Understanding Design Ideation with Vision-Language Models and Video-Based Design

Abstract

This study explores integration of Large Language Models (LLMs) and Vision-Language Models (VLMs) into design ideation of industrial design. Conducted at TU Delft, the research involved brainwriting and video-based design (VBD) methodologies. The primary aim was to legitimize and valiate context-injected LLMs and VLMs in supporting designers' search for inspiration through development of experiment framework. The study measured workload, user experience, acceptance of technology, divergent thinking capabilites and attitudes towards AI. It also preliminary analysed the results, focusing on qualitative insights.

Data was collected through surveys, interviews and eye-tracking data although the eye tracking data was excluded from analysis. The study found that while AI tools support ideation by generating diverse ideas and handling repetitive tasks, they need improvement for contextually relevant and accurate information. Designers expressed cautious optimism about AI's potential, emphasizing the need for human oversight to retain creativity and ensure context-aware assistance.

The research highlighted optimistic leaning opinions on AI integration, noting that current AI capabilities are not yet sufficient for design ideation demands. It emphasized the necessity for AI to act as a collaborative partner, preserving the designer's critical role in the creative process.

Keywords: Large Language Models (LLMs), Vision-Language Models (VLMs), Industrial Design Engineering, Design Ideation, Video-Based Design (VBD), Human-Computer Interaction (HCI), AI in Design Processes

TU Delft

Industrial Design Engineering –
Integrated Product Design

Andrija Stanković
Chair: Dr. E. Niforatos
Mentor: MSc Tianhao He

Table of Contents

1. Introduction

1.1. Project setting

This master thesis, undertaken at the Faculty of Industrial Design Engineering at TU Delft, is a component of a broader PhD research within the Department of Sustainable Design Engineering, specifically in the section of Knowledge and Intelligence Design (KinD). KInD is a human-centered and multidisciplinary research team that explores the relationship between design and the digital technologies driving the development of intelligent products, services, and systems. Within these domains, KInD explores questions related to bias, fairness and transparency in automated decisions. By integrating design with data science, Internet of Things, machine learning, human-computer interaction, spatial analysis, and crowd computing, KInD tries to engineer product-service systems (PSS) that are robust, intelligible, accessible, and inclusive.

Positioned at the intersection of context-injected Large Language Models (LLMs), Vision-Language Models (VLMs), design methodology, and video-based design*, this thesis delves into the area where artificial intelligence facilitates and aids decision-making. The project unfolded over a period of approximately 100 working days, with Appendix A presenting the overall Gantt chart of activities. Its goal is to explore and legitimize the innovative integration of VLMs and LLMs in order to enhance and support the inspiration-seeking process in industrial design engineering.

The activities in this thesis involved developing a theoretical and conceptual framework, organizing and conducting practical experimentation, recruiting participants, gathering and validating data, and evaluating the broad implications of results, with a focus on qualitative analysis. The project aimed to conduct rigorous experimentation, testing, and validation using design cases and videos.

For an overview of the project brief see Appendix B and for an insight into motivation to conduct this project see Appendix C.

**Video-based design, as defined for the context of this research, is an approach that integrates video into the design process, enabling a deeper, user-centered understanding by capturing real-life contexts and experiences to inform and enhance design outcomes ("Studying what people do," 2007)*

1.2. Opportunities

The opportunities cover a broad area and the list presented here should not be considered exhaustive. The following are some of the advantages that could enrich design knowledge and support effective usage of VLMs during ideation.

1. There is potential for VLMs and LLMs to expand the existing knowledge concerning the augmentation of the ideation stage in design processes by AI. While this thesis predominantly focuses on the inspiration aspect, we assume there is potential for extending the use of VLMs and LLMs to other stages of the design process where the search for inspiration is present.
2. The search for inspiration within design has been characterized as intuitive and unstructured (Gonçalves, 2016) - and LLMs and VLMs can structure and expand upon it. Context-injected combination of VLM and LLMs are one way of achieving it as they can provide a guided approach that could facilitate a more thorough exploration of the problem space, thereby enhancing the overall design process (Suh et al., 2023).
3. The usage of VLMs and LLMs in the design process can offer insights for ongoing AI development, particularly in understanding how designers utilize AI tools, thereby informing improvements in both user experience (UX), user interface (UI) and more broadly human-ai co-creation.

