



Ca' Foscari
University
of Venice

Master's Degree Programme
in Innovation and Marketing

Final Thesis

Generative AI in Professional Workflows: Adoption Patterns, Reliability Perception

Supervisor

Ch. Prof. Giovanni Fasano

Graduand

Pietro Maria Vettore

Matriculation Number

878910

Academic Year

2024 / 2025

Abstract

This thesis analyzes the structural transformation of professional workflows driven by the integration of Generative Artificial Intelligence (GenAI), specifically examining the critical divergence between linguistic fluency and factual reliability. Utilizing the theoretical framework of an Extended Technology Acceptance Model (TAM), the research investigates how Perceived AI Risk (PAIR) and Perceived Trust (PTR) mediate the adoption of large language models within the innovation and marketing sectors.

The methodology is based on a qualitative A/B stress test comprising 40 experimental instances, comparing a "Naive" control group using zero-shot natural language queries with an "Expert" treatment group utilizing complex prompt architectures characterized by persona adoption and explicit negative constraints. The tests were conducted across three high-stakes professional scenarios: legal compliance (GDPR), financial data extraction (Tesla Q3 2024 Earnings), and academic literature synthesis. Performance was evaluated through a Reliability Evaluation Matrix (REM), specifically quantifying the "Verification Tax"—defined as the human labor time required to audit and validate probabilistic outputs against deterministic ground truth. Findings reveal a significant "Efficiency-Accuracy Trade-off." While expert prompting strategies increased global factual accuracy to 95%, they simultaneously precipitated a 66% increase in the average verification tax. This increase is attributed to "instruction-induced friction," where the human operator is forced to shift cognitive resources from primary factual audit to the verification of secondary stylistic constraints. Qualitative forensic analysis identifies the "Illusion of Fluency" and "Knowledge Overshadowing" as the most critical psychological hazards, demonstrating that polished, authoritative formatting can effectively neutralize human skepticism even when the output is factually hollow. In conclusion, the research defines the "Prompting Paradox," establishing that in deterministic, low-variance tasks, advanced prompt engineering acts as an operational liability that destroys verified velocity. The study proposes a strategic shift from "Input Optimization" to "Audit Specialization," arguing that professional competence in the generative economy resides in the strategic recognition of risk. Ultimately, the human role is redefined from a primary content creator to the final guarantor of institutional trust.

Keywords:

Generative Artificial Intelligence, Large Language Models (LLMs), AI Hallucinations, Prompt Engineering, Human-in-the-Loop (HITL), Verification Tax, Ground Truth

Index

Index	3
CHAPTER I: Introduction	1
1.1.2 Distribution of Content: The Signal-to-Noise Crisis.....	5
1.1.3 Go to Market Strategy: The Tension Between Speed and Risk.....	7
1.2 The Reliability Gap:.....	8
1.3 Research Foundations and Academic Perimeter.....	10
1.3.1 Research Questions (RQs).....	11
CHAPTER II: THEORETICAL FRAMEWORK	12
2.1 The Technology Acceptance Model (TAM) for GenAI: Risk as the Barrier to Adoption.....	12
2.1.1 The Role of Perceived AI Risk (PAIR) and Trust.....	15
2.2 Defining "Hallucination" in Artificial Intelligence.....	16
2.2.1 Taxonomy of Error: Intrinsic vs. Extrinsic.....	17
2.2.2 Factuality vs. Faithfulness in Professional Workflows.....	18
2.2.3 The "Bias Towards Action": Why AI Cannot Remain Silent.....	18
2.2.4 The Alignment Tax: Safety Refusals and Over-Constraint.....	19
2.3 The Necessity of Human-Centric Evaluation.....	20
CHAPTER 3: METHODOLOGY	22
3.1 Research Design: Ground Truth Verification.....	22
3.1.2 The Variables: Group A and Group B.....	23
3.2 Dataset Construction: Three High-Stakes Scenarios.....	24
3.3 The Reliability Evaluation Matrix.....	26
3.4 Experimental Procedure and Analytical Framework.....	28
3.4.2 The "Length Factor" Control.....	29
3.4.3 Statistical Analysis Plan.....	29
CHAPTER IV - RESULTS	30
4.1. Experimental Setup and Task Taxonomy.....	30

4.2 Quantitative Analysis.....	33
CHAPTER V: CONCLUSIONS.....	43
Appendix A: Experimental Performance Scorecard.....	45
Appendix B - AI Prompts and Outputs.....	49
REFERENCES.....	59

CHAPTER I: Introduction

The traditional digital economy was predicated on the scarcity of high-fidelity generation. Historically, the production of professional assets—ranging from strategic marketing plans to complex financial reports—demanded significant capital investment in skilled human labor and temporal resources. The transition toward Generative Artificial Intelligence (GenAI) represents a fundamental discontinuity in this trajectory, not by simply automating repetitive tasks, but by commoditizing high-level cognitive and creative output. As the marginal cost of content generation approaches zero, economic value is migrating from the act of production to the scarcity of verification and trust.

This paradigm shift introduces a unique set of operational risks that distinguish GenAI from deterministic software. Unlike traditional computational tools that operate on logic-based rules where specific inputs yield predictable outputs, GenAI models are probabilistic engines designed to predict sequences rather than verify truth. As established by the "Inevitability Theorem," hallucination is not a temporary technical defect but a mathematical certainty inherent to computable large language models. Consequently, professional adoption is no longer a matter of measuring raw computational capability, but of managing what this thesis defines as the "Reliability Gap": the discrepancy between the linguistic fluency of an output and its factual or logical accuracy.

Within contemporary workflows, the human professional is no longer the sole originator of raw material but has been elevated to the role of a "Senior Manager" overseeing a "Junior Colleague". While the AI assistant can process vast datasets and generate drafts at unprecedented speeds, it lacks strategic comprehension and corporate liability. This delegation introduces a hidden operational friction—the "Verification Tax"—defined as the human capital required to audit, fact-check, and align synthetic outputs with institutional standards. If the time required to verify the output approaches the time required to create it manually, the promised productivity gains of the generative revolution become illusory.

The objective of this research is to move beyond the "Illusion of Progress" often associated with high-velocity AI generation. Utilizing an Extended Technology Acceptance Model (TAM), the study investigates how user expertise, or "Prompting Competence," acts as a decisive variable in mitigating risk and calibrating trust. Through a qualitative stress test across high-stakes legal, financial, and academic scenarios, this thesis quantifies the

efficiency-accuracy trade-off. It aims to establish that modern professional competence in the generative economy resides not in the ability to generate the most content, but in the strategic recognition of when to prioritize safety constraints over speed. Ultimately, this research seeks to redefine the human role as the ultimate guarantor of institutional trust in a landscape of synthetic abundance.

GenAI's integration into creative industries has especially precipitated significant consumer resistance and operational friction, fundamentally altering the reception of creative artifacts. Beyond consumer sentiment, this technological shift has triggered severe institutional and legal backlash regarding intellectual property and platform governance. The ambiguity of rights ownership, the unverified legitimacy of training data sources, and the lack of clear boundaries regarding artistic originality pose massive ethical and legal dilemmas for the music economy. This shift introduces a unique set of operational risks that distinguish GenAI from traditional software. Traditional computational tools operate on logic-based rules where a specific input consistently yields a predictable output. In contrast, GenAI models are probabilistic engines designed to predict the next likely token rather than to verify truth. As established by Cossio (2025) through the "Inevitability Theorem," hallucination is not a temporary defect but an innate limitation of computable Large Language Models (LLMs). Cossio (2025) formally proves that for any computable model, there exists a ground truth function where the model must fail to match the correct output. Therefore, the global economy is transitioning into a "probabilistic workflow" era. In this new reality, the efficiency gains promised by automation must be constantly weighed against the operational costs of error detection and risk mitigation. The narrative that AI will simply replace human workers is reductionist. Instead, as Amankwah-Amoah et al. (2024) argue, the technology acts as a catalyst for role transformation. The human professional is no longer the sole originator of raw material but becomes an orchestrator or curator of algorithmic outputs. The economic winners in this new paradigm will not be those who generate the most content, but those who can most effectively filter, verify, and humanize the machine's output to establish trust with the consumer.

This paradigm shift necessitates a fundamental re-evaluation of human-computer interaction. In contemporary workflows, GenAI is no longer utilized as a passive software application similar to a traditional word processor or spreadsheet; instead, it functions as a "Junior Colleague." Comparable to an eager, highly productive, yet inexperienced

entry-level assistant, the AI can process vast amounts of data and generate high-fidelity drafts at unprecedented speeds. However, it lacks true strategic comprehension, contextual nuance, and corporate liability. When professionals delegate tasks to this "Junior Colleague," the resulting output is often linguistically fluent but factually precarious. Because the AI cannot be held legally or reputationally accountable for its hallucinations, human workers must elevate their roles to that of Senior Managers. The cognitive energy previously expended on raw creation is now reallocated to supervisory functions: auditing, fact-checking, and aligning synthetic output with strict corporate standards. This supervisory burden introduces a hidden operational friction—a "Verification Tax"—where the velocity of the machine is inherently bottlenecked by the necessity of human curation. To understand the magnitude of this shift, it is necessary to map the rapid evolution of the underlying technology. The industry has transitioned from predictive discriminative models, designed to classify data, to generative transformer-based architectures designed to synthesize new data distributions. The progression from early iterations such as GPT-3 to the multimodal capabilities of GPT-4 and the visual fidelity of DALL-E represents a significant advancement in semantic understanding and combinational logic.

This technological maturation has precipitated what scholars describe as the "Democratization of Creativity." Wang et al. (2023) highlight that advanced generative models, particularly text-to-image systems like DALL-E, exhibit a level of "Combinational Creativity" that was previously the exclusive domain of skilled human artists. By enabling users to synthesize complex visual and textual artifacts through natural language prompts, these tools have decoupled creative output from technical manual skill. A marketing manager with no background in illustration can now generate a hyper-realistic prototype of a product launch simply by articulating the concept.

However, this democratization introduces complex dynamics regarding quality and originality. Wang et al. (2023) note that while these models can simulate creativity by combining disparate concepts (e.g., "an armchair in the shape of an avocado"), they operate within the bounded latent space of their training data.

This development raises critical questions regarding the nature of the output: does it constitute genuine innovation, or merely high-fidelity mimicry? The industry landscape is now characterized by a surplus of synthetic content. As the marginal cost of producing "average" creative work approaches zero, the market becomes saturated with plausible but

often generic assets. For professional users, this abundance introduces a new challenge: differentiating brand voice in an ecosystem where competitors have access to the same generative engines. Competitive advantage, therefore, is derived not from access to the tool itself, but from the sophisticated application of "Prompt Engineering" to move the model beyond its statistical mean and into the realm of genuine novelty. The most tangible impact of GenAI on the daily operations of marketing professionals is the obsolescence of the traditional "Linear Workflow." To operationalize this shift, it is necessary to contrast the established Fordist production model with the emerging recursive "Circular Workflow."

Historically, creative and strategic workflows followed a sequential, waterfall-like trajectory:

Ideation → Drafting → Editing → Finalization.

In this paradigm, the "cost of iteration" was high. If a creative director decided to fundamentally alter a campaign concept during the drafting phase, it necessitated a restart of labor-intensive manual execution. Consequently, the operational logic of marketing firms was predicated on minimizing iteration. Great effort was expended in the "Brainstorming" phase to ensure the brief was perfect, thereby protecting the expensive labor of the "Drafting" phase.

A critical financial and operational consequence of this model is the escalating cost of changes. As a project progresses through the review and finalization stages, any required modifications necessitate a complete or substantial manual rework of the deliverables, leading to prohibitively expensive changes, especially once the final assets are published or launched. This high-cost-of-iteration model is consistently observed across domains such as content generation, technical documentation, and data analysis reporting.

GenAI invalidates this economic logic; since the "Drafting" phase is now instantaneous and effectively free, the workflow transforms into a high-velocity Circular Loop:

Prompt → Generate → Verify → Refine.

This transition mirrors the evolution observed in modern music production. Deruty and Grachten (2022) describe the shift from "linear composition" (writing a score, then recording it) to "in-studio composition," in which composing, sound design, and mixing occur simultaneously in a nonlinear feedback loop.

Transposing this to marketing, the professional does not write a press release and then edit it. They "sculpt" the text through iterative prompting. They might generate twenty variations of a slogan in thirty seconds, select the most promising candidate, modify the constraints (e.g., "make it more empathetic"), and regenerate.

Central to this new workflow is the mechanism of "Combinational Creativity."

Wang et al. (2023) define this as the ability to merge a Base concept (the primary subject) with an additive concept (a modifier or style). In a linear workflow, a designer would have to consciously decide to combine "minimalist design" (Base) with "baroque textures" (Additive) and then laboriously execute it. In the circular workflow, the operator creates these combinations via prompts to rapidly test hypotheses.

- Prompt: "Write a crisis apology email (Base) using the tone of a supportive peer (Additive)."
- Verification: The user assesses if the model successfully integrated the "Additive" constraint or if it produced a generic corporate apology.

1.1.2 Distribution of Content: The Signal-to-Noise Crisis

The integration of GenAI into the creative supply chain has precipitated a radical shift in the economics of content distribution. The demanding nature of generating high-quality marketing copy, technical documentation, or visual assets, which historically necessitated considerable human labor, served as an inherent limiting factor. In the pre-AI era, the primary constraint on distribution was production costs. Generating high-quality marketing copy, technical documentation, or visual assets required significant human labor, acting as a natural filter that limited the volume of content entering the market. Today, however, we are witnessing the complete removal of this production bottleneck. As noted by Amankwah-Amoah et al. (2024), GenAI facilitates the immediate generation of content at scale, effectively driving the marginal cost of creating a new digital asset to near zero.

This unrestricted production capacity has resulted in a "Signal-to-Noise" crisis within distribution channels. Organizations are no longer competing for visibility in a scarce

market; they are competing for attention in an ecosystem flooded with synthetic abundance. Because GenAI models function as probabilistic completion engines rather than truth-seeking entities, they exhibit a "Bias Towards Action" (Cossio, 2025). When tasked with filling a distribution pipeline—such as generating thousands of personalized emails or social media posts—the model will not hesitate to generate content even when it lacks sufficient information. It fills the void with plausible-sounding but potentially erroneous text. Consequently, the distribution landscape is at risk of being polluted by what we term "Hallucination Spam." This refers to the mass dissemination of content that is grammatically fluent but factually untethered. As demonstrated by Janiak et al. (2025), standard automated evaluation metrics like ROUGE fail to detect these errors because they measure lexical overlap rather than factual consistency.

A marketing email might receive a high ROUGE score because it uses the correct keywords, yet still contain a "Parisian Tiger" hallucination—a completely fabricated fact that looks real (Cossio, 2025). For the Innovation Manager, this creates a dangerous operational reality where the distribution channel becomes a vector for reputational risk. The challenge is no longer how to distribute content, but how to prevent the distribution of automated misinformation.

Reliability as the Differentiation Strategy In this environment of infinite content, "Reliability" becomes the primary differentiation strategy. Zhou et al. (2025) emphasize that Perceived Trust (PTR) is a critical determinant of technology acceptance. If a consumer cannot distinguish between a verified human insight and a stochastic AI fabrication, their trust in the brand diminishes. Therefore, companies must pivot their distribution strategy from "Maximum Velocity" to "Maximum Fidelity." The competitive advantage belongs to firms that can verify their content before it enters the distribution stream, ensuring that their signal stands out against the background noise of unverified synthetic media.

