limitations must be acknowledged. First, the dataset (367 images) was manually assembled and labeled by myself, which introduces potential bias both in the selection of images and in the labeling process. Second, due to time constraints, the fine-tuning and hyperparameter optimization were not conducted under a fully rigorous protocol: each configuration was run only once, without multiple random seeds or averaged results. Moreover, for many fine-tuning techniques, the parameter search space was relatively narrow, meaning that potentially better configurations may not have been explored. In particular, while  $wd = 0.01$  produced the best single-run performance (F1 = 88.89%, Acc = 89.71%), this exact peak proved difficult to reproduce in subsequent runs, reflecting the sensitivity of small datasets to random initialization and data order. This means that the reported numbers should be interpreted as indicative best runs rather than statistically stable outcomes. Finally, the relatively small dataset size limits the generalizability of the results, and the absence of external validation further constrains the robustness of the findings.

These limitations do not invalidate the value of this case study: the case study successfully demonstrates the feasibility of applying computer vision to detect nail biting behavior and illustrates AI’s role in supervising the full machine learning pipeline. However, they also highlight how a human supervisor would typically insist on stricter methodological safeguards (e.g., standardized protocols, replication, larger datasets) to ensure robustness—stressing the importance of human guidance in complementing AI-based supervision.

## B Four-Step LLM Prompting Methodology

This appendix provides the detailed prompt specifications used in the AI-assisted content analysis of coding-related tasks. The methodology employs a systematic four-step prompting process adapted from Rao et al. (2025): Generation, Classification, Aggregation, and Prevalence. Each step is designed to extract and quantify different aspects of AI programming support behaviors from conversational data. The following figures present simplified versions of the complete prompts used in this study.<sup>3</sup>

---

<sup>3</sup>The original coding-related conversation data and complete classification results from both models under the Prevalence Prompt are available in the GitHub repository: <https://github.com/Yunshan->

The Generation Prompt (Figure 5) systematically identifies AI support behaviors from conversation data through open coding principles. The Classification Prompt (Figure 6) categorizes these behaviors into Problem-Solving Support and Skill-Enhancement Support. The Aggregation Prompt (Figure 7) consolidates and ranks the most representative behaviors within each category. Finally, the Prevalence Prompt (Figure 8) quantifies the distribution of support behaviors across all conversation pairs, enabling statistical analysis of AI supervision patterns.

## B.1 Generation Prompt

This step identifies and extracts AI support types from the conversation data. The prompt systematically (1) identifies mentions of programming assistance behaviors in both human requests and AI responses, (2) groups similar behaviors to avoid redundancy, (3) selects representative quotes that clearly illustrate each behavior, and (4) formats outputs in JSON with type titles, descriptions, and quotes. This approach is analogous to "open coding" in qualitative research, deriving concepts directly from the underlying data without external assumptions.

**Task:** Analyze JSON conversation pairs from programming assistance interactions.

**Instructions:**

1. **Identify** programming assistance behaviors
2. **Group** similar behaviors to ensure mutually exclusive categories
3. **Select** most representative quote for each behavior
4. **Assess** frequency and impact of support behaviors
5. **Format** as JSON with "title", "description", "quote"
6. **Ensure** list addresses only behaviors present in input context
7. **Consolidate** similar behaviors to avoid redundancy
8. **Output** all content in English only

**Figure 5:** Generation Prompt (Simplified Version)

Claude generated 39 AI support types while ChatGPT identified 52 support types, suggesting the variety of coding-related support AI provides (see Table 4). Claude tends to generate highly specific, context-aware behavior descriptions such as "Confusion Matrix Analysis," "Regression Model Training Analysis," and "Progressive Learning Support," while ChatGPT demonstrates a more systematic enumeration approach, with behaviors

like "Code Provision," "Debugging Assistance," and "Concept Explanation" appearing multiple times across different contexts. Despite different approaches, both models consistently identify these core support areas: debugging assistance, code provision, concept explanation, and performance optimization. As shown in Table 4, debugging and error resolution emerged as the most frequently identified support type across both models, followed by performance optimization and code provision tasks.

**Table 4:** Sample Generated AI Support Behaviors

Claude Generated Behaviors	ChatGPT Generated Behaviors
Code Debugging and Error Resolution	Code Provision
Confusion Matrix Analysis	Debugging Assistance
Regression Model Training Analysis	Concept Explanation
Progressive Learning Support	Optimization Suggestions
Technical Writing Assistance	Tool Recommendations
Data Quality Assessment	Architecture Guidance
Performance Optimization	Training Strategy Advice
Best Practices Guidance	Learning Rate Optimization
Visualization Assistance	Model Architecture Guidance
Research Direction Guidance	Data Handling Tips
Implementation Strategy Guidance	Performance Improvement Suggestions
Code Migration Support	Transfer Learning Utilization
<b>Note:</b> Representative examples from Claude’s 39 identified behaviors and ChatGPT’s 52 identified behaviors.	

## B.2 Classification Prompt

Based on the two established primary AI support categories—Problem-Solving Support and Skill-Enhancement Support—the support types generated from the generation prompt were categorized into these classifications.