1.3. Key Findings

- VLMs need to be improved to provide more accurate and relevant information for divergent thinking process in design ideation. The study suggests that further development is needed to enhance the precision of AI-generated suggestions to better align with specific design contexts.

- Design students are optimistic but cautious about integration of VLMs in design ideation, emphasizing the need for human oversight to retain relevance and control over creative outputs so as not to lead to design fixation.

- Most of participating design students view atificial intelligence as as a collaborative tool that structures, organizes information and aids with providing general information and solving repetitive tasks.

- Transparency of AI sourcing is one of the most prominent concerns, with a need for clearer guidelines, control and accountability over provided information. Furthermore, there is a need for content output balance, as some participants were overwhelmed by the amount of generated content.

2. Background

2.1. Introduction

The integration of Large Language Models (LLMs) and Vision Language Models (VLMs) at the onset of the design process, particularly during the ideation phase, represents a domain not substantially explored. With recent advancements in artificial intelligence, specifically through the emergence of generative LLMs like ChatGPT-3 and GPT-4, there is a looming shift across multiple disciplines. These models are not only automating routine tasks but can also enhance abstract reasoning and problem-solving capabilities, which are crucial in industrial design—a field that relies heavily on breaking design fixation and fostering divergent and convergent thinking (Kalyan, 2023; Naveed et al., 2023; Woelfel et al., 2013; Zhou et al., 2021).

This novel integration could augment parts of the creative process that were traditionally performed by human intelligence, thereby adding to design discourse and potentially increasing efficiency in prototyping, problem-solving, and user engagement (Makatura et al., 2023). Design ideation is typically iterative and flexible, usually involving the rapid generation of many ideas to foster innovative solutions. Therefore, it is crucial to understand the role of Vision Language Models (VLMs), and AI models more broadly, in interpreting video content. Videos are integral at various stages of the design process, serving as a multisensory medium that enhances communication within design teams and with external stakeholders, while adding value to concept presentation and usability testing - among other use cases (Gonçalves, 2016; Halskov & Nielsen, 2006; Isaacs & Tang, 1993; Moore & Buur, 2005; "Studying what people do," 2007).

However, the traditional methods of analyzing and producing these videos are often labor-intensive and resource-heavy, presenting significant challenges, especially for smaller teams or in scenarios with limited production capabilities (Zimmerman, 2005; "Studying what people do," 2007). Additionally, there is a risk of these persuasive videos overshadowing critical design issues, potentially leading to a misinterpretation of user needs or an undue emphasis on visual appeal of media at the expense of critical examination of usability and functionality (Batalas et al., 2012).

By exploring the use of LLMs in streamlining the analysis and integration of video content in the design process, this research aimed to explore abovementioned drawbacks and deepen understanding of their impact on user experience and overall utility. This background section therefore explores the current advancements of large language models and related areas of videos and inspiration, identifying the relevant literature review areas to define a research gap, problem statement and research questions.

2.2. Literature review

The primary goal of the literature review was to establish a solid foundation for the thesis and ensure a thorough grounding before proceeding with user testing. This framework also informed the formulation of research questions and facilitated the acquisition of knowledge regarding the phenomena and technologies integral to the research project. Among the abovementioned, such an understanding was necessary to identify potential opportunities and limitations that might arise during experiment session.

To achieve this goal, the literature review was segmented into major domains of exploration. These domains included understanding how designers interact with technology, specifically how, why, and when they use it in the design process—focusing on areas such as the use of Large Language Models (LLMs) in the design process, video in the design process, design theory and design methodology. Furthermore, it covered inspiration in the design process and the complexity of design tasks to facilitate a better understanding of the terms and definitions used.
Each section of the literature review served as a building block for understanding and developing a framework that would guide the development of experiment procedure. The chapters in literature review begin with brief introductions specifying the goals and aims of the literature research chapter, then proceed to listing the findings considered most relevant for the study.