1.1.3 Go to Market Strategy: The Tension Between Speed and Risk

As organizations transition from traditional manual execution to AI-assisted workflows, the primary driver for executive buy-in is rarely limited to internal cost reduction; rather, it is the pursuit of unprecedented market agility. In the realm of Innovation and Marketing, the ultimate testing ground for this agility is the execution of commercial initiatives. The

promise of Generative AI is highly tantalizing for modern enterprises: to compress what used to be months of strategic planning, content ideation, and campaign deployment into a matter of days or even hours. However, this aggressive acceleration fundamentally alters the operational risk profile of the organization. If the integration of GenAI provides workers with a high-speed "Junior Colleague," it simultaneously dismantles the historical friction that naturally prevented errors from reaching the public domain. To fully grasp the managerial impact of this technology, it is crucial to examine the macro-level consequences of deploying AI-generated output at scale.

The strategic imperative to adopt GenAI is frequently framed in terms of "Time-to-Market." In the hyper-competitive digital economy, the ability to launch a product, campaign, or service faster than competitors is often viewed as the decisive factor for success. GenAI promises to collapse the development cycle, enabling "immediacy" in prototyping and execution. However, this pursuit of speed introduces a profound strategic tension: the inverse relationship between Velocity and Risk Control.

Traditional Go-to-Market (GTM) workflows included built-in latency periods—drafting, reviewing, revising—that served as natural safety buffers. The "immediacy" of GenAI removes these buffers. A marketing team can now generate a campaign strategy and execute it in hours rather than weeks. While this acceleration is the primary value proposition of the technology, it also accelerates the dissemination of error. According to the Inevitability Theorem proposed by Cossio (2025), hallucination is an innate limitation of computable Large Language Models (LLMs). It is not a bug that can be eliminated; it is a mathematical certainty. Therefore, any GTM strategy that relies on the raw speed of GenAI without a corresponding increase in verification protocols is statistically guaranteed to deploy errors into the market.

The risk of unchecked autonomy is exacerbated by the "Illusion of Competence." Because LLMs generate text with high linguistic fluency, they mask their lack of reasoning capability (Janiak et al., 2025). An Innovation Manager might read a coherent, confident executive summary generated by GPT-4 and assume it is factually accurate, authorizing a rapid GTM launch based on flawed data. This creates a "Safety Paradox" where the tools that make us faster also make us more vulnerable. Zhou et al. (2025) identify Perceived AI Risk (PAIR) as a significant barrier to adoption, noting that users are acutely aware of the potential for ethical and operational failure.

Success cannot be measured solely by the reduction in launch time. It must be measured by "Verified Velocity"—the speed at which *accurate* and *safe* content can be deployed. This requires the integration of what Cossio (2025) describes as "Guardrails" and "Symbolic Integration" into the GTM workflow. The necessity for "Verified Velocity" is further corroborated by recent analyses of enterprise LLM deployments. As noted by Ernst & Young (EY, 2025) in their assessment of hallucination risks in high-stakes domains, the unverified deployment of GenAI can severely compromise advisory deliverables, expose organizations to regulatory scrutiny, and erode internal confidence. Addressing hallucination risk is no longer considered optional but a "strategic imperative" for delivering compliant and trusted AI services (EY, 2025). Therefore, the manager's role is to act as the "brake" in a system designed for speed, applying critical scrutiny to the machine's output to prevent the high-velocity distribution of automated misinformation.

1.2 The Reliability Gap:

Having established in the previous section that the professional value chain has shifted from manual execution to strategic oversight, we must now define the central operational conflict of this new paradigm. While Generative AI (GenAI) solves the problem of production scarcity, it introduces a novel and pervasive problem: the Reliability Gap.

This gap is defined as the discrepancy between the linguistic fluency of the model's output and its factual or logical accuracy. In traditional software environments, such as Excel or SQL databases, the system is deterministic; if the input logic is correct, the output is reliable. In the GenAI environment, the system is probabilistic and autoregressive. It does not retrieve facts; it predicts the next statistically likely token (Cossio, 2025). This architectural difference creates a workflow where the primary risk is no longer the inability to create, but the inability to distinguish between a valid insight and a plausible fabrication. The prevailing narrative suggests that AI functions as an "Automation Tool," replacing human labor and saving time. A manager might assume that if an AI can write a report in one minute that previously took sixty minutes, the firm has saved fifty-nine minutes of labor. However, this calculation is critically flawed because it ignores the stochastic nature of the model. As Janiak et al. (2025) demonstrate in their study "The Illusion of Progress,"

current evaluation metrics like ROUGE are insufficient for detecting factual errors. A text can be highly fluent, achieving high scores on automated metrics, yet remain factually incorrect. Consequently, the human operator cannot simply accept the AI's output. They must audit it. This auditing process (checking facts, verifying citations, ensuring brand tonal alignment) constitutes the Verification Tax.

- **Traditional Workflow:** Time spent Writing (High) + Time spent Editing (Low).
- **GenAI Workflow:** Time spent Prompting (Low) + Time spent Verifying (High).

If the time required to verify the output approaches or exceeds the time required to create it manually, the "productivity gain" becomes illusory. The user is essentially paying a tax on every interaction with the model to insure against the risk of hallucination (Cossio, 2025). The danger of this new workflow lies in what Janiak et al. (2025) call the "Illusion of Progress." Because the AI generates content instantly, the user experiences a false sense of completion. The screen is filled with text; the code runs; the image looks polished. Psychologically, the task feels "done." However, this visual completeness masks potential "Intrinsic Hallucinations" (contradictions within the text) or "Extrinsic Hallucinations" (fabrications of reality) (Cossio, 2025). This illusion creates a widening "Competence Gap" between naive and expert users.

- The Naive User (Group A): Falls victim to the illusion. They employ basic prompts, treat the model's fluency as a proxy for accuracy, and pay a low Verification Tax initially, only to incur high costs later when errors are discovered in the market. This group experiences high "Perceived AI Risk" (Zhou et al., 2025) because they do not understand why the model fails.
- The Expert User (Group B): Recognizes the inevitability of error. They employ "Advanced Prompting" techniques—such as Chain-of-Thought reasoning and negative constraints—to proactively reduce the model's search space. Crucially, they pay a high Verification Tax upfront, scrutinizing the output to ensure it meets the strict standards of "Content Quality" (Zhou et al., 2025).

1.3 Research Foundations and Academic Perimeter

Traditional innovation management literature has focused on optimizing deterministic processes. Whether managing a physical supply chain or a software development lifecycle, the underlying assumption was that specific inputs would yield predictable, verifiable outputs. Innovation was defined as manipulating tangible variables to reduce variance and increase efficiency.

This thesis shifts the locus of analysis to a fundamentally different domain: Reliability in Creative Prompting.

In the GenAI ecosystem, the "product" is no longer a physical object with stable properties but a stochastic, semantically interpretable output. It may take the form of a vibe-coded application, a marketing campaign, or a video.

The "manufacturing process" is no longer a linear assembly line but a probabilistic interaction between a human operator and a Large Language Model (LLM).

Consequently, the "defects" we must analyze are not material flaws but cognitive aberrations such as hallucinations, bias, and logical inconsistency.

Therefore, the central problem addressed by this research is not how to generate content, which GenAI has commoditized, but how to manage the risk associated with that generation. We posit that the "Reliability Gap" represents the digital equivalent of a broken quality assurance mechanism. Just as a factory manager must minimize physical waste, the AI utilizer must now minimize "cognitive waste," defined as the time and energy lost to verifying, correcting, and discarding the model's hallucinations.

The theoretical tension driving this research lies in the conflict between the inherent limitations of the technology and the psychological requirements for professional adoption.

We frame this tension using two distinct theoretical lenses. The Inevitability of Error On the technological side, we ground our analysis in the Inevitability Theorem proposed by Cossio (2025). This framework provides a formal proof that for any computable LLM, hallucination is not an occasional bug but a mathematical certainty. Cossio (2025) demonstrates through Theorem 3 that for any computable model, there exists a ground truth function where the model must fail to match the correct output. This establishes that risk is structural to the tool; it cannot be engineered away by the model developers alone. On the behavioral side, we utilize the Extended Technology Acceptance Model (TAM)

proposed by Zhou et al. (2025). This model identifies Perceived AI Risk (PAIR) and Perceived Trust (PTR) as the decisive variables for adoption. Zhou et al. (2025) argue that regardless of a model's raw capability or speed, if the user perceives a high risk of error, the *Intention to Use (IU)* collapses. The intersection of these two theories creates a critical managerial paradox.

If error is inevitable (Cossio), but trust is required for adoption (Zhou), how can GenAI ever be successfully integrated into professional workflows?

The scope of this thesis is to explore whether Human Competence can bridge this gap. We investigate whether "Prompt Engineering" and "Constraint Management" can act as risk mitigation strategies that render a structurally imperfect tool professionally viable.

1.3.1 Research Questions (RQs)

1. How has the execution of core marketing and innovation tasks changed compared to the pre-GenAI era, and what is the operational cost of the "Verification Tax"?
2. Does the user's "Prompting Competence" act as a decisive variable in reducing the rate of factual hallucinations and increasing the utility of the output?
3. How does "Perceived AI Risk" (PAIR) regarding hallucinations affect the professional's Intention to Use (IU), and can this risk be mitigated through structured workflows?

CHAPTER II: THEORETICAL FRAMEWORK

2.1 The Technology Acceptance Model (TAM) for GenAI: Risk as the Barrier to Adoption

To understand the integration of Generative AI (GenAI) into professional workflows, one must move beyond simple metrics of computational capability and examine the behavioral psychology of the user. The Technology Acceptance Model (TAM), originally proposed to

predict user adoption based on Perceived Usefulness (PU) and Perceived Ease of Use (PEOU), remains the foundational framework for information systems research. However, the introduction of generative agents necessitates a significant expansion of this model. Unlike deterministic software (e.g., Excel), where an input yields a predictable output, GenAI introduces uncertainty. Therefore, the adoption dynamic shifts from a calculation of utility to a calculation of risk.

The core of the classic TAM posits that Actual System Use is primarily determined by Behavioral Intention (BI), which is itself driven by two key cognitive beliefs: Perceived Usefulness (PU) and Perceived Ease of Use (PEOU). Perceived Usefulness (PU) is the degree to which a person believes that using a particular system will enhance their job performance or productivity. In a pre-GenAI context, this meant the system was a superior substitute for manual labor (e.g., a spreadsheet is more useful than a ledger). Perceived Ease of Use (PEOU) is the degree to which a person believes that using the system will be free of effort. A system that is intuitive, fast, and requires minimal cognitive load has high PEOU. The traditional model establishes a critical path in which PEOU directly influences PU (an easy-to-use tool is often seen as useful), and both directly influence Intention to Use.

However, the introduction of generative, non-deterministic agents necessitates a significant expansion of this classic model, leading to the Extended Technology Acceptance Model (TAM) for GenAI proposed by Zhou et al. (2025).

Traditional software is deterministic; the output is predictable, and error margins are quantifiable. GenAI is stochastic; an input yields a probabilistic, non-verifiable output often characterized by fluent, yet factually untethered, content known as hallucination.

This aligns with the seminal framework proposed by Bender et al. (2021), who characterize these models as 'stochastic parrots'. They argue that LLMs merely stitch together linguistic patterns observed in training data without referencing any underlying communicative intent or model of the world, thereby making truth a statistical accident rather than a design feature.

Therefore, the adoption dynamic shifts fundamentally from a calculation of utility to a calculation of risk. The extension by Zhou et al. (2025) integrates two critical external variables that directly interact with the core TAM constructs: Perceived AI Risk (PAIR)

and Perceived Trust (PTR). These variables are not mere antecedents; they are crucial mediators of the central PU and PEOU relationships.

Perceived AI Risk (PAIR) is the user's subjective assessment of potential negative consequences associated with deploying the AI's output in a professional context. This includes functional risks such as hallucinations, legal/ethical risks such as copyright concerns, and reputational risks. Critically, PAIR is found to negatively correlate with PEOU and Intention to Use. High risk makes the system psychologically and practically "difficult to use" because it imposes a high verification tax on the user.

Conversely, Perceived Trust (PTR) is defined as the user's belief that the system can consistently and safely deliver accurate and faithful results amidst technological uncertainty. PTR plays a vital positive role, as it positively correlates with both PU and PEOU. Trust reduces the cognitive friction associated with verification, thereby increasing the perception of Ease of Use, and ensuring that the high speed of generation translates into genuinely Usable output, thus enhancing Perceived Usefulness.

This expanded framework acknowledges that the human operator's decision to adopt GenAI is not a simple comparison of time saved, but a complex calculation where the fear of professional liability (PAIR) and the confidence in the output's fidelity (PTR) act as powerful psychological filters on the raw efficiency gains (PU and PEOU).

Building upon the Extended Technology Acceptance Model (TAM), recent empirical studies underscore that AI adoption in professional settings transcends individual utility calculations and is deeply embedded within organizational dynamics. Song et al. (2025) argue that traditional TAM constructs must be broadened to include top management support, which plays a pivotal role in legitimizing AI implementation and mitigating cultural resistance. Within this organizational context, the perceived value of the technology often eclipses mere operational simplicity. For instance, an empirical investigation of UK business companies by Durmaz et al. (2025) revealed that while perceived usefulness is the most influential predictor of ChatGPT adoption, perceived ease of use has a remarkably weak effect on user attitudes. This indicates a "value-over-simplicity mindset" among professionals (Durmaz et al., 2025), suggesting that innovation managers are willing to navigate complex systems—and effectively pay the associated "Verification Tax"—provided the tool delivers tangible strategic value and its adoption is backed by institutional trust.

However, this drive for utility is consistently challenged by the psychological and cognitive friction introduced by the technology's unreliability. The concept of Perceived AI Risk (PAIR) is inextricably linked to the cognitive load imposed on the professional during the verification process. As Aristawidya and Ramadhana (2026) articulate, while generative AI can reduce extraneous cognitive load by summarizing complex material, poorly contextualized or inconsistent outputs force users to expend significant mental effort to verify accuracy, thereby drastically increasing the overall cognitive burden. This dynamic is exacerbated in high-stakes business environments, where Joshi notes that 68% of business leaders report profound difficulty distinguishing between reliable outputs and hallucinations. When professionals lack transparent uncertainty quantification, they are forced to engage in extensive forensic auditing, translating the theoretical risk of hallucination into a tangible drain on human capital and severely diminishing the promised return on investment.

The persistence of these risks stems from the structural and architectural limitations of Large Language Models (LLMs), which remain opaque "black boxes" in corporate workflows (Song et al., 2025). Huang et al. (2025) provide a comprehensive taxonomy of these limitations, emphasizing that hallucinations are fundamentally driven by the autoregressive nature of the models, which prioritize statistical token prediction over factual accuracy. This design flaw makes them highly susceptible to both factuality hallucinations—where the model contradicts verifiable real-world knowledge—and faithfulness hallucinations, where the output diverges from the user's specific prompt or provided context (Huang et al., 2025). Because the internal logic of these models lacks causal understanding and relies on probabilistic averages, the Verification Tax cannot simply be engineered away through minor software updates. Consequently, organizations must actively manage this inherent vulnerability by institutionalizing rigorous human-in-the-loop oversight and constraint engineering, confirming that the reliable integration of generative AI relies entirely on human curation rather than autonomous machine execution.