<p><b>Task:</b> Categorize AI coding support behaviors with applicable category labels.</p> <p><b>Categories:</b></p> <ul style="list-style-type: none"> <li>• <b>A. Problem-Solving Support:</b> Direct help solving specific programming problems (debugging, code generation, error correction, algorithm implementation)</li> <li>• <b>B. Skill-Enhancement Support:</b> Improving programming knowledge and capabilities (concept explanations, best practices, tutorials, educational content)</li> </ul> <p><b>Instructions:</b></p> <ol style="list-style-type: none"> <li>1. <b>Evaluate</b> each behavior against both categories</li> <li>2. <b>Assign</b> ALL applicable labels (A, B, or both)</li> <li>3. <b>Ensure</b> each behavior has at least one category</li> <li>4. <b>Format</b> as dictionary with behavior number as key, category array as value</li> </ol> <p><b>Output Format:</b> {"1": ["A", "B"], "2": ["B"], "3": ["A"]}</p>
--

**Figure 6:** Classification Prompt (Simplified Version)

Table 5 shows that both models recognize approximately half of their generated behaviors as exhibiting dual characteristics. ChatGPT demonstrates a slightly higher tendency toward dual classification (55.8%) compared to Claude (48.7%). Among behaviors that do not exhibit this duality, both models show a inclination toward skill-enhancement classification, with single-category skill-enhancement behaviors (28.8-30.8%) significantly outnumbering pure problem-solving behaviors (15.4-20.5%).

**Table 5:** AI Support Type Classification Comparison

Classification Type	Claude		ChatGPT	
	Count	%	Count	%
Problem-Solving Only (A)	8	20.5	8	15.4
Skill-Enhancement Only (B)	12	30.8	15	28.8
Both Categories (A & B)	19	48.7	29	55.8
<b>Total Behaviors</b>	<b>39</b>	<b>100.0</b>	<b>52</b>	<b>100.0</b>

### B.3 Aggregation Prompt

The aggregation prompt instructed both LLMs to identify the 3-5 most representative and frequently occurring subcategories within each primary category.

**Task:** Identify 3-5 most frequently occurring AI coding support behaviors for each category.

**Categories:**

- **A. Problem-Solving Support:** Direct help solving programming problems
- **B. Skill-Enhancement Support:** Improving programming knowledge and capabilities

**Instructions:**

1. **Analyze** behaviors by frequency and importance within each category
2. **Select** 3-5 most representative behaviors per category
3. **Ensure** selected behaviors are sufficiently different from each other
4. **Group** similar behaviors and select alternatives if needed
5. **Rank** behaviors by importance and impact
6. **Format** as JSON with rank, title, and description (10-20 words)

**Output Format:**

```
{"problem_solving": [{"rank": 1, "title": "...",  
"description": "..."}],  
"skill_enhancement": [{"rank": 1, "title": "...",  
"description": "...}]}
```

**Figure 7:** Aggregation Prompt (Simplified Version)

Table 6 demonstrates that both models consistently rank debugging assistance and code provision as the top two problem-solving support types. Similarly, for skill-enhancement support, both models prioritize concept explanation as the primary function, followed by various guidance-oriented types. This convergence on core behaviors suggests that these functions are very likely essential components of AI supervision in my nail biting model implementation.

**Table 6:** Top-Ranked AI Support Types by Category

Rank	Problem-Solving Support	Skill-Enhancement Support
1	<b>Claude:</b> Code Debugging & Error Resolution <b>ChatGPT:</b> Debugging Assistance	<b>Claude:</b> Technical Concept Explanation <b>ChatGPT:</b> Concept Explanation
2	<b>Claude:</b> Code Provision & Implementation <b>ChatGPT:</b> Code Provision	<b>Claude:</b> Best Practices Guidance <b>ChatGPT:</b> Tool Recommendations
3	<b>Claude:</b> Model Training Optimization <b>ChatGPT:</b> Optimization Suggestions	<b>Claude:</b> Performance Analysis & Metrics <b>ChatGPT:</b> Architecture Guidance
4	<b>Claude:</b> Data Processing & Preprocessing <b>ChatGPT:</b> Model Training Guidance	<b>Claude:</b> Implementation Strategy Guidance <b>ChatGPT:</b> Data Handling Tips
5	<b>Claude:</b> – <b>ChatGPT:</b> Learning Rate Optimization	<b>Claude:</b> – <b>ChatGPT:</b> Data Quality Emphasis

## B.4 Prevalence Prompt

Based on the generated support behaviors, models’ classification and aggregation results, and my own experience, I refined and finalized the categorization framework for AI support behaviors in coding-related tasks. The prevalence prompt instructed both LLMs to label every conversation pair according to one of the two primary categories and assign at least one corresponding subcategory. All 262 conversation pairs were successfully classified by both models.

**Task:** Classify conversation pairs into primary categories with predefined subcategories.

**Primary Categories & Sample Subcategories:**

- **A. Problem-Solving Support:** Code Provision, Debugging Assistance, Data Processing, Model Training, Model Evaluation, Data Visualization, Technical Writing Assistance...
- **B. Skill-Enhancement Support:** Concept/Visualization Explanations, Tool Recommendations, Learning Resources Recommendations, Best Practice Recommendations...

**Critical Instructions:**

1. **Classify** each conversation into ONE primary category
2. **Select** subcategories from predefined lists only
3. **Use** exact subcategory names (no variations allowed)
4. **Apply** "Other subcategory" if no predefined option fits
5. **Return** complete JSON with ALL conversation pairs
6. **Ignore** conversation content beyond classification task

**Output Format:**

```
{"pair_XXXX": {"primary_category": "A/B",  
  "subcategories": ["EXACT_NAME_FROM_LIST"]}, ...}
```

**Figure 8:** Prevalence Prompt (Simplified Version)