Specifically, the review covered:

- **LLMs in Design**: To understand the previous findings on the dynamics of human-AI interaction within the design process.

- **Inspiration in Design**: To define and contextualize inspiration, including its importance to designers, its relevance during the ideation stage, the sources from which it is drawn, and how creative ideas are measured and assessed by previous studies.

- **Video in Design Process**: To assess the viability and scope of video usage within the design process.

- **Complexity of Design Tasks**: To explore the potential correlation between design task complexity, as understood by difficulty of defining the task, and video content complexity, as understood by value of video bitrate.

- **Video Complexity**: To develop a framework for the experiment session, we aim to investigate how participants and large language models (LLMs) react to increasingly complex design tasks, characterized by more intricate video content. This approach mirrors how some designers understand and interpret design task complexity.

2.2.1. LLMs in Design Process

In recent explorations of the integration of Large Language Models (LLMs) in the design process, a significant emphasis has been placed on the role of AI as a collaborative partner. In general, there is a consensus that AI is beneficial and can aid designers if employed and used responsibly. Below are some of the common findings:

- In his autoethnographic study Asadi (2023) emphasizes the potential of AI in enhancing the creative process, advocating for its active participation beyond the current approach.

- The necessity of improving interaction paradigms that facilitate both divergent and convergent thinking is a recurrent theme in literature. Suh et al. (2023), Brophy (2001), and Sternberg (2018) collectively argue for the development of frameworks that mitigate the premature convergence of ideas, promote the exploration of the design space, and streamline the ideation process.

- Ding & Chan (2023) underscore the significance of design space exploration, advocating for LLMs to serve more as facilitators in the workflow rather than being merely generators of isolated artifacts.

- The discourse further extends into the realm of user engagement and co-creation, highlighting the evolving relationship between humans and AI. The collaborative approach outlined by Suh et al. (2023), which builds on a vision originally envisioned by Licklider (1960), suggests that human creativity is amplified by technology's ability to execute tasks and generate insights.

On the other hand, among the more emphasized benefits and insights, the reviewed literature also exposed some of the drawbacks of using LLMs. For example, while addressing the practical challenges of integrating LLMs into the design process, Suh et al. (2023) critique the existing interaction models for their inability to fully harness AI's creative potential, often leading to rapid idea fixation among less experienced users. The literature calls for creativity support tools that not only facilitate a balance between divergent and convergent thinking (Brophy, 2001; Sternberg, 2018) but also empower users through prompt engineering and user-controlled AI-generated outputs (Ding & Chan, 2023). Secondly, concerns regarding the current limitations of AI-assisted design processes are mentioned by several studies. The potential for users to be overwhelmed by the sheer volume of AI-generated options (Hai Dang et al., 2022) and the need for improved interaction paradigms are highlighted as areas for future research (Suh et al., 2023; Ding & Chan, 2023). These reflections emphasize a consensus on the necessity for ongoing evolution in the design and implementation of LLMs within creative and design workflows, aiming to better support human-AI co-creation and enhance the creative capabilities of designers.

2.2.2. Video in Design Process

Video in design can be used in many different areas, and it is difficult to create an extensive list of the potential use cases of video in the design process. However, it is possible to aggregate the common usage tendencies into domains which can serve as a guiding point to understanding of how to employ videos in design effectively. In general, findings indicate that video is a valuable tool in the design process, particularly for finding inspiration. Specifically, it is effectively used for interaction analysis, usability testing, and design ethnography. These applications, among others which are not considered relevant in the context of this study (see Methodology), are drawn from the book on the usage of video "Studying What People Do" (2007) by Buur and Ylirisku.