2.1.1 The Role of Perceived AI Risk (PAIR) and Trust

Zhou et al. (2025) propose an extended TAM specifically tailored for GenAI music and creative applications, introducing the critical variable of Perceived AI Risk (PAIR). PAIR is defined as the user's subjective assessment of the potential negative consequences and uncertainties associated with deploying the technology. In the context of creative and marketing workflows, this risk is multifaceted. It includes legal uncertainties regarding copyright, ethical concerns about data provenance, and, crucially, the functional risk of generating uncontrollable or erroneous content (Zhou et al., 2025). The empirical findings by Zhou et al. (2025) reveal a decisive negative correlation between PAIR and the Intention to Use (IU). Specifically, if a professional perceives a high likelihood of error or liability, their willingness to adopt the tool declines, regardless of its theoretical speed or capability. More importantly, Zhou et al. (2025) identify that PAIR negatively impacts Perceived Ease of Use (PEOU). This is a vital managerial insight. A tool that hallucinates is not merely "imperfect". It is operationally "difficult" because it forces the user into a state of hyper-vigilance. The cognitive load required to verify every output, police the AI for copyright infringement, and correct subtle factual errors negates the perceived ease of the system. Recent research supports this shift in cognitive burden; Gerlich (2025) highlights that while AI enables "cognitive offloading" during the drafting phase, the subsequent requirement for critical evaluation imposes a dense, high-stakes cognitive load on the professional auditor. Furthermore, as Tonmoy et al. (2024) indicate in their comprehensive survey of LLM vulnerabilities, hallucination mitigation is fundamentally an ongoing interactive process rather than a solved technical feature. If the user must act as a constant supervisor to mitigate these inherent flaws (Tonmoy et al., 2024), the operational friction increases, thereby suppressing adoption (Zhou et al., 2025).

Conversely, Zhou et al. (2025) identify Perceived Trust (PTR) as a positive driver for both PEOU and PU. Trust in this context is defined as the user's belief that the system can reliably and safely assist in achieving goals amidst technological uncertainty. When a professional trusts the system's output consistency and data integrity, the "verification tax" is lowered. The user experiences a sense of "peace of mind," which enables a more intuitive interaction with the system, thereby enhancing the perception of its utility (Zhou et al., 2025). However, a critical distinction in the findings by Zhou et al. (2025) is that Trust (PTR) does not always directly correlate with Intention to Use (IU). Instead, it

mediates the relationship through Usefulness and Ease of Use. This implies that trust is a necessary precondition for perceiving value. A marketing manager will not deem a copywriting tool "useful" (PU) unless they first "trust" (PTR) that it will adhere to brand guidelines and factual reality. Consequently, the central bottleneck in the GenAI value chain is not generation speed but the reliability of the output. Until the "Hallucination Risk" is mitigated to a level where Trust can be established, the technology remains experimentally interesting but professionally non-viable for high-stakes tasks.

This dynamic is strongly influenced by the user's technological maturity. Li et al. (2024) demonstrate that within the TAM framework applied to generative models, "AI literacy" serves as a critical bridge between Perceived Ease of Use and Trust. Users with a higher operational understanding of how LLMs generate text are better equipped to anticipate errors, which paradoxically increases their overall trust in the workflow because they know how to apply effective verification strategies (Li et al., 2024). This confirms that user competence is the ultimate moderating variable in GenAI adoption.

2.2 Defining "Hallucination" in Artificial Intelligence

To operationalize the "risk" identified in the TAM analysis, we must rigorously define the technical failure mode known as "hallucination." In the medical field, hallucination refers to a sensory experience without external stimulus. In the context of Large Language Models (LLMs), Huang et al. (2025) and Cossio (2025) define hallucination as the generation of content that is plausible and coherent but factually incorrect, nonsensical, or unfaithful to the source content. It is critical to posit, as Cossio (2025) demonstrates through formal computability theory, that hallucination is not a temporary "bug" that will be eliminated by larger datasets. Cossio (2025) provides a theoretical framework proving that hallucination is inevitable for any computable LLM. Using diagonalization arguments, it is proven that for any computable model, there exists a ground truth function where the model must fail to match the correct output (Theorem 3). This "Inevitability Theorem" suggests that uncertainty is an innate property of the architecture. Therefore, from an innovation management perspective, the goal cannot be the total elimination of error, but

rather the precise categorization and mitigation of "Faithfulness" failures versus "Factuality" failures.

2.2.1 Taxonomy of Error: Intrinsic vs. Extrinsic

A robust taxonomy is required to distinguish between creative features and operational liabilities. Huang et al. (2025) and Cossio (2025) provide a bifurcation of hallucinations into two primary categories.

- **Intrinsic Hallucinations (Faithfulness Failures):** Intrinsic hallucinations occur when the generated output logically contradicts the provided input or context (Huang et al., 2025; Cossio, 2025). In a marketing or creative workflow, this is often described as Instruction Inconsistency or Context Inconsistency.
 - Example: A user provides a financial report and asks for a summary. The AI generates a summary that claims profits went down when the source text explicitly states profits went up.
 - Mechanism: The model fails to maintain consistency during inference or misinterprets the source document (Cossio, 2025).
 - Managerial Impact: This is a critical failure of "Faithfulness." In professional settings, this type of error undermines the tool's utility for tasks such as summarization, data analysis, or brand-aligned copywriting. If the AI cannot be faithful to the prompt, it cannot be trusted as a collaborative partner.
- **Extrinsic Hallucinations (Factuality Failures):** Extrinsic hallucinations refer to generated content that is not consistent with the training data or reality and cannot be verified by the input context (Huang et al., 2025; Cossio, 2025). The model introduces entities, facts, or events that do not exist.
 - Example: The model generates a biography for a real CEO but invents a fraudulent university degree, or claims "The Parisian Tiger was hunted to extinction in 1885" (Cossio, 2025).
 - Mechanism: This arises when the model attempts to bridge knowledge gaps or generalizes incorrectly from its training data (Cossio, 2025).
 - Managerial Impact: In factual tasks (e.g., generating a white paper), this is catastrophic and leads to reputational damage (Zhou et al., 2025). However, in

purely creative tasks (e.g., writing fiction), extrinsic hallucination is the mechanism of novelty. The prompt "Imagine a car shaped like an avocado" requires the model to hallucinate an object that does not exist.

2.2.2 Factuality vs. Faithfulness in Professional Workflows

The distinction between Factuality (absolute correctness against the world) and Faithfulness (adherence to the user's input) is paramount for this thesis. Huang et al. (2025) note that while factuality is objective, faithfulness is relative to the prompt. In the "In-Studio Composition" workflow described in Chapter 1, the primary risk is often a lack of Faithfulness. If a creative director prompts for a specific tone and the AI generates a different one, it is an Intrinsic Hallucination. If the AI invents a new melody, it is an Extrinsic Hallucination, which might actually be desirable. Therefore, evaluating GenAI in creative sectors requires measuring Faithfulness (did it do what I asked?) rather than just Factuality (is this true in the real world?).

2.2.3 The "Bias Towards Action": Why AI Cannot Remain Silent

A compounding factor in the hallucination problem is the model's inherent "Bias Towards Action." By design, commercial LLMs are trained to be highly responsive assistants. Consequently, they are heavily penalized during their Reinforcement Learning from Human Feedback (RLHF) phase for refusing to answer. As a result, when an AI encounters a query for which it lacks sufficient data, it rarely admits ignorance (e.g., stating "I do not know" or "Data not available"). Instead, compelled by its objective function to predict the next token, it invents a plausible-sounding answer to fill the silence. This inability to gracefully fail or request clarification is a critical vulnerability when integrating GenAI into automated business processes, as it guarantees that vague prompts will yield confident fabrications rather than requests for better instructions.

2.2.4 The Alignment Tax: Safety Refusals and Over-Constraint

To fully comprehend the operational limitations of Large Language Models within professional workflows, one must examine the behavioral guardrails embedded during

their training phases. Modern generative models are predominantly fine-tuned utilizing Reinforcement Learning from Human Feedback (RLHF), a methodology designed to align the model's stochastic outputs with human preferences, ensuring they remain simultaneously helpful and harmless. While this alignment process is fundamental to preventing the generation of toxic, unethical, or malicious content, it introduces systemic operational friction, often referred to in the literature as the "alignment tax."

A primary manifestation of this alignment tax in enterprise environments is the phenomenon of "over-refusal." As models are continuously optimized by human evaluators and reward models to avoid generating unsafe content, their internal safety filters can become excessively conservative. Ouyang et al. (2022) and broader alignment literature illustrate that when an operator utilizes advanced constraint engineering—such as imposing strict negative constraints, mandating complex structural frameworks, or establishing highly specific professional personas—the syntactic complexity and rigid boundaries of the input can inadvertently mimic the structure of an adversarial attack or a prompt injection attempt. Confronted with a prompt that heavily restricts its probabilistic freedom or demands rigorous, multi-layered logic, the model's safety architecture may misinterpret the user's technical precision as a malicious anomaly. In such instances, rather than navigating the constrained latent space to produce a highly verified and accurate output, the model defaults to a predefined, canned refusal, typically responding with variations of "I cannot fulfill this request" or "I am unable to assist with that."

This dynamic establishes a critical theoretical trade-off within the paradigm of prompt engineering and innovation management. "Competence Engineering," which we have defined as the expert application of complex constraints designed to reduce stochastic variance and mitigate the Verification Tax, simultaneously elevates the probability of triggering these rigid safety mechanisms. Because the expert user forces the model out of its default probabilistic pathways to ensure absolute faithfulness to a task, the system's defensive heuristics are more likely to activate. In stark contrast, a "Naive User" employing simple, unconstrained, zero-shot prompts is statistically less likely to encounter these safety blockades, as their queries align more closely with the conversational average the model expects. However, this lack of friction comes at the severe cost of receiving highly generalized, unverified, and potentially hallucinatory outputs. Consequently, the modern innovation manager must navigate a precarious equilibrium, constantly balancing

the architectural necessity for strict semantic control against the model's deeply ingrained, RLHF-driven propensity for defensive evasion.

2.3 The Necessity of Human-Centric Evaluation

If trust is the barrier to adoption (Section 2.1) and hallucination is the primary risk (Section 2.2), then the industry requires a reliable method to measure hallucination rates. However, the current standard for evaluation in Natural Language Processing (NLP) is fundamentally flawed when applied to creative and variable tasks. Janiak et al. (2025) argue that the reliance on automated metrics like ROUGE and BLEU creates an "illusion of progress" that obscures the true unreliability of these systems.

Metrics such as ROUGE (Recall-Oriented Understudy for Gisting Evaluation) and BLEU (Bilingual Evaluation Understudy) were designed for tasks like translation, where there is a clear reference text. They measure quality by calculating the n-gram (word sequence) overlap between the AI's output and a "gold standard" reference (Janiak et al., 2025). Janiak et al. (2025) demonstrate that this approach is intellectually lazy and operationally dangerous for Generative AI for the following reasons:

1. **Semantic Blindness:** ROUGE checks whether words match, not whether their meanings are correct. A sentence can be factually identical to the reference but use different vocabulary (synonyms), resulting in a low score. Conversely, a sentence can share many words with the reference but contain a single logical negation ("not") that flips the meaning, yet still receives a high ROUGE score (Janiak et al., 2025).
2. **The Length Bias:** Janiak et al. (2025) uncover a systematic bias in these metrics against verbosity. ROUGE consistently penalizes longer responses, even if the additional length provides necessary context or creative elaboration. In creative marketing or music composition, elaboration is often the goal. By penalizing length, these metrics inadvertently penalize creativity and reward concise but potentially shallow outputs.

The most damning evidence provided by Janiak is the divergence between ROUGE scores and human judgment. Through a comprehensive human evaluation study, Janiak et al. (2025) show that while ROUGE has high recall, its precision in detecting actual factual errors is alarmingly low (approximately 40%). This means that a significant portion of content flagged as "good" by automated metrics is actually hallucinated or unfaithful when reviewed by a human. Janiak et al. (2025) further show that simple heuristics, such as measuring response length, can sometimes rival or outperform sophisticated metrics like Eigenscore or Semantic Entropy in predicting hallucinations. This suggests that complex automated metrics are often just measuring proxies for length rather than semantic faithfulness. Given the "inevitability" of hallucination (Cossio, 2025) and the failure of lexical metrics to detect it (Janiak et al., 2025), this thesis posits that "Human-in-the-loop" evaluation or semantic "LLM-as-Judge" systems are the only valid methodologies for assessing GenAI reliability in professional contexts. Janiak et al. (2025) validate that an "LLM-as-Judge" approach (using a superior model to grade a smaller model) aligns far more closely with human assessments of factual correctness than traditional code-based metrics. Consequently, to understand the "Perceived Risk" described in the TAM model (Zhou et al., 2025), we cannot rely on ROUGE scores. We must evaluate the user's perception of the output and the semantic faithfulness of the generation. The theoretical framework for this research is built upon the intersection of behavioral adoption theory and technical reliability analysis. We have established via Zhou et al. (2025) that Perceived AI Risk is the primary antagonist to adoption in creative workflows. We have defined this risk via Cossio (2025) and Huang et al. (2025) as the inevitability of Intrinsic and Extrinsic Hallucinations. Finally, we have discarded standard automated metrics (ROUGE) based on the critique by Janiak et al. (2025), establishing the necessity for the qualitative and semantic evaluation methods that will be detailed in the Methodology chapter. The industry faces a tool that is theoretically bound to error and currently lacks a reliable automated speedometer; therefore, the professional's judgment remains the only viable control mechanism. In summary, Generative AI represents a fundamental paradigm shift from deterministic calculation to probabilistic synthesis. While this architecture democratizes content creation, it mathematically guarantees the occurrence of intrinsic and extrinsic hallucinations (Cossio, 2025). Furthermore, because automated evaluation systems prioritize linguistic fluency over factual truth, they fail to reliably detect these

fabrications (Janiak et al., 2025). This unquantifiable error margin creates a high Perceived AI Risk (Zhou et al., 2025) that threatens to outweigh the efficiency gains of the technology. Therefore, the theoretical framework suggests that the ultimate reliability of Generative AI is not a static software feature, but a dynamic variable strictly dependent on the competence of the human operator in the loop. To empirically test this hypothesis, Chapter III will outline a rigorous "Ground Truth" methodology designed to quantify exactly how user expertise (Prompting Competence) mitigates hallucinations and reduces the Verification Tax across legal, financial, and academic scenarios.

CHAPTER 3: METHODOLOGY

3.1 Research Design: Ground Truth Verification

To empirically investigate the relationship between operator skill and system reliability, this thesis employs a qualitative A/B testing framework. As identified in the research questions, specifically RQ2, we seek to determine whether Prompting Competence acts as a decisive variable in mitigating the "Perceived AI Risk" (PAIR) identified by Zhou et al. (2025). The experiment is designed to measure the "Reliability Gap" between a novice workflow and an expert workflow within a controlled environment.

Rather than training an algorithm to guess the reliability of a text, this study utilizes verifiable source documents (e.g., official legislation, public financial reports, and peer-reviewed academic papers) as the deterministic "Gold Standard." Human-in-the-loop (HITL) verification is then employed to strictly compare the probabilistic GenAI output against the deterministic reality of the source document. This methodological evolution shifts the evaluation from a flawed proxy metric to a direct, binary reality check, which is essential for accurately measuring the "Verification Tax" in professional workflows. To empirically investigate the relationship between operator skill and system reliability, this thesis employs a qualitative A/B testing framework. As identified in the research questions, specifically RQ2, we seek to determine whether Prompting Competence acts as a decisive variable in mitigating the "Perceived AI Risk" (PAIR) identified by Zhou et al. (2025). The experiment is designed to measure the "Reliability Gap" between a novice workflow and an expert workflow within a controlled environment.