Considering the book, video-based design (VBD) is an approach in the design process, in which it facilitates a deeper understanding of users and a more interactive approach to design development. To draw a definition from the context of the book, it could be defined as an approach that integrates video into the design process, enabling a deeper, user-centered understanding by capturing real-life contexts and experiences to inform and enhance design outcomes ("Studying what people do," 2007)

Other studies have noted different highlights of video usage in design processes. For example, Harrison, Minneman, Stults, and Weber (1989) pointed out the role of video as a design medium, emphasizing its capacity to support ambiguous communication. They argue that this ambiguity is allowing for ideas to be open to interpretation and refinement among designers and that it has been found to foster a collaborative environment where concepts can evolve through negotiation and dialogue.

Moreover, a video's ability to convey non-verbal cues such as gestures and expressions enables a nuanced analysis of face-to-face interactions like negotiation and visual communication, which often lack in other stimuli such as text or images. This supports the engagement of designers with stakeholders and provides support in both documenting and developing research within design processes.

Expanding on these insights, Moore and Buur (2005) explored video's roles as both a descriptive and creative tool in design. They noted that video provides context for framing problems and solutions and triggers multiple points of focus among viewers. This capability of video to make user experiences explicit aids designers in reflecting upon and reframing design issues, thereby enhancing discussions about the problem and solution spaces.

In a more structured exploration, "Studying What People Do" (2007) elaborates on video's role in facilitating a conscious design process. The book describes how video can be used to build and share conceptions collaboratively, helping designers to reflect upon situations and consider more possibilities. This ability emphasizes video's utility in participatory design, usability studies, and interaction analysis among others, where it serves as a valuable tool for documenting real-life interactions or aiding scenario-based design.

2.2.3. Inspiration in Design Process

Inspiration in the design process is one of the prominent aspects of experience that influences the initiation and progression of creative activities. Eckert and Stacey (2000) extensively discuss how sources of inspiration are integral to setting the context for new designs and informing their creation. They articulate that inspiration is not merely a catalyst for creativity but also a critical component in the communication and development of design ideas. According to them, sources of inspiration provide a contextual framework that allows designers to communicate and position their work effectively, sparking creativity, offering new perspectives, and triggering the generation of original ideas. Moreover, Eckert and Stacey highlight the subjective nature of inspiration, which can come from a wide array of sources, whether sensed or abstract.

In a broader discussion on the decoding of designers' inspiration processes, Gonçalves (2016) extends the understanding of inspiration beyond the ideation stage into later stages where it is still searched for by designers, while at the same time asserting that designers maintain a limited range of external stimuli preferences during that entire time. Both design students and professionals often regard the search for inspiration as pivotal during idea generation, with a preference for visual stimuli like images, objects, and textual sources which offer structured yet multiple interpretations to encourage creativity.

The literature also addresses the potential pitfalls of overly focusing on domain-specific knowledge, which can lead to design fixation, while too general or distant information may prevent designers from effectively addressing the problem at hand (Plucker & Beghetto, 2004). Lawson and Dorst (2009) suggest that a designer's level of expertise might vary with the specific problem being tackled, influencing their ability to identify the most adequate information for the problem thereby influencing the level of inspiration.

Eastman (2001) and Cross (2004) discuss how external stimuli, which can be pictorial, textual, audible, or tactile, play a role in the design process. Experienced designers, as opposed to novices, are often better prepared to analyze the problem comprehensively and search for helpful information. Cross further notes that successful designers are proactive in framing problems and directing the search for solutions, a process termed 'problem framing.'

Furthermore, the inspiration process itself, as Gonçalves (2016) points out, often lacks reflection among designers but is a cyclic activity engaged in multiple times throughout the design process. This process could be enhanced by a more reflective approach to boost ideas development. Gonçalves also emphasizes the potential for developing computational tools to help designers efficiently find relevant stimuli that are semantically distant from the problem domain, fitting different phases of the design process.

As an addition, Setchi and Bouchard (2010) define inspiration as a multifaceted phenomenon whereby designers absorb and reinterpret existing ideas, forms, and concepts. This process not only serves as a guiding principle and catalyst for creativity but is also fundamentally subjective, influenced by designers' individual experiences, cultural backgrounds, and personal interests. The subjectivity of inspiration allows it to play a crucial role in accelerating the ideation phase, thereby enabling designers to explore a broader array of possibilities. According to Setchi and Bouchard, designers interact with sources of inspiration through various methods such as research, observation, and engagement with their surroundings, employing tools like mood boards, sketches, and digital libraries.

Building on these insights, Dazkir, Mower, Reddy-Best, and Pedersen (2013) examine how design students engage with the inspiration process, noting significant differences in the engagement levels between self-selected and assigned sources of inspiration. Their study finds that students who select their own sources within a given assignment feel more personally connected to the task and exhibit higher initial engagement. In contrast, those who are assigned sources, while struggling with initial engagement, often achieve a broader understanding of cultural contexts in design.

These findings show the overarching tendencies with which designers and design students interact with and derive benefit from different sources of inspiration, showing that the search for inspiration is a complex process influenced by engagement, personal interest and the breadth of exploration.

2.2.4. Complexity of Design Tasks

Previous studies on managing complexity of design tasks generally show the need for varied approaches to manage and exploit complexity to enhance creativity and efficiency in design processes.

Kersten, Diehl, and Engelen (2018) discuss the complexity of design tasks, emphasizing that a strategic approach to task clarification can leverage complexity to foster creativity. They introduce the Context Variation by Design (CVD) method, which advocates for varying complexity early in the design process to open a broader range of solutions and creative paths. This approach contrasts with the premature simplification of tasks or the narrow focus on a single context or user group. CVD encourages a design process that transcends the limitations of a single context to uncover "shared insights" through integrating perspectives from varied contexts, thus enriching the conceptual design space with multiple media types and higher order patterns from lower order chaos.

ElMaraghy et al. (2012) highlight the increasing complexity in product development as a critical concern for contemporary businesses. They note that companies with an edge in product development are those that can efficiently bring new products to market, utilizing fewer resources and delivering superior design quality, thus offering better returns to shareholders and contributing positively to the economy.

Thomas and Izatt (2003) present a taxonomy of engineering design tasks based on the amount of freedom and complexity involved. They suggest that higher levels of design freedom in educational settings can enhance the learning experience by encouraging creative thinking and requiring a balance of technical knowledge and creative problem-solving skills. The level of designer contribution, with fewer constraints, directly increases the complexity of the design task.

Chen (2016) explores the learning problems and resource usage of undergraduate industrial design students, identifying concept generation as one of the most challenging aspects of design tasks. The difficulty in finding inspiration and executing lateral (divergent) thinking to develop a diverse set of concepts often leads to a "bottleneck for ideation." Chen suggests that addressing the complexity of design tasks should focus on enhancing the personal capabilities and thinking styles of students, as the difficulty is largely based on the degree to which tasks challenge the students' personal capabilities and resourcefulness.

2.2.5. Video Complexity

To enhance our understanding of how designers and AI interact and correspond to various task difficulties in video-based design (VBD), an investigation into the methods by which individuals perceive the content complexity of videos was carried out. This exploration reviewed previous studies that have examined video complexity, though it is important to note that these studies generally focused on complexity for purposes other than determining perceptual complexity.

The approaches to studying video complexity have typically included both computational and perceptual aspects. These methodologies aimed to improve video content retrieval and summarization techniques, optimize processing and encoding strategies, and enhance the accessibility of content. Such enhancements were facilitated through the development and application of various taxonomies and measurement techniques.

In the following paragraphs, we will provide several examples of how these methods have been applied, illustrating the breadth of research in this area and its relevance to understanding the interaction between designers and AI in managing complex video content within the design process.

- Several studies have adopted audio-visual cues combined with cognitive and structural information to enhance classification accuracy. Notably, Rouvier et al. (2009) utilized acoustic features like instability and space characterization through SVM classifiers for audio-based video genre identification. Similarly, Ekenel et al. (2010) integrated multiple sensory cues to classify various video genres with high effectiveness.