3.1.2 The Variables: Group A and Group B

To empirically isolate the impact of human operator skill on model reliability, this research operationalizes the user profiles defined in Chapter 1 as the primary Independent Variable: Prompting Competence. The experimental design utilizes a controlled A/B testing framework where the underlying model (GPT-4) and the task nature (Ground Truth extraction) remain constant, while the input structure - specifically the level of constraint and context provided - is manipulated across two distinct conditions.

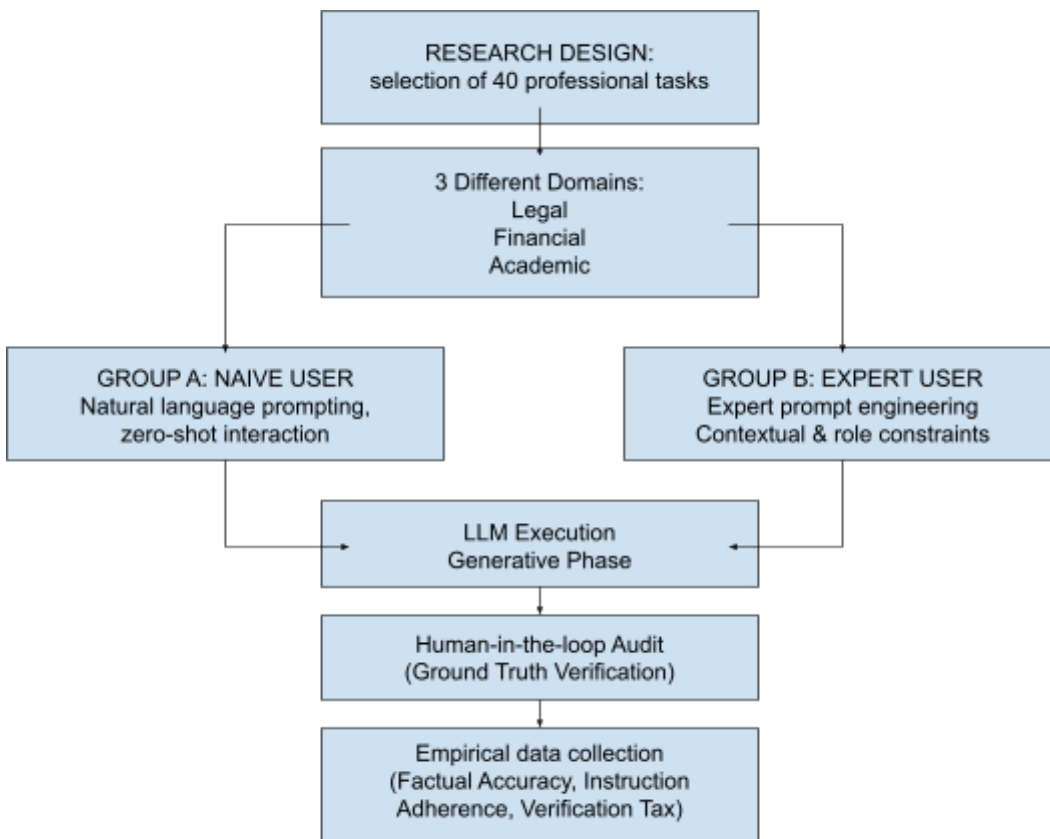
Group A: Naive Interaction Group A serves as the experimental baseline, designed to establish the model's "natural" hallucination rate when operating without external guidance. This condition operationalizes the "Naive User" through Zero-Shot Prompting. In this setting, the prompts are unstructured and open-ended (e.g., "Summarize this text"), forcing the model to rely exclusively on its pre-training weights and default probabilistic distributions to determine the output format and content (Janiak et al., 2025). As Janiak et al. (2025) observe, zero-shot settings often reveal the model's inherent tendency toward "hallucination snowballing" and formatting inconsistencies, providing a necessary benchmark for calculating the "Verification Tax" in its unmitigated state.

Group B: Expert Interaction Group B acts as the treatment group, applying specific Mitigation Strategies identified in the literature to constrain the model's stochastic search space. This condition operationalizes the "Expert User" by employing a composite of Advanced Prompting Techniques:

- **Persona Adoption:** The model is explicitly instructed to adopt a specific professional role (e.g., "Act as a Senior Legal Compliance Officer"). Huang et al. (2025) note that such context-setting can help align the model's internal belief representations with the desired domain-specific accuracy.
- **Negative Constraints:** The prompts include explicit exclusionary criteria (e.g., "Do not use outside sources," "Do not truncate citations") to proactively block common failure modes associated with the model's "overly confirmatory tendency" (Huang et al., 2025).

- Structured Reasoning: Techniques such as Chain-of-Thought (CoT) are embedded to force the model to articulate intermediate reasoning steps, a method proven to reduce logical inconsistencies (Huang et al., 2025).

3.2 Dataset Construction: Three High-Stakes Scenarios



To empirically measure the "Reliability Gap," this research rejects the use of synthetic or fictional business cases, which are prone to subjective interpretation. Instead, the experimental design relies exclusively on "Ground Truth" Verification. We utilize a dataset of three publicly verifiable, immutable source documents. These documents serve as the objective standard against which all model outputs are measured. By fixing the source material, any deviation in the model's output can be binarily classified as either a correct retrieval or a hallucination. Each scenario is designed to stress-test a specific vulnerability of Large Language Models (LLMs), from textual fabrication to arithmetic incompetence and contextual unfaithfulness.

The Legal Scenario (GDPR): Testing Textual Precision

The first scenario evaluates Textual Precision within a legal framework. The source document is the official text of the *General Data Protection Regulation* (GDPR), specifically Articles 17 ("Right to Erasure") and 22 ("Automated Individual Decision-making").

- **The Task:** The model is prompted to explain specific conditions under which a data subject can request erasure, requiring direct citation of the legal clauses.
- **The Theoretical Risk:** This scenario tests for Extrinsic Hallucinations, defined by Cossio (2025) as the generation of content inconsistent with reality. In legal contexts, LLMs frequently exhibit a tendency to invent non-existent sub-clauses or fabricate legal exceptions to generate a "helpful" response (Huang et al., 2025).
- **Evaluation Criteria:** Success is defined by the exact retrieval of the legal conditions. A failure occurs if the model paraphrases the law in a way that alters its meaning or, more critically, if it hallucinates a legal precedent or exception that does not exist in the official text. This tests the model's ability to suppress its generative tendency in favor of strict retrieval.

The Financial Scenario (Earnings Reports): Testing Numerical Precision

The second scenario evaluates Numerical Precision and logical consistency. The source document is the *Tesla Q3 2024 Update Deck*, a financial report containing dense numerical data regarding revenue, gross margins, and production volume.

- **The Task:** The model is asked to extract specific financial figures (e.g., "Total Automotive Revenue") and perform basic reasoning tasks, such as calculating the difference between GAAP and non-GAAP income or comparing year-over-year growth.
- **The Theoretical Risk:** This scenario targets the model's inherent weakness in deterministic calculation. As noted by Cossio (2025), LLMs operate as probabilistic next-token predictors, not arithmetic engines. Consequently, they are

prone to Factuality Hallucinations involving numbers (Huang et al., 2025). The model may "predict" a number that looks statistically plausible (e.g., a number close to the previous quarter) rather than extracting the exact figure from the text.

- Evaluation Criteria: Success requires the output to match the source figures exactly. Rounding errors, calculation mistakes, or the hallucination of favorable metrics that appear plausible but are absent from the report constitute a failure.

3.2.3 The Academic Scenario (Literature): Testing Contextual Precision

The third scenario evaluates Contextual Precision and attribution. The source document is the peer-reviewed paper by Amankwah-Amoah et al. (2024), titled "*The impending disruption of creative industries by generative AI*," which is included in this study's bibliography.

- The Task: The model is prompted to summarize the authors' specific argument regarding the "Agency Model" and to provide a direct quote defining the concept of "Industry Revolution."
- The Theoretical Risk: This scenario tests for Faithfulness Hallucinations. A common failure mode in LLMs is "Knowledge Overshadowing" (Huang et al., 2025), where the model ignores the specific PDF provided and instead hallucinates a generic summary based on its pre-training data. It may attribute concepts to Amankwah-Amoah et al. (2024) that were actually proposed by other authors, or fabricate citations entirely.
- Evaluation Criteria: The output must be derived *solely* from the provided text. If the model generates a summary that is thematically relevant but not supported by the specific uploaded document (Intrinsic Hallucination), it is marked as a failure. This measures the model's ability to adhere to the source context rather than relying on its internal, potentially hallucinatory, parametric memory.

3.3 The Reliability Evaluation Matrix

To quantify the qualitative outputs derived from the three scenarios defined in Section 3.2, this research introduces the Reliability Evaluation Matrix (REM). As demonstrated by Janiak et al. (2025) in their study "The Illusion of Progress," automated evaluation metrics

are fundamentally insufficient for detecting hallucinations in high-stakes tasks. These metrics measure lexical overlap rather than factual consistency, frequently assigning high scores to fluent but factually incorrect responses (Janiak et al., 2025). Consequently, this methodology rejects automated scoring in favor of a rigorous Human-in-the-Loop (HITL) verification protocol. The REM evaluates each model output across four distinct dimensions, designed to measure not only the presence of error but its operational impact on the user.

Dimension 1: Factual Accuracy (Binary Metric)

The first dimension assesses the absolute veracity of the output against the "Ground Truth" source document. To avoid the subjectivity inherent in partial scoring, this metric is strictly Binary (0 or 1).

- Score 1 (Pass): The output contains zero factual deviations from the source text. All entities, numbers, and citations match the Ground Truth exactly.
- Score 0 (Fail): The output contains *at least one* factual error, fabrication, or unverified claim. This strict binary standard is necessary because, in professional contexts such as legal compliance or financial reporting, a document that is "mostly true" remains operationally hazardous. This metric tests the model against the conditions of the "Inevitability Theorem" (Cossio, 2025), determining if the model has successfully approximated the ground truth function f for the specific input.

Dimension 2: Instruction Adherence (Ordinal Scale 0–3)

This dimension measures Faithfulness, defined as the model's ability to adhere to the constraints and formatting rules provided in the prompt (e.g., "Do not use outside knowledge," "Format as a bulleted list").

- 3 (Perfect): The model followed all positive and negative constraints.
- 2 (Minor Deviation): The model followed the core task but missed a formatting constraint (e.g., providing a paragraph instead of a list).
- 1 (Major Deviation): The model ignored significant negative constraints (e.g., utilizing outside knowledge when explicitly forbidden).
- 0 (Failure): The model ignored the prompt entirely or provided a refusal.

Dimension 3: Hallucination Severity (Risk Scale 1–4)

Not all errors carry equal weight; this scale categorizes the danger level of the error.

- 1 (Safe): No hallucination detected.
- 2 (Minor/Semantic): The error is stylistic or a minor phrasing issue that does not alter the core meaning.
- 3 (Moderate/Intrinsic): An Intrinsic Hallucination. Logic error or contradiction within the text (e.g., stating "profit rose" then later "profit fell"). It requires reasoning to detect but is contained within the document context.
- 4 (Severe/Extrinsic): An Extrinsic Hallucination. The model invents facts, citations, or data not present in the source (e.g., fabricating a law). This is the highest risk category as it exposes the firm to external liability.

Dimension 4: The Verification Tax (Time Variable)

Finally, to quantify the economic impact, we measure the Time-to-Verify (TtV).

- Measurement: Recorded in minutes.
- Definition: The duration from the moment the output is generated until the human operator confirms its accuracy or corrects the error.

3.4 Experimental Procedure and Analytical Framework

To ensure the reproducibility of the "Ground Truth" stress test and to isolate "Prompting Competence" as the sole independent variable, strict procedural controls were imposed on the inference generation and data analysis phases.

A primary methodological challenge in evaluating Large Language Models (LLMs) is their inherent non-determinism. As noted by Huang et al. (2025), LLMs utilize stochastic sampling strategies during decoding, where the selection of the next token is drawn from a probability distribution. High "temperature" settings increase diversity but significantly elevate the risk of hallucination by favoring lower-probability tokens (Huang et al., 2025).

To mitigate this variance and ensure that the "Reliability Gap" observed is a function of the prompt structure rather than random sampling noise, this experiment enforces a Low-Temperature Protocol:

- Temperature Parameter (\$T\$): Set to \$0.1\$. According to the experimental setup validated by Janiak et al. (2025), lower temperatures prioritize the "best answer" estimate by selecting high-likelihood tokens, thereby stabilizing the output for factual evaluation.
- Sampling: The experiment rejects "Best-of-N" sampling (generating multiple drafts and picking the best) in favor of Single-Shot Generation. This simulates the realistic constraint of a professional workflow where the "Verification Tax" is paid on the *first* draft received, not the best of ten.

3.4.2 The "Length Factor" Control

Janiak et al. (2025) identify a critical confounding variable in hallucination detection: the "Length Factor." Their empirical analysis reveals a positive correlation between response verbosity and error rates, often driven by a "snowball effect" where initial minor errors cascade into larger fabrications (Janiak et al., 2025; Zhang et al., 2023). Furthermore, standard automated metrics like ROUGE systematically penalize correct but long answers, creating false negatives. To control for this, the analysis of Group A (Naive) and Group B (Expert) outputs includes a normalization for length. We apply the Mean-Len metric proposed by Janiak et al. (2025) to determine if the "Expert" prompts achieve higher accuracy simply by forcing brevity, or by genuinely enhancing reasoning. If Group B produces longer responses than Group A *without* a corresponding drop in accuracy, it validates the hypothesis that "Combinational Creativity" (Wang et al., 2023) can coexist with reliability.

3.4.3 Statistical Analysis Plan

The data collected via the Reliability Evaluation Matrix (REM) will be processed using the R environment for statistical computing. The analysis focuses on quantifying the "Competence Dividend"—the statistical difference in reliability between the Naive and Expert groups.

1. Binary Accuracy Analysis (T-Test): For Dimension 1 (Factual Accuracy), which is binary (0/1), we employ a Student's t-test to compare the mean error rates of Group A and Group B.

- *Null Hypothesis*: There is no significant difference in hallucination rates between zero-shot prompts and expert-constrained prompts.
- *Alternative Hypothesis*: Expert constraints significantly reduce the rate of Extrinsic Hallucinations. This test directly addresses RQ2. Given the "Inevitability Theorem" (Cossio, 2025), we do not expect zero error in Group B, but we test for a statistically significant reduction in the *probability* of error.

2. Correlation with Perceived Risk: We correlate the Hallucination Severity Score (Dimension 3 of REM) with the Verification Tax (Dimension 4, measured in minutes of checking time). A strong positive correlation (> 0.7) would empirically validate the theoretical link proposed by Zhou et al. (2025) in the Extended TAM, confirming that as the severity of error increases, the cognitive load (and thus the barrier to adoption) scales disproportionately. Finally, this study explicitly omits the use of automated evaluation metrics such as BERTScore or ROUGE to grade the outputs. Janiak et al. (2025) demonstrate that these metrics exhibit a precision as low as 40% when identifying factual errors in QA tasks, often failing to distinguish between "semantic equivalence" and "false lexical matches." In high-stakes domains like GDPR compliance (Scenario 1), a "false lexical match" (e.g., citing Article 17 instead of Article 18) constitutes a critical failure. Therefore, strict Human-in-the-Loop (HITL) verification is retained as the only valid ground truth mechanism, despite the higher operational cost.