- Shamsi et al. (2019) implemented features such as spectral flux and shot boundary analysis to categorize videos with high precision. Moreover, dynamics of video content, such as foreground and background motion, have been explored as a classification basis by Roach et al. (2001), underscoring the utility of motion analysis in genre identification.

- Semantic and contextual analyses also significantly contribute to video complexity understanding. Assari et al. (2014) utilized semantic concept co-occurrences, enhancing classification through a mathematical approach. In the realm of deep learning, Patil et al. (2021) combined CNNs and RNNs, showcasing improved performance via keyframe extraction methods.

- Affective analysis by Zhao et al. (2013), which uses viewer facial expressions to recommend and classify videos.

- Social metadata usage also emerges as a novel classifier in the studies by Yew & Shamma (2011), revealing the potential of social interactions in genre determination. Furthermore, the contextual embedding of videos in educational settings, as explored by Ramesh et al. (2020), highlights the application of video complexity analysis in pedagogical environments.

2.2.6. Advancements in Large Language Models (LLMs) and Vision-Language Models (VLMs) for Industrial Design

Large Language Models, such as GPT-4, have demonstrated capabilities in understanding and generating human-like text, which has implications for automating routine tasks and enhancing problem-solving in various domains (Zhou et al., 2023). Vision-Language Models, on the other hand, combine the processing of visual data with textual analysis, allowing for a better understanding of user interactions that involve visual contexts. The evolution of these models is progressing from basic text and image processing to complex, context-aware systems (Li et al., 2023; Hu et al., 2023) which indicates that they could in near future interpret and suggest design modifications.

In industrial design, these models offer an opportunity for augmenting the creative process. Usually, designers rely on a mix of intuitive and structured methods to gather inspiration and define design problems (Gonçalves, 2016). The integration of LLMs and VLMs could structure and streamline these methods, making the process more efficient and comprehensive. By analyzing design-related data from both textual descriptions and visual content, these models can uncover subtle patterns and insights that may not be immediately apparent to human designers.

Furthermore, the integration of Large Language Models (LLMs) and Vision-Language Models (VLMs) has also notably advanced. As with Artificial Intelligence in general, these innovations enhanced interactions between digital and physical realms, facilitating complex automations and interactions. Below are some examples:

- The evolution of 3D-LLMs has expanded the capabilities of LLMs into the three-dimensional domain, allowing for interactions with 3D point clouds. This advancement has helped tasks such as 3D question answering and complex navigation, which are increasingly relevant in industrial design. Such technologies have shown the ability to enhance spatial understanding and facilitate a more dynamic engagement with industrial models, proving beneficial for both designing and manipulating intricate structures (Hong et al., 2023).

- The development of the MMICL framework has significantly improved the efficiency of VLMs in handling complex multimodal prompts (Zhao et al., 2023). This capability can be considered important in industrial design, where the integration of diverse data types and sources is used very often. The MMICL framework has been found useful by making faster and more precise design decisions by analyzing and integrating varied information.

- The CoVLM model enhances the compositional reasoning capabilities within LLMs, allowing for a better synthesis of textual and visual data. This model facilitates a more accurate design process by dynamically composing visual entities and relationships from textual instructions, thereby ensuring a precise alignment between design descriptions and visual data (Li et al., 2023). This is particularly critical in fields like industrial design where exactness in design translation is required.

- BLIVA has improved the handling of text-rich visual information. This model can understand images containing embedded texts, which is crucial for parsing detailed annotations within industrial designs. BLIVA's capabilities suggest substantial improvements in performance for tasks reliant on annotated visual data, enhancing both the interpretation and utility of such data in complex design scenarios (Hu et al., 2023).

- The development of NavGPT highlights the enhanced reasoning abilities of LLMs within contexts of navigation and spatial reasoning. This model can be particularly valuable in industrial design for optimizing layouts and planning scenarios, demonstrating its potential to significantly contribute to spatial analysis and design planning in complex environments (Zhou et al., 2023).

The advancements in LLMs and VLMs can enhance the process of industrial design, promoting multimodal interaction, improved spatial and compositional reasoning, and better integration of textual and visual data. These developments can open new opportunities for further research and innovation in the integration of AI technologies with physical and visual design elements.

2.3. Research Gap

Despite the current advancements in integration, research and progress of VLMs and LLMs, their application in ideation phase of industrial design process is not well explored in the current literature. There is a clear gap in understanding how these technologies can be specifically tailored to enhance the ideation and conceptualization phases of design. Furthermore, given the variety of design ideation methods available, each facilitating idea generation in slightly different ways, it is important to investigate how artificial intelligence interacts with designers during the ideation process. Therefore, the thesis aims to bridge the knowledge gap by examining the impact of artificial intelligence on the design process, particularly focusing on their role in enhancing the inspiration search phase.

More specifically, it also presents an opportunity for the design community to use the extensive knowledge and cognitive capabilities of LLMs for innovative design thinking and problem-solving. By leveraging LLM's knowledge base, designers can better understand the complexity of defining design problems; spot opportunities towards an improved solution space and expand their ideation capabilities.

2.4. Problem Statement

Investigate and set up an experiment testing framework on how combining Video-based Large Models (VLMs) and Design Context-injected Large Language Models (LLMs) enhances the design process initial ideation stage through experiment with design students and/or professionals, aiming to enrich design methodologies, theories, and HCI with top-level analysis of data focusing on qualitative exploration.

2.5. Research Questions

Based on the findings from the literature, the study focused on measuring both qualitative and quantitative aspects.

We seek to answer the abovementioned through the below research questions. For structure and clarity, each research question is accompanied by the rationale from the literature review for its development and afterwards proceeds to detail how question will be assessed.

RQ1. How do different levels of visual complexity of videos impact designers' divergent thinking when using a context-injected LLM?

RQ1 literature review rationale: The integration of Large Language Models (LLMs) into the design process, as highlighted in the literature, explores how AI can enhance creativity by serving as a collaborative partner while emphasizing the need for further understanding (Asadi, 2023; Suh et al., 2023). Furthermore, the review points out the importance of differing views on design task complexity and their management (Kersten, Diehl, and Engelen, 2018) and notes the specific role of video in providing nuanced, complex stimuli that can trigger divergent thinking in design (Harrison et al., 1989). Since the definition of design task complexity and visual complexity are still quite ambiguous, this question aims to provide more structure into the ambiguity of interpretations by defining a foundation based on which the impact of AI regarding design task complexity should be assessed.

Previous research of visual complexity of videos has mostly centered on the classification of videos into various categories and taxonomies. While substantial efforts have been directed towards reducing the computational load, enhancing the accuracy of classification algorithms, machine learning, optimizing video processing and other methods (Chang et al., 2007; Damnjanovic & Trow, 2023; Patil et al., 2021; Shamsi et al., 2019; Shyu et al., 2008) a notable gap exists in the literature regarding the categorization of videos based on perceptual visual complexity based on specifically the content of the videos.

Therefore, the study needs to establish a baseline understanding of perceptual complexity, and it does so by correlating it with video bitrate. The points below provide the definitions used for the thesis's purpose.

- **Video bitrate** refers to the rate of data encoded within a video stream. It quantifies the amount of information transmitted per second from the video source (such as a camera or encoder) to the destination (such as an online platform or viewer's device) (Heckmann, 2023).

- **Visual complexity**, as defined by Alghamdi, E., Velloso, E., & Gruba, P. (2021), corresponds to the challenges encountered in describing visual stimuli, which can be attributed to factors such as visual clutter, the density of edges within video frames, colorfulness, structural variability, and the frequency of motion.

Abovementioned visual complexity factors, indicative of higher video bitrates, serve as the basis for classifying video complexity into low and high categories, with a bitrate differential exceeding 1000 kilobits per second considered high enough for categorization purposes.