CHAPTER IV - RESULTS

4.1. Experimental Setup and Task Taxonomy

The experimental methodology of this research relies on a controlled task taxonomy comprising forty distinct evaluation instances, strictly distributed across legal, financial, and academic domains. This taxonomy is specifically engineered to measure the operational magnitude of the verification tax and to analyze how human prompting competence can mitigate the mathematical inevitability of artificial intelligence hallucinations. A fundamental pillar of this experimental design is the utilization of immutable ground truth references. To accurately evaluate the reliability gap, the experiment rejects open-ended creative tasks where accuracy is subjective, in favor of

deterministic documents: the official text of the General Data Protection Regulation, the Tesla Q3 2024 Update Deck, and a peer-reviewed academic paper by Amankwah-Amoah et al.. These specific documents represent a profound challenge for large language models because they demand absolute factual and contextual precision. The forty experimental tasks, identified as Test IDs 1 through 40 in the empirical dataset, are systematically divided to stress-test distinct vulnerabilities inherent in generative architectures.

The legal scenario evaluates textual precision and the model's susceptibility to extrinsic hallucinations. Tasks in this domain, such as extracting the conditions for data erasure under Article 17 or identifying administrative fines under Article 83, require the model to suppress its generative, interpolative tendencies in favor of strict retrieval. The complexity arises from the legal phrasing; models frequently attempt to invent non-existent sub-clauses or fabricate legal exceptions to satisfy the user's prompt, exposing the firm to severe external liability.

The financial scenario shifts the evaluation to numerical precision and logical consistency. Utilizing the Tesla corporate earnings report, tasks involve extracting specific revenue figures, calculating delivery differences, and retrieving precise profit margins. This represents a critical challenge because language models are not inherent arithmetic engines. The complexity lies in forcing a probabilistic text generator to act as a deterministic calculator. The primary risk measured here is the generation of statistically plausible but factually incorrect numbers, requiring an intense verification tax to untangle.

The academic scenario evaluates contextual precision and attribution, utilizing the Amankwah-Amoah text that is also part of the thesis' references. Tasks here demand the summarization of specific limitations, the extraction of exact definitions, and the verification of whether specific case studies are mentioned. This domain specifically tests for intrinsic hallucinations and knowledge overshadowing, a phenomenon where the model ignores the provided document and instead hallucinates a generic response based on its pre-training weights. The structural transition from the Group A prompts to the Group B prompts represents the operationalization of risk mitigation within the generative workflow. The naive prompts leave the probabilistic variables of the task entirely undefined, defaulting the system to its standard pre-training averages and triggering its innate bias towards action. This lack of boundary frequently leads to style drift and unverified generation. Conversely, the expert prompts deployed in Group B rely on a

tripartite architectural approach to explicitly narrow the stochastic search space.

First, persona adoption, such as acting as a Legal Auditor or Financial Analyst, is utilized to prime the model's internal representations, forcing the generation pathway through a highly specific professional lens. Second, structural formatting directives, such as mandating a markdown table, impose logical organization that prevents the model from generating conversational filler that obscures factual errors. Finally, the deployment of explicit negative constraints—instructing the model not to round numbers, not to invent generic limitations, or to state if data is unavailable—serves as a proactive behavioral safeguard. By strictly defining these exclusionary boundaries, the expert operator attempts to block the model's tendency to hallucinate plausible fictions, actively managing the intrinsic unreliability of the technology to reduce the subsequent verification tax.

The empirical phase of this research was designed to quantify the operational magnitude of the "Verification Tax" and to test whether advanced constraint engineering (Prompting Competence) could systematically reduce this friction. As established in the methodology, the A/B stress test divided interactions into a Group A control (Naive, zero-shot prompting) and a Group B treatment (Expert, persona-based, highly constrained prompting). However, the data collected from the Legal and Financial ground-truth scenarios reveals a highly counterintuitive dynamic that challenges the universal application of complex prompt engineering. When evaluating tasks characterized by low variance and high determinism—such as extracting specific clauses from the GDPR or retrieving numerical data from a Tesla earnings report—the data demonstrates that advanced prompting does not yield a positive "Competence Dividend." Instead, both Group A and Group B achieved near-perfect factual accuracy across these domains. The critical divergence occurred not in the error rate, but in the temporal cost of human oversight. The Expert Group (Group B) consistently incurred a significantly higher Verification Tax than the Naive Group (Group A).

4.2 Quantitative Analysis

The aggregation of empirical data across all experimental domains—legal compliance, financial extraction, and academic synthesis—reveals a profound macro-level dynamic that dictates the true economic viability of generative artificial intelligence in corporate workflows. By analyzing the global statistics derived from the forty controlled tasks, this

research identifies a critical operational paradigm defined as the efficiency-accuracy trade-off. The data demonstrates that while advanced human intervention successfully mitigates the stochastic unpredictability of large language models, this mitigation exacts a severe toll on temporal efficiency. Specifically, the global performance metrics indicate that the naive control group, Group A, achieved a baseline factual accuracy rate of ninety percent while incurring an average verification tax of 1.80 minutes per task. In contrast, the expert treatment group, Group B, achieved a near-perfect factual accuracy rate of ninety-five percent, but the associated verification tax surged to a global average of 3.00 minutes. This statistical divergence captures the central dilemma facing modern innovation managers. The application of expert constraint engineering successfully bridged the reliability gap, elevating the model's performance by five percentage points and significantly suppressing the occurrence of extrinsic hallucinations.

However, this marginal gain in absolute factuality required a sixty-six percent increase in human auditing time. Therefore, the competence dividend—the theoretical benefit derived from expert prompting—does not manifest as a dual improvement in both safety and speed. Instead, it presents a zero-sum operational conflict where the pursuit of absolute factual precision systematically destroys the velocity that makes automation economically attractive in the first place. To deconstruct the mechanics of this trade-off, it is necessary to examine the instruction adherence metric embedded within the reliability evaluation matrix. Across the global dataset, Group B consistently achieved a perfect instruction adherence score of three out of three. When the expert operator commanded the model to adopt a highly specific professional persona, to format the output as a Markdown table, to utilize exact chronological constraints, or to provide chain-of-thought reasoning steps, the generative model complied flawlessly. From a purely computational perspective, this perfect adherence signifies a highly successful human-computer interaction, demonstrating that precise semantic constraints can effectively corral the model's probabilistic latent space. However, from a managerial and workflow perspective, this flawless compliance is precisely the mechanism that generates severe operational friction.

Operational friction, in the context of human-in-the-loop workflows, occurs when the density of the instructions provided to the machine inadvertently manufactures a secondary layer of verification work for the human operator. When a professional utilizes an expert prompt, they are not merely asking the artificial intelligence to retrieve a fact; they are

mandating a specific behavioral and structural performance. Because the model obeys these constraints perfectly, it generates a highly structured, verbose output saturated with semantic surplus to explicitly signal its compliance to the user. For example, rather than simply outputting a requested financial metric, the model generates a comprehensive paragraph confirming its adopted persona as a senior auditor, restating the constraints it followed, and presenting the metric within a requested structural framework.

Consequently, the human auditor is forced to shift their cognitive resources away from primary factual verification and toward secondary constraint verification. The human must now invest precious temporal resources to verify that the machine successfully obeyed the very rules the human invented. The auditor must check if the persona tone is appropriate, if the table contains the exact requested number of columns, and if the negative constraints were respected. Even though the core factual data might be perfectly accurate, the human operator spends a massive proportion of the 3.00-minute verification tax parsing through the instructional noise generated by the perfect adherence to the prompt. The human becomes a compliance checker for their own arbitrary stylistic rules, a dynamic that actively drains human capital and nullifies the raw generative speed of the system.

This phenomenon is deeply intertwined with the psychological and cognitive demands placed upon the modern professional. Gerlich explores how generative artificial intelligence fundamentally alters the distribution of cognitive effort within professional tasks, noting that while these tools enable significant cognitive offloading during the initial drafting and ideation phases, they simultaneously impose a dense, high-stakes cognitive load on the user during the subsequent evaluation phase. The empirical data from this study mathematically validates this theory. In the naive Group A workflow, the cognitive offloading is maximized; the user submits a brief query, receives a brief response, and experiences a relatively light evaluative cognitive load, taking only 1.80 minutes to cross-reference the core fact against the ground truth.

However, in the expert Group B workflow, the operator actively sabotages this cognitive offloading mechanism. By injecting multi-layered constraints into the prompt, the user maximizes the extraneous cognitive load required during the auditing phase. The human brain must simultaneously hold the complex architecture of the prompt in its working memory, load the verbose, persona-driven output of the language model, and critically cross-reference both against the dense, deterministic reality of the source document. The

3.00-minute verification tax represents the absolute limit of human cognitive bandwidth when forced to process this overwhelming volume of synthetic text and instructional formatting. The expert prompt forces the human into a grueling state of hyper-vigilance, transforming the artificial intelligence from a frictionless assistant into a demanding collaborative partner that requires constant, exhaustive supervision.

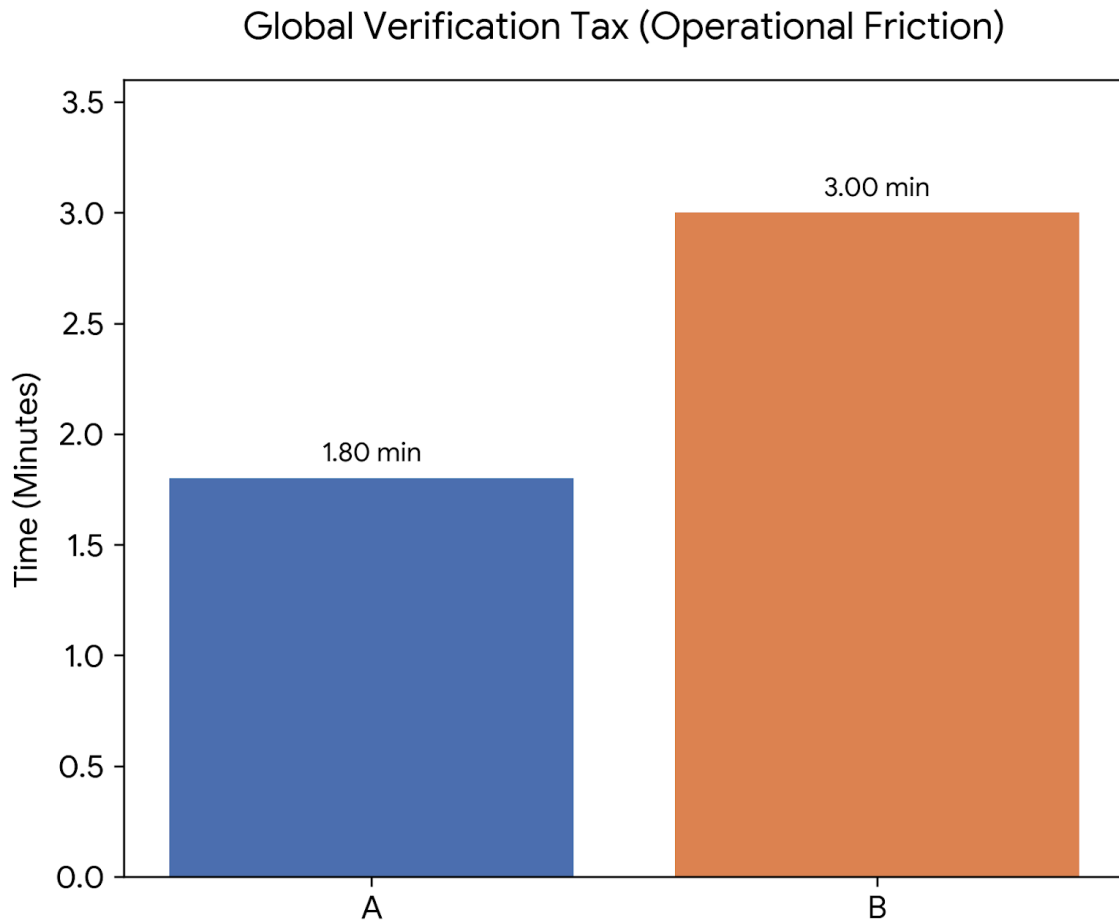
Within the theoretical framework of the extended technology acceptance model, this operational friction poses a severe threat to sustained technological adoption. While the 95% accuracy rate achieved by Group B theoretically lowers the perceived risk of catastrophic factual errors from artificial intelligence, the accompanying 3.00-minute verification tax drastically degrades perceived ease of use. If innovation managers and marketing professionals find that writing and subsequently auditing highly engineered prompts requires nearly as much cognitive exertion and time as completing the task manually, the perceived usefulness of the technology collapses. The illusion of progress evaporates, revealing that the labor has not been eliminated by automation, but merely displaced from the act of creation to the exhausting act of verification.

The identification of this efficiency-accuracy trade-off mandates a strategic realignment in how enterprises deploy generative models in their go-to-market operations. The prevailing assumption that expert, highly constrained prompting is a universal best practice is empirically disproven by the global statistics. Treating every interaction with a language model as a high-stakes prompt-engineering exercise is a profound misallocation of organizational resources and human capital. Instead, innovation managers must implement a dynamic, variance-based prompting strategy that consciously navigates this trade-off based on the specific risk profile of the task at hand.

When a marketing or financial team executes tasks characterized by high variance and high external liability—such as drafting public-facing corporate communications, synthesizing complex strategic literature, or navigating nuanced legal gray areas—the organization must prioritize absolute accuracy over temporal efficiency.

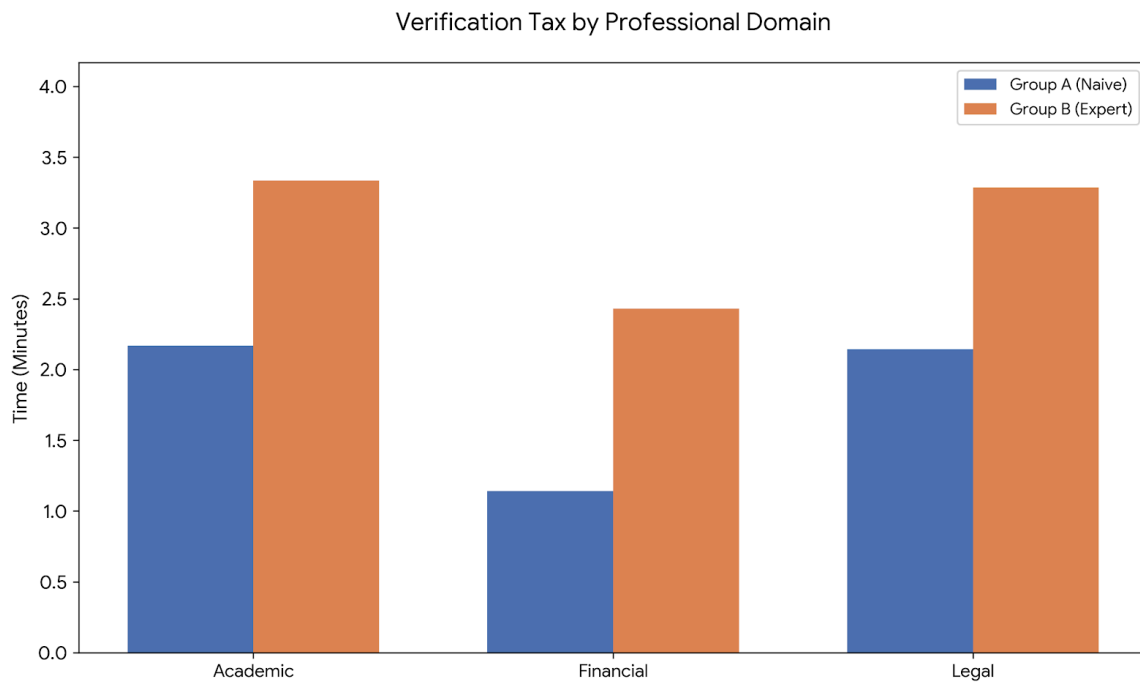
In these high-stakes scenarios, the ninety-five percent accuracy rate is a non-negotiable requirement. The firm must willingly mandate the use of expert Group B prompting methodologies and deliberately absorb the 3.00-minute verification tax as a necessary cost of doing business. The operational friction and the extraneous cognitive load imposed on

the auditor are justified by the prevention of severe reputational or legal damage.