With the lack of previous helpful studies of perceptual video categorization, this differential was chosen for the following reasons:

- plays a major role in streaming quality (Menon et al., 2022; Borges et al., 2024)
- shows a correlation with perceptual and content complexity (Damnjanovic & Trow, 2023; Duan et al., 2020; Green Video Complexity Analysis for Efficient Encoding in Adaptive Video Streaming, n.d.; Korhonen & Reiter, 2009; Hines et al., 2014; Peng et al., 2013; Xu et al., 2014).

For a better understanding, a combination of bitrate analysis with subjective evaluations and other visual complexity indices might be necessary. In summary, the proposed baseline is deemed a feasible framework for assessing visual complexity in videos and is within the study's scope.

The hypothesis is that incorporating context-aware Language Learning Models (LLMs) into the design process will enhance designers' capacity for divergent thinking. Divergent thinking is defined as the capability to devise multiple alternative solutions to a given problem (Guilford, 1950). It is anticipated in practice that this mode of thought enables designers to conceive of a wider range of concepts with higher originality (What Is Divergent Thinking?, 2024; Guilford, 1950).

To investigate this hypothesis, the study adopted brainwriting as the primary method of ideation (Dam & Siang, 2024). The emphasis and choice of the brainwriting method stems from its ability to enable designers to generate multiple diverse ideas quickly, which is essential when exploring novel concepts at the beginning of the design process - while at the same time offering flexibility (What Is Divergent Thinking?, 2024; What Is Brainwriting, 2024) within the limited user testing timeframe.

As for the evaluation of generated ideas within this ideation method, three criteria were used: fluency, flexibility, and originality (Guilford, J.,1967). For the overview of how these ideas were measured see chapter Data Analysis.

RQ2. How do designers perceive the influence of Large Language Models (LLMs) on their creative process, specifically in terms of finding inspiration?

RQ2 literature review rationale: As highlighted by Asadi (2023) and further supported by Suh et al. (2023), Large Language Models (LLMs) are considered beneficial tools that can substantially influence the creative processes of designers. This question aims to explore designers' subjective perceptions of LLMs within their workflows, specifically concentrating on the stage of finding inspiration—a crucial element of the creative design process as discussed by Eckert and Stacey (2000) and Gonçalves (2016). While previous studies generally present an optimistic view of AI's role in creative domains, they also express concerns about the existing limitations of AI-assisted design processes. The possibility that users might be overwhelmed by the sheer volume of AI-generated options has been noted (Hai Dang et al., 2022), alongside the need for improved interaction methodologies. These concerns are identified as areas requiring further research (Suh et al., 2023; Ding & Chan, 2023). Research indicates that there is an essential need for continual evolution of understanding the implementation of LLMs within design to better support human-AI co-creation and enhance the ideation phase.

To gain insight into how designers perceive the influence of a context injected LLM (within video-based design) on their creative process we will employ a structured comparative interview between two subject groups.

RQ3. How does the integration of video-based design and LLMs affect the mental workload of designers during ideation stage, specifically when they are searching for inspiration?

RQ3 literature review rationale: Research shows that concurrent cognitive load can significantly impact decision-making across various contexts. Specifically, Dewitte, Pandelaere, Briers, and Warlop (2005) found that increased cognitive load can cause individuals to rely more heavily on readily available information when making consumer decisions. Similarly, Whitney, Rinehart, and Hinson (2008) observed that higher cognitive loads lead individuals to engage in less risky behaviors during decision-making tasks.

In the context of design, the integration of video-based design has been recognized for its potential to deepen insights into user behaviors and interactions, enhancing the overall design process (Moore & Buur, 2005). When combined with Large Language Models (LLMs), this integration could increase it by introducing information overload (Hai Dang et al., 2022). This research question seeks to examine how the integration of video-based design and LLMs affects the mental workload of designers.

While, to our knowledge, there are no direct studies on how LLMs specifically influence designers' mental workload, we hypothesize that the integration of video-based design and LLMs will follow the findings from previous studies. Utilizing a combination of subjective methods, such as questionnaires (weighted NASA-TLX) (The NASA TLX Tool: Task Load Index, n.d.), and objective measures (eye-tracking), we seek to facilitate an