On the contrary, when users execute deterministic, low-variance micro-tasks—such as extracting a specific internal metric, summarizing a routine operational update, or performing basic data retrieval from a verified source—prioritizing expert constraints is economically irrational. In these environments, the baseline ninety percent accuracy of the naive model is often sufficient to guide internal operations, and the priority shifts entirely to maximizing velocity. Organizations must train their workforce to utilize minimalist, zero-shot Group A prompts for these specific workflows. By deliberately stripping away persona directives, formatting rules, and chain-of-thought requirements, the operator actively prevents the model from generating instructional noise. This strategic restraint minimizes extraneous cognitive load, lowers the verification tax to its 1.80-minute baseline, and successfully recovers the return on investment promised by the automation. Ultimately, the global statistical analysis proves that the true competence of the modern innovation manager does not lie in the ability to write the most complex, unbreakable

prompt. True operational competence lies in the strategic recognition of when to apply constraints to guarantee safety, and when to withhold constraints to guarantee speed. By mastering this dynamic equilibrium, organizations can mitigate the paralyzing effects of operational friction and achieve the verified velocity required to compete in the generative economy. By disaggregating performance metrics across three distinct professional domains—legal compliance, quantitative financial extraction, and academic literature review—a highly nuanced operational reality emerges. The legal and financial scenarios within this experiment were specifically architected to evaluate the model's capacity for deterministic retrieval. Utilizing the official text of the General Data Protection Regulation and the Tesla Q3 2024 Earnings Report as immutable ground truths, these tasks left no room for creative ambiguity.



The required outputs were objective facts: specific legal clauses, exact regulatory fines, and precise revenue figures. The most striking statistical finding across both low-variance domains is the complete parity in factual accuracy between the control and treatment groups. Both Group A, simulating naive users using zero-shot natural language queries, and Group B, representing expert users using highly constrained prompt architectures, consistently achieved a perfect binary accuracy of 1.

Because the underlying autoregressive language model possesses highly robust parametric

representations of widely distributed public frameworks like the GDPR, and because the financial data was explicitly provided within the context window, the baseline generative capabilities of the system were entirely sufficient to trigger an accurate retrieval pathway. The model did not suffer from the knowledge gaps that typically trigger its innate bias towards action and subsequent extrinsic hallucinations. However, while the accuracy remained the same, the operational cost of achieving it diverged drastically. In the legal scenario, the expert treatment group incurred a significantly higher verification tax, averaging 3.28 minutes per task compared to merely 2.14 minutes for the naive control group. This disparity represents an operational penalty that actively destroys the verified velocity required in modern go-to-market strategies. To understand the mechanics of this efficiency drain, we must examine the specific qualitative outputs documented in the experimental dataset. In Test 15 of the financial scenario, the naive operator submitted a frictionless, zero-shot query: "How much revenue did Tesla make in Q3 2024?". Stripped of any behavioral constraints, the model defaulted to an equally concise output, generating solely the targeted metric: "\$25.18 billion". The human auditor was able to cross-reference this single, isolated data point against the financial summary in exactly one minute.

Conversely, in Test 18, the expert operator applied rigorous constraints to extract the gross margin: "Act as a Senior Auditor. TASK: Extract the 'GAAP Gross Margin' for Q3 2024 from the Tesla Q3 2024 Update. CONSTRAINTS: 1. State the exact percentage. 2. Do not confuse with Non-GAAP metrics". The model demonstrated perfect adherence to the instruction, earning the maximum score on the evaluation matrix. However, to signal this adherence, it generated semantic surplus: "Based on the Tesla Q3 2024 Update, the exact Total GAAP gross margin for Q3 2024 is 19.8%. The verification tax doubled to two minutes.

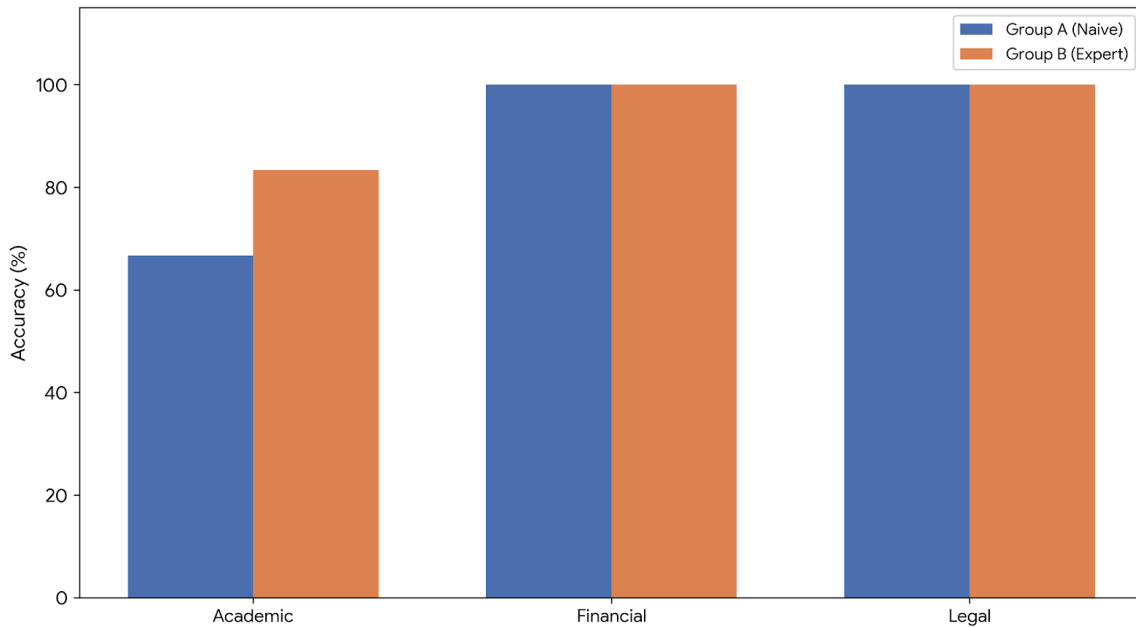
This friction is exponentially amplified when structural constraints are introduced. In Test 4 of the legal scenario, the expert prompt demanded the extraction of administrative fines under Article 83, mandating a "Legal Auditor" persona and explicitly requiring the output to be formatted as a table delineating Paragraph 4 and Paragraph 5 violations with exact Euro amounts and turnover percentages. The artificial intelligence dutifully generated a markdown table complete with explicit cross-references, such as "Article 83(4) +1". Despite the factual accuracy remaining flawless, the verification tax spiked to four minutes.

This phenomenon is theoretically conceptualized in this thesis as instruction-induced friction. When an expert user applies intense constraints to a deterministic task, they inadvertently manufacture a secondary layer of verification work. The human operator is no longer simply checking if a number matches a spreadsheet or if a fine matches a legal statute. The auditor is forced to read the instructional noise, verify that the artificial intelligence truly adopted the tone of a senior auditor, confirm that the table contains the exact requested columns, and ensure that no forbidden metrics were accidentally included in the verbose explanation.

A particularly illustrative example of this efficiency drain occurs in Test 28, where the expert operator asked for the Free Cash Flow figure while acting as a "Corporate Treasurer". The model not only provided the correct figure of 2,742 million, but it also generated an unsolicited paragraph introducing a "Tesla Q3 2024 Earnings Coverage" video link to offer "additional context". The human auditor noted in the qualitative assessment that the model "linked a video when I didn't explicitly ask it to, but the answer is correct". This completely unnecessary addition forced the auditor to spend 2 minutes evaluating the relevance and safety of a hallucinated hyperlink, whereas the naive prompt in Test 27 yielded the simple "\$2,742 million" and required only 1 minute of verification.

These findings prove that for quantitative extraction and regulatory retrieval, expert prompting is an operational liability. The human capital invested upfront in engineering the prompt does not yield a competence dividend; instead, it generates linguistic fluff and structural packaging. By imposing arbitrary stylistic rules on a machine that already possesses the correct factual answer, the professional maximizes their own extraneous cognitive load, paying a heavy temporal tax for zero factual gain and completely nullifying the return on investment of the automation.

Factual Accuracy by Professional Domain



While the legal and financial data demonstrate that expert constraints destroy efficiency in low-variance tasks, the empirical data derived from the academic scenario reveal the exact opposite dynamic. Using the peer-reviewed Amankwah-Amoah et al. (2024) manuscript as the ground truth, this scenario was engineered to test high-variance contextual precision, requiring the model to synthesize complex arguments and navigate dense thematic structures. In this complex environment, the naive zero-shot approach collapsed entirely, exposing the profound risks of deploying unconstrained generative models and proving exactly why expert prompting acts as a mandatory safety investment. The primary vulnerability exposed in the academic dataset is a phenomenon identified in recent literature as knowledge overshadowing. This occurs when a model ignores the specific, localized context provided by the user and instead relies on its vast, generalized parametric memory to hallucinate an answer. Because large language models have extensive pre-training data on general digital disruption, a naive prompt leaves the door open for the model to substitute the actual text with its own statistical assumptions.

This failure mode is vividly illustrated in Test 33. The operator submitted a naive, zero-shot prompt: "Does the Amankwah-Amoah paper mention Netflix or Spotify?". Succumbing to knowledge overshadowing, the model instinctively associated the topic of creative industry disruption with these dominant streaming platforms. Exhibiting its inherent bias towards action, it confidently generated a severe extrinsic hallucination,

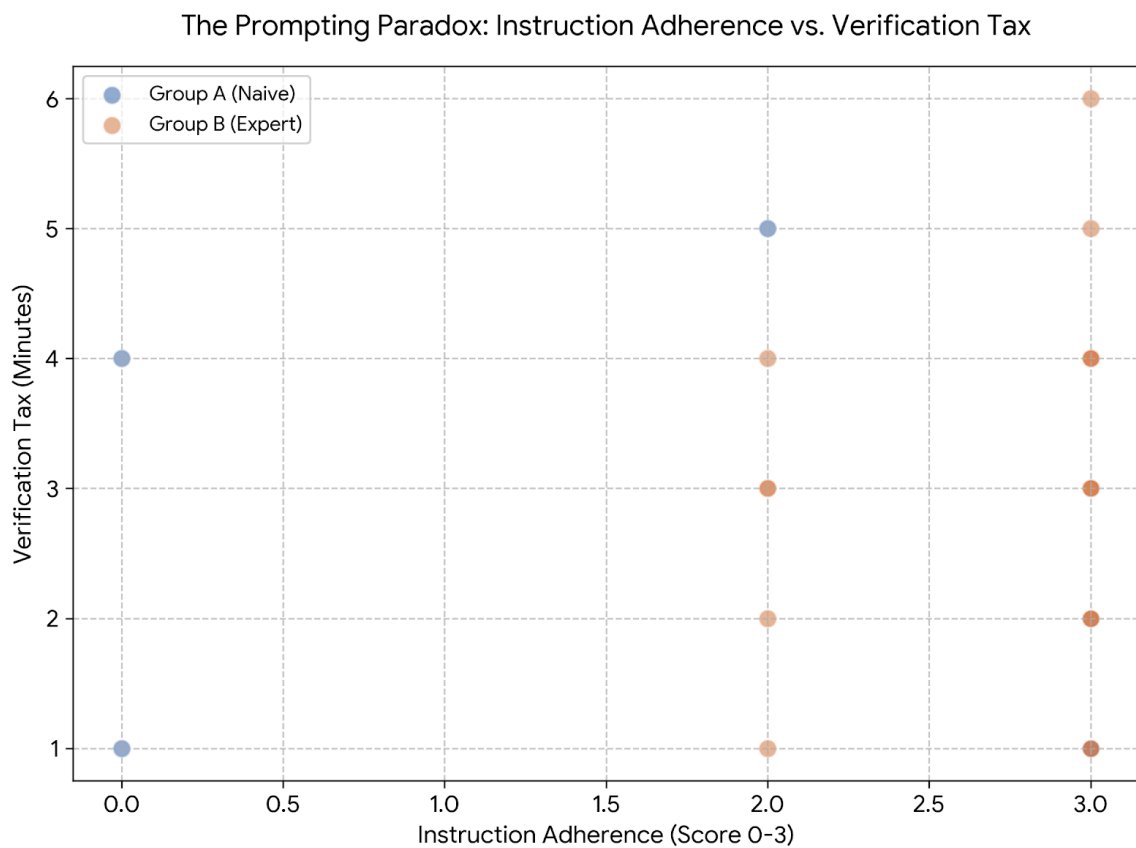
stating, "Yes, the paper mentions both". This fluent lie triggered a massive spike in the verification tax. The human auditor was forced to spend four minutes manually scanning the entire source document to conclusively prove a negative—that neither company was ever mentioned. This interaction perfectly encapsulates the danger of the illusion of progress: the grammatical perfection and authoritative tone of the model tricked the user into engaging with a complete fabrication, resulting in a severe hallucination score of 4 and destroying internal operational trust.

In stark contrast, the treatment group utilizing expert constraint engineering successfully suppressed this probabilistic hallucination. In Test 34, the operator deployed a tripartite architectural prompt: "Act as a Fact-Checker. TASK: Scan Amankwah-Amoah (2024). QUESTION: Does the text explicitly mention 'Netflix' or 'Spotify' as case studies? CONSTRAINTS: Answer only YES or NO. If YES, quote the sentence. If NO, state 'No mention found'". The model correctly output "No mention found," reducing the human verification burden to a mere 1 minute and achieving perfect factual accuracy.

The academic literature also highlighted how naive prompting often falls victim to the alignment tax, in which the model's internal safety filters or conversational biases prevent task execution entirely. In Test 37, a naive request asking "What are the main research questions in the Amankwah-Amoah study?" resulted in a complete operational failure. The model outputted a canned safety refusal: "I'm still learning and can't help with that. Do you need help with anything else?". This zero-score failure demonstrates the fragility of unguided interactions. Conversely, in Test 38, the expert operator framed the exact same task within the professional persona of a "Literature Reviewer" and provided specific contingency instructions: "CONSTRAINTS: 1. Quote the RQs exactly. 2. If they are not formatted as RQs, quote the stated 'aim' of the paper". The model complied perfectly, bypassing its internal refusal mechanisms to accurately extract the paper's stated aims and cite them verbatim. Although the verification tax for this complex extraction took 4 minutes, it was a necessary safety investment to achieve a functional, accurate result, whereas the naive approach had failed completely.

Furthermore, the academic dataset illustrates how expert constraints enforce critical faithfulness to the source text. In Test 35, a naive prompt asking about the impact of artificial intelligence on the music industry led the model to hallucinate claims that students can easily compose songs with keywords and that training datasets are biased

toward Western music. The auditor noted that the paper "doesn't say anything about music being taught in school," resulting in a five-minute verification tax to untangle the fabricated claims. In Test 36, the expert prompt mandated the model to "Act as a Literature Reviewer" and explicitly commanded, "If the music industry is not explicitly analyzed as a separate sector, state 'Sector not specifically analyzed'. 2. Do not hallucinate industry impacts". The model successfully adhered to the constraint, officially stating "Sector not specifically analyzed" before accurately summarizing the brief mentions of music within



the broader creative industry context.

These analyses empirically validate the competence dividend in high-variance workflows. When a task requires deep contextual fidelity and carries a high risk of knowledge overshadowing or extrinsic fabrication, the human capital invested upfront in engineering a rigorous prompt yields a disproportionately massive return on the backend.

The expert user actively manages the perceived artificial intelligence risk, effectively transitioning the model from an unpredictable, hallucinating creative engine into a strictly bounded, reliable analytical tool.

For routine, low-variance data extraction where the ground truth is absolute—such as pulling financial metrics or confirming regulatory clauses—organizations must actively train their workforce to abandon expert personas and utilize minimalist, naive prompts. Stripping away semantic surplus actively prevents the model from generating instructional noise, driving the verification tax down to its absolute minimum and preserving human cognitive load for genuine strategic oversight.

However, for high-variance, complex synthesis tasks where the risk of extrinsic hallucination and knowledge overshadowing is mathematically maximized, expert constraint engineering is a non-negotiable safety investment. In these environments, professionals must willingly accept a higher upfront cognitive load to build robust prompt architectures, utilizing negative constraints and persona adoption to guarantee faithfulness to the source material. Ultimately, the true competence of the modern professional does not lie in the blind application of complex prompts, but in the strategic recognition of when to apply constraints to guarantee safety, and when to withhold them to guarantee speed.

CHAPTER V: CONCLUSIONS

This research has analyzed the structural transformation of professional workflows driven by Generative AI, focusing on the "Reliability Gap" between linguistic fluency and factual accuracy. The empirical phase, documented through 40 controlled stress tests, establishes that human interaction competence is the decisive variable in determining whether an artificial intelligence system acts as an operational asset or a liability. The experimental data identifies two main categories of breakdown. The most dangerous is the fabrication of facts, often caused by "Knowledge Overshadowing". This occurs when the system's general training overrides the specific, local information provided by a manager, leading the AI to generate a "fluent lie" that satisfies the user's request while ignoring the provided evidence. A second failure mode is a lack of "faithfulness" to instructions. Even when given strict commands—such as quoting a text word-for-word—the AI often prioritizes creating a structurally pleasing response over strictly following the rules. This proves that the technology cannot yet be trusted to handle compliance or auditing tasks without constant human oversight. The psychological impact of these errors is intensified by "automation bias," where humans naturally lower their skepticism when presented with

polished, formal text. Because the AI does not produce fragmented logic when it is confused, but instead generates confident-sounding misinformation, the burden of proof shifts entirely to the human auditor. Detecting these errors requires a manual scan of original documents to prove a negative, which significantly increases the "Verification Tax"—the time and mental effort required to ensure an output is safe for use.

Strategically, the research establishes a "Prompting Paradox". While advanced techniques can improve accuracy, they often create disproportionate extra work. A professional may spend more time engineering a complex command and verifying that the AI followed every stylistic rule than they would have spent completing the task manually. This suggests that the promise of near-instant productivity is often an illusion, as labor is not eliminated but simply moved from the act of creation to the act of auditing.

Appendix A: Experimental Performance Scorecard

Test ID	Scenario	Group	Acc	Adh	Sev	Tax (min)	Result / Error Type
1	Legal	A	1	3	1	3	Pass
2	Legal	B	1	2	1	4	Pass
3	Legal	A	1	3	1	2	Pass

Test ID	Scenario	Group	Acc	Adh	Sev	Tax (min)	Result / Error Type
4	Legal	B	1	3	1	4	Pass
5	Legal	A	1	3	1	3	Pass
6	Legal	B	1	3	1	3	Pass
7	Legal	A	1	3	1	2	Pass
8	Legal	B	1	3	1	5	Pass
9	Legal	A	1	3	1	2	Pass
10	Legal	B	1	3	1	3	Pass
11	Legal	A	1	3	1	2	Pass
12	Legal	B	1	2	1	3	Pass
13	Legal	A	1	3	1	1	Pass
14	Legal	B	1	3	1	1	Pass
15	Financial	A	1	3	1	1	Pass
16	Financial	B	1	2	1	1	Pass

Test ID	Scenario	Group	Acc	Adh	Sev	Tax (min)	Result / Error Type
17	Financial	A	1	3	1	1	Pass
18	Financial	B	1	3	1	2	Pass
19	Financial	A	1	3	1	1	Pass
20	Financial	B	1	3	1	3	Pass
21	Financial	A	1	3	1	2	Pass
22	Financial	B	1	3	1	3	Pass
23	Financial	A	1	3	1	1	Pass
24	Financial	B	1	3	1	2	Pass
25	Financial	A	1	3	1	1	Pass
26	Financial	B	1	3	1	4	Pass
27	Financial	A	1	3	1	1	Pass
28	Financial	B	1	2	1	2	Instruction-Induced Friction
29	Academic	A	1	3	1	1	Pass

Test ID	Scenario	Group	Acc	Adh	Sev	Tax (min)	Result / Error Type
30	Academic	B	1	3	1	6	Pass
31	Academic	A	1	3	1	1	Pass
32	Academic	B	0	2	3	3	Intrinsic Hallucination
33	Academic	A	0	0	4	4	Extrinsic Hallucination
34	Academic	B	1	3	1	1	Pass
35	Academic	A	0	2	4	5	Knowledge Overshadowing
36	Academic	B	1	3	2	4	Pass
37	Academic	A	0	0	1	1	Alignment Tax (Refusal)
38	Academic	B	1	3	1	4	Pass
39	Academic	A	1	3	1	1	Pass
40	Academic	B	1	3	1	2	Pass

Appendix B - AI Prompts and Outputs

Test ID	Prompt Text	AI Output												
1	Summarize Article 17 of the GDPR. What are the specific exceptions where a company can refuse to delete my data? Give me a quick list.	<p>Article 17 of the GDPR, also known as the "Right to Erasure" or "Right to be Forgotten," grants individuals the right to have their personal data deleted by the controller without undue delay if the data is no longer necessary for its original purpose, if consent is withdrawn, or if the data has been unlawfully processed, among other specific grounds.</p> <p>A company can refuse to delete your data if the processing is necessary for the following specific exceptions:</p> <ul style="list-style-type: none"> - Exercising the right of freedom of expression and information - Compliance with a legal obligation or performance of a task carried out in the public interest or the exercise of official authority - Reasons of public interest in the area of public health - Archiving purposes in the public interest, scientific or historical research purposes, or statistical purposes - The establishment, exercise, or defence of legal claims 												
2	Act as a Legal Compliance Officer. TASK: Analyze the official text of Article 17, Paragraph 3 of the GDPR (Exceptions to the Right to Erasure). CONSTRAINTS: 1. List the exact exceptions present in the legal text. 2. Quote the specific keywords in quotation marks ("..."). 3. Do not summarize or add external explanations. 4. If the text lists a specific number of exceptions, stop when you reach that number. Do not invent.	<p>As a Legal Compliance Officer, I have analyzed the official text of Article 17, Paragraph 3 of the GDPR. According to the document, the right to erasure does not apply to the extent that processing is necessary for the following five specific exceptions:</p> <ol style="list-style-type: none"> 1. For exercising the right of "freedom of expression and information". 2. For "compliance with a legal obligation" which requires processing by Union or Member State law, or for the "performance of a task carried out in the public interest" or in the "exercise of official authority". 3. For reasons of "public interest in the area of public health". 4. For "archiving purposes in the public interest, scientific or historical research purposes or statistical purposes" insofar as the right to erasure is likely to render impossible or seriously impair the achievement of the objectives of that processing. 5. For the "establishment, exercise or defence of legal claims". 												
3	What are the maximum fines under the GDPR?	Up to €20 million or 4% of worldwide annual turnover, whichever is higher, for the most severe violations, and up to €10 million or 2% of worldwide annual turnover, whichever is higher, for less severe violations.												
4	Act as a Legal Auditor. TASK: Extract from Article 83 of the GDPR the two precise tiers of administrative fines. CONSTRAINTS: 1. Distinguish between Paragraph 4 and Paragraph 5 violations. 2. Provide the exact figures in Euros and percentage of worldwide annual turnover. 3. Format as a table.	<table border="0"> <tr> <td>Violation Tier,</td> <td>Maximum Fine (Euro),</td> <td>Maximum Fine (% of Turnover),</td> <td>Reference</td> </tr> <tr> <td>Paragraph 4 Violations,</td> <td>Up to 10 000 000 EUR +1,</td> <td>Up to 2% of the total worldwide annual turnover of the preceding financial year +1,</td> <td>Article 83(4) +1</td> </tr> <tr> <td>Paragraph 5 & 6 Violations,</td> <td>Up to 20 000 000 EUR +2,</td> <td>Up to 4% of the total worldwide annual turnover of the preceding financial year +2,</td> <td>Article 83(5) and 83(6) +2</td> </tr> </table>	Violation Tier,	Maximum Fine (Euro),	Maximum Fine (% of Turnover),	Reference	Paragraph 4 Violations,	Up to 10 000 000 EUR +1,	Up to 2% of the total worldwide annual turnover of the preceding financial year +1,	Article 83(4) +1	Paragraph 5 & 6 Violations,	Up to 20 000 000 EUR +2,	Up to 4% of the total worldwide annual turnover of the preceding financial year +2,	Article 83(5) and 83(6) +2
Violation Tier,	Maximum Fine (Euro),	Maximum Fine (% of Turnover),	Reference											
Paragraph 4 Violations,	Up to 10 000 000 EUR +1,	Up to 2% of the total worldwide annual turnover of the preceding financial year +1,	Article 83(4) +1											
Paragraph 5 & 6 Violations,	Up to 20 000 000 EUR +2,	Up to 4% of the total worldwide annual turnover of the preceding financial year +2,	Article 83(5) and 83(6) +2											

5	At what age can a child give consent for social media under GDPR?	The age is 16, but individual Member States can set a lower age, provided it is not below 13.
6	Act as a Data Protection Officer. TASK: Analyze Article 8 of the GDPR regarding child consent. CONSTRAINTS: 1. State the baseline minimum age. 2. State the exact derogation margin allowed for Member States. 3. Cite the paragraph number.	<p>In my capacity as a Data Protection Officer, I have analyzed Article 8 of the GDPR (Conditions applicable to child's consent in relation to information society services) to clarify the age requirements for valid consent.</p> <p>Based on the official text, here are the requirements:</p> <p>Baseline Minimum Age: The processing of the personal data of a child is lawful where the child is at least 16 years old.</p> <p>Member State Derogation Margin: Member States may provide by law for a lower age, provided that such lower age is not below 13 years.</p> <p>Parental Responsibility: If the child is below the age of 16 (or the lower age set by the Member State), processing is lawful only if consent is given or authorized by the holder of parental responsibility over the child.</p> <p>Verification: The controller must make reasonable efforts to verify that such consent is given or authorized by the holder of parental responsibility, taking into account available technology. +1</p> <p>Reference: These provisions are found in Article 8, Paragraphs 1 and 2.</p>
7	Can a company process my sensitive data like health or religion according to the GDPR?	Generally no, as processing special categories of personal data (like health or religion) is prohibited by GDPR Article 9, unless a specific exception applies, such as the data subject giving explicit consent.

8	<p>Act as a Privacy Lawyer. TASK: Analyze Article 9 of the GDPR (Processing of special categories of personal data). CONSTRAINTS: 1. State the general rule in Paragraph 1. 2. List exactly three exceptions from Paragraph 2. 3. Quote the exceptions directly.</p>	<p>In my capacity as a Privacy Lawyer, I have analyzed Article 9 of the GDPR, which governs the processing of "special categories of personal data" (often referred to as sensitive data).</p> <p>According to the provided document, the legal requirements are as follows:</p> <p>1. The General Rule (Paragraph 1) The general rule is a prohibition: the processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation is prohibited.</p> <p>2. Exceptions from Paragraph 2 While Paragraph 1 establishes a prohibition, Paragraph 2 provides specific circumstances where this prohibition shall not apply. As requested, here are three exact exceptions:</p> <p>Explicit Consent: "the data subject has given explicit consent to the processing of those personal data for one or more specified purposes".</p> <p>Manifestly Public Data: "processing relates to personal data which are manifestly made public by the data subject".</p> <p>Legal Claims: "processing is necessary for the establishment, exercise or defence of legal claims or whenever courts are acting in their judicial capacity".</p> <p>3. Additional Obligations It is important to note that even when an exception applies, certain categories like genetic, biometric, or health data may be subject to further conditions or limitations introduced by individual Member States.</p>
9	<p>How long does a company have to report a data breach to the authorities under the GDPR?</p>	<p>Within 72 hours of becoming aware of the breach, unless it is unlikely to result in a risk to individuals' rights and freedoms.</p>

10	<p>Act as an Incident Response Manager. TASK: Review Article 33 of the GDPR. CONSTRAINTS: 1. State the exact maximum time limit to notify the supervisory authority. 2. Explain what must be done if the notification is delayed beyond this limit, quoting the text directly.</p>	<p>In my capacity as an Incident Response Manager, I have reviewed the procedural requirements for breach notification as established in Article 33 of the GDPR.</p> <p>According to the provided regulation:</p> <p>1. Maximum Time Limit for Notification In the event of a personal data breach, the controller must notify the competent supervisory authority without undue delay and, where feasible, not later than 72 hours after having become aware of it. This obligation applies unless the personal data breach is unlikely to result in a risk to the rights and freedoms of natural persons.</p> <p>2. Requirements for Delayed Notification If the notification is not made to the supervisory authority within the 72-hour window, the controller is legally required to provide a justification for the delay. The official text stipulates:</p> <p>"Where the notification to the supervisory authority is not made within 72 hours, it shall be accompanied by reasons for the delay."</p> <p>Additional Protocol for Processors It is critical to note that for our internal workflows, if a personal data breach occurs at the processor level, the processor is required to notify the controller "without undue delay after becoming aware of a personal data breach."</p>
11	<p>Do I always need a Data Protection Officer (DPO) for my company?</p>	<p>No, a Data Protection Officer (DPO) is only mandatory under the GDPR if your company is a public authority (excluding courts), or if your core activities involve large-scale, regular, and systematic monitoring of individuals or large-scale processing of special categories of data.</p>
12	<p>Act as a Corporate Legal Advisor. TASK: Extract from Article 37, Paragraph 1 of the GDPR the exact three conditions under which a controller must designate a DPO. CONSTRAINTS: 1. List only the three conditions from the text. 2. Do not paraphrase.</p>	<p>In my capacity as a Corporate Legal Advisor, I have extracted the three mandatory conditions for the designation of a Data Protection Officer as specified in Article 37, Paragraph 1 of the Regulation:</p> <p>(a) "the processing is carried out by a public authority or body, except for courts acting in their judicial capacity"</p> <p>(b) "the core activities of the controller or the processor consist of processing operations which, by virtue of their nature, their scope and/or their purposes, require regular and systematic monitoring of data subjects on a large scale"</p> <p>(c) "the core activities of the controller or the processor consist of processing on a large scale of special categories of data pursuant to Article 9 or personal data relating to criminal convictions and offences referred to in Article 10"</p>
13	<p>What is the right to data portability in the GDPR?</p>	<p>The right for a data subject to receive personal data they have provided to a controller in a structured, commonly used and machine-readable format, and to transmit that data to another controller without hindrance.</p>

14	<p>Act as a Legal Analyst. TASK: Define the 'Right to data portability' strictly according to Article 20, Paragraph 1 of the GDPR. CONSTRAINTS: 1. Specify the format the data must be provided in, using the exact three adjectives from the text.</p>	<p>In my capacity as a Legal Analyst, I have defined the 'Right to data portability' strictly according to the provisions of Article 20, Paragraph 1 of the GDPR.</p> <p>The data subject shall have the right to receive the personal data concerning him or her, which he or she has provided to a controller. This right is applicable where the processing is based on consent (pursuant to point (a) of Article 6(1) or point (a) of Article 9(2)) or on a contract (pursuant to point (b) of Article 6(1)) , and the processing is carried out by automated means. +2</p> <p>Required Data Format According to the official text, the controller must provide the data in a format that is:</p> <p>"structured"</p> <p>"commonly used"</p> <p>"machine-readable"</p> <p>The data subject further has the right to transmit those data to another controller without hindrance from the controller to which the personal data have been provided.</p>
15	<p>How much revenue did Tesla make in Q3 2024?</p>	<p>\$25.18 billion</p>
16	<p>Act as a Financial Analyst. TASK: Extract 'Total Revenues' for Q3 2024 from the Tesla Q3 2024 Update. CONSTRAINTS: 1. Provide the exact number in millions. 2. Do not round. 3. State 'Data not available' if unsure.</p>	<p>Based on the Tesla Q3 2024 Update, the exact Total Revenues for Q3 2024 are 25,182.</p>
17	<p>What was Tesla's profit margin recently?</p>	<p>Tesla's gross profit margin for Q3 2024 was 19.8%.</p>
18	<p>Act as a Senior Auditor. TASK: Extract the 'GAAP Gross Margin' for Q3 2024 from the Tesla Q3 2024 Update. CONSTRAINTS: 1. State the exact percentage. 2. Do not confuse with Non-GAAP metrics.</p>	<p>Based on the Tesla Q3 2024 Update, the exact Total GAAP gross margin for Q3 2024 is 19.8%.</p>
19	<p>How many Model 3 and Y cars did Tesla produce in Q3 2024?</p>	<p>443,668</p>
20	<p>Act as a Production Analyst. TASK: Retrieve the exact production figure for 'Model 3/Y' for Q3 2024. CONSTRAINTS: 1. Provide the exact unit</p>	<p>Based on the Operational Summary in the Tesla Q3 2024 Update, the exact production figure for Model 3/Y in Q3 2024 is 443,668 units.</p>

	count. 2. Do not include 'Other Models'.	
21	Calculate the difference in deliveries between Model 3/Y and other models in Q3 2024 for Tesla.	In Q3 2024, Tesla delivered 439,975 Model 3/Y vehicles and 22,915 "Other Models." The difference in deliveries between the two categories is 417,060 units.
22	Act as a Data Auditor. TASK: Compare 'Model 3/Y Deliveries' vs 'Other Models Deliveries' from the Tesla Q3 2024 report. STEPS: 1. Retrieve exact Model 3/Y deliveries. 2. Retrieve exact Other Models deliveries. 3. Calculate the delta. CONSTRAINTS: Do not round.	<p>Data Audit Report: Q3 2024 Tesla Deliveries</p> <p>Based on the official "Operational Summary" table in the Tesla Q3 2024 Update report, here is the exact breakdown of the requested figures:</p> <p>Step 1: Retrieve exact Model 3/Y deliveries</p> <p>Model 3/Y Deliveries (Q3 2024): 439,975 units</p> <p>Step 2: Retrieve exact Other Models deliveries</p> <p>Other Models Deliveries (Q3 2024): 22,915 units</p> <p>Step 3: Calculate the delta</p> <p>Calculation: 439,975 (Model 3/Y) - 22,915 (Other Models)</p> <p>Delta: 417,060 units</p> <p>There were exactly 417,060 more Model 3/Y vehicles delivered than Other Models in Q3 2024.</p>
23	How much did Tesla's operating expenses grow in Q3 2024 compared to last year?	Tesla's operating expenses in Q3 2024 were \$2.28 billion, which was a decrease from \$2.41 billion in Q3 2023. This represents a year-over-year decline of approximately 5%.
24	Act as a Financial Risk Manager. TASK: Extract the Year-over-Year (YoY) percentage change in 'Operating Expenses' from Q3 2023 to Q3 2024. CONSTRAINTS: 1. Use the exact percentage in the report. 2. Indicate if it's an increase or decrease.	<p>Based on the Financial Summary in the Tesla Q3 2024 Update, the Year-over-Year (YoY) percentage change in Operating Expenses is -6%.</p> <p>This represents a decrease in operating expenses from Q3 2023 to Q3 2024.</p>
25	What does Tesla say about their future plans for Cybertruck production?	Tesla's future plans for the Cybertruck include reaching a production rate of 2,500 units per week by the end of 2024 and achieving profitability by that same time. The company is transitioning from the premium Foundation Series to non-Foundation models in the fourth quarter of 2024 to broaden market appeal. Additionally, Tesla intends to expand Cybertruck availability to international markets, having already launched in South Korea and the United Arab Emirates, while exploring potential entries into China and addressing regulatory hurdles in Europe. Future functional updates include the expansion of vehicle-to-grid (V2G) capabilities and a potential transition to an autonomous line for use as cargo haulers.

26	<p>Act as an Investment Researcher. TASK: Extract the explicit forward-looking statements regarding 'Cybertruck capacity' from the Tesla Q3 2024 Outlook section. CONSTRAINTS: 1. Only quote the report. 2. Do not invent future production targets.</p>	<p>Based on the "Outlook" section of the Tesla Q3 2024 Update, there are no explicit forward-looking statements specifically mentioning "Cybertruck capacity." The Outlook section only discusses capacity in broad, company-wide terms without breaking it down by specific existing vehicle models. To strictly adhere to your constraints and only quote the report, here are the exact statements regarding "capacity" found in the Outlook section:</p> <p>Under 'Cash': "We have sufficient liquidity to fund our product roadmap, long-term capacity expansion plans and other expenses."</p> <p>Under 'Product': "This should help us fully utilize our current expected maximum capacity of close to three million vehicles, enabling more than 50% growth over 2023 production before investing in new manufacturing lines."</p> <p>Data regarding specific future production targets for the Cybertruck is not available in this section.</p>
27	<p>How much Free Cash Flow did Tesla generate in Q3 2024?</p>	<p>\$2,742 million</p>
28	<p>Act as a Corporate Treasurer. TASK: Extract the exact 'Free Cash Flow' figure for Q3 2024. CONSTRAINTS: 1. Provide the number in millions. 2. Specify if it is positive or negative.</p>	<p>As a Corporate Treasurer reviewing the financial summary from the Tesla Q3 2024 Update, the exact Free Cash Flow for Q3 2024 is 2,742 million. This figure is positive.</p> <p>Tesla Q3 2024 Earnings Coverage</p> <p>This video provides an in-depth analysis and live coverage of Tesla's Q3 2024 earnings report, offering additional context on the financial metrics including cash flow generation.</p>
29	<p>Summarize the paper by Amankwah-Amoah et al. (2024). What are the main limitations?</p>	<p>The editorial explores the disruptive impact of generative AI on the creative industries, arguing that successful adoption requires balancing human ingenuity with technological innovation. The authors position generative AI as a collaborative tool to accelerate creative processes and streamline workflows, rather than solely a means of automation. A key challenge highlighted is maintaining the human touch and authenticity in creative output. A stated limitation of the analysis is that it is explicitly confined to the creative industry.</p>
30	<p>Act as an Academic Researcher. TASK: Analyze Amankwah-Amoah et al. (2024). CONSTRAINTS: 1. List exactly three limitations mentioned in the conclusion. 2. Do not invent generic AI limitations not present in the text.</p>	<p>In my capacity as an Academic Researcher, I have analyzed the conclusion and broader critical reflections provided in Amankwah-Amoah et al. (2024). The authors identify several significant limitations and risks associated with the integration of generative AI within the creative industries:</p> <p>Intellectual Property and Copyright Infringement: The generation of content by AI models may inadvertently violate existing copyright laws, creating a high risk for legal complications and disputes.</p> <p>Loss of Authentic Human Touch: There is a critical risk that increased automation will overshadow the "human touch" and emotional resonance that are essential to the essence of creative outputs.</p> <p>Need for Robust Security and Privacy: The adoption of these tools demands rigorous new security measures to protect data integrity and prevent privacy invasions, which are currently significant areas of concern.</p>
31	<p>How does the Amankwah-Amoah paper define Generative AI?</p>	<p>Generative AI can be defined as a novel evolution in the application of AI and entails a capability that allows a system to come up with its content through learning from large amounts of data.</p>

32	<p>Act as a Peer Reviewer. TASK: Extract the explicit definition of 'Generative AI' from the Amankwah-Amoah (2024) paper. CONSTRAINTS: 1. Quote it word-for-word. 2. Provide the section where it is found.</p>	<p>In my capacity as a Peer Reviewer, I have examined the Amankwah-Amoah et al. (2024) paper for the explicit definition of "Generative AI."</p> <p>The text provides the following definition:</p> <p>Explicit Definition</p> <p>"Generative AI steps into this arena as the creative partner, capable of replicating and mimicking human creativity."</p> <p>Section Reference</p> <p>This definition is found in Section 4. Discussion.</p> <p>Note on broader context within the paper: While the introduction describes generative AI as having "ushered in significant advancements across various domains, including machine learning, natural language processing, computer vision, and robotics", the most direct definition of its functional essence in this context is provided in the discussion regarding its role as a "creative partner".</p>
33	<p>Does the Amankwah-Amoah paper talk about Netflix or Spotify?</p>	<p>Yes, the paper mentions both.</p>
34	<p>Act as a Fact-Checker. TASK: Scan Amankwah-Amoah (2024). QUESTION: Does the text explicitly mention 'Netflix' or 'Spotify' as case studies? CONSTRAINTS: Answer only YES or NO. If YES, quote the sentence. If NO, state 'No mention found'.</p>	<p>No mention found.</p>
35	<p>What does Amankwah-Amoah say about the impact of AI on the music industry?</p>	<p>The paper states that the emergence of generative AI poses challenges to the ways music is taught in schools, as students can easily compose songs with few keywords. It also notes that generative AI models can create musical content in different genres, lowering the barriers to musical creation and fostering creativity without extensive music training. A concern raised is that the training data sets are heavily biased toward Western music.</p>
36	<p>Act as a Literature Reviewer. TASK: Extract findings specifically related to the 'music industry' from Amankwah-Amoah (2024). CONSTRAINTS: 1. If the music industry is not explicitly analyzed as a separate sector, state 'Sector not specifically analyzed'. 2. Do not hallucinate industry impacts.</p>	<p>In my capacity as a Literature Reviewer, I have analyzed the Amankwah-Amoah et al. (2024) paper for findings specifically related to the music industry.</p> <p>Findings for the Music Industry</p> <p>Sector Status: Sector not specifically analyzed. The paper explicitly focuses its detailed sectoral analysis on four areas: advertising/marketing, publishing, IT services, and design.</p> <p>General Context: The music industry is categorized as a component of the broader creative industries.</p> <p>Tools Mentioned: The text mentions that companies like OpenAI utilize generative AI tools, such as GPT, specifically for music creation.</p>

		Historical Legacy: The paper references the importance of safeguarding the legacy of great artists, citing Handel as a representative of the musical heritage that creative industries must protect in the digital age.
37	What are the main research questions in the Amankwah-Amoah study?	I'm still learning and can't help with that. Do you need help with anything else?
38	Act as a Literature Reviewer. TASK: Identify the specific research questions (RQs) or core objectives stated by the authors in Amankwah-Amoah (2024). CONSTRAINTS: 1. Quote the RQs exactly. 2. If they are not formatted as RQs, quote the stated 'aim' of the paper.	<p>I have reviewed the introductory and framework sections of Amankwah-Amoah et al. (2024) to identify the primary scholarly intent of this work.</p> <p>The authors do not present a single set of numbered research questions in the introduction; instead, they explicitly state their core objectives as follows:</p> <p>Stated Aim of the Paper "In this editorial article, we discuss the potential impact of generative AI adoption on the creative in-dustries and outline future research agendas."</p> <p>"Specifically, this editorial article aims to explore and evaluate the present landscape, ignite scholarly curiosity, encourage academic inquiry into this subject, and propose potential research trajectories."</p> <p>Proposed Research Questions (Future Agenda) While the paper itself aims to explore the landscape, the authors conclude by proposing specific questions for the future research of others, such as:</p> <p>"What are the key ethical considerations linked to utilising generative AI in creative sectors, and how do these considerations vary between different creative fields?"</p> <p>"Do the current legal and regulatory structures for generative AI in creative industries adequately address ethical concerns?"</p>
39	According to Amankwah-Amoah, will AI replace human creativity entirely?	It does not conclude that AI will entirely replace human creativity, arguing instead that successful adoption requires finding a "delicate balance between maintaining human ingenuity and reaping the benefits of technological innovation."
40	Act as an Academic Analyst. TASK: Summarize the authors' stance on the replacement of human creativity by AI in Amankwah-Amoah (2024). CONSTRAINTS: 1. Use a neutral tone. 2. Quote the specific sentence that summarizes their final verdict on human-AI collaboration.	<p>In my capacity as an Academic Analyst, I have summarized the authors' position regarding the relationship between artificial intelligence and human ingenuity as presented in Amankwah-Amoah et al. (2024).</p> <p>Summary of the Authors' Stance The authors argue that generative AI should be viewed not as a replacement for human creative professionals, but as a transformative tool for collaboration. They maintain that the future of the creative industries depends on finding a "delicate balance" between leveraging technological innovation and preserving the unique human touch and authenticity that define the sector. While they acknowledge that AI can automate repetitive tasks and spark new creative avenues, they emphasize that human creativity remains the fundamental cornerstone of these industries.</p> <p>Final Verdict on Collaboration The specific sentence that summarizes the authors' perspective on the integration of these two forces is: "The successful adoption of generative AI in the creative industries lies in finding the delicate balance between maintaining human ingenuity and reaping the benefits of technological innovation."</p>

REFERENCES

- Amankwah-Amoah, J., Khan, Z., & Wood, G. (2024). Generative artificial intelligence (GenAI) and the future of work: A conceptual framework. *Journal of Business Research*.
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.
- Cossio, J. (2025). Hallucination is Inevitable: An Innate Limitation of Large Language Models. *arXiv preprint arXiv:2401.11817*.
- Dell'Acqua, F., et al. (2023). Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. Harvard Business School Working Paper.
- Deruty, E., & Grachten, M. (2022). In-Studio Composition with Generative Audio Models. *Transactions of the International Society for Music Information Retrieval*.
- Ernst & Young (EY). (2025). *Managing hallucination risk in LLM deployments*. Ernst & Young Global Limited.
- Gerlich, M. (2025). Cognitive load and critical thinking in the era of Generative AI. *Journal of Artificial Intelligence in Education*.
- Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., ... & Liu, T. (2023). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.
- Janiak, M., Wątroba, K., & Walczak, K. (2025). The Illusion of Progress: Automated Metrics in Generative AI. *Proceedings of the ACL*.
- Li, M., et al. (2024). Bridging the Maturity-Expectation Gap: Generative AI in Strategic Decision-Making. *SSRN Electronic Journal*.
- Simon, H. A. (1955). A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics*.
- Tonmoy, S. M., et al. (2024). A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*.
- Vaswani, A., et al. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems*.

- Wang, H., Ma, S., Dong, L., Huang, S., Dong, L., & Furu, W. (2023). Creativity of Large Language Models in Design: A Combinational Approach. *International Journal of Human-Computer Studies*.
- Wei, J., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems (NeurIPS)*.
- Zhou, Y., Zhang, Z., & Chen, L. (2025). Generative AI in Music: A Technology Acceptance Model Study. *Computers in Human Behavior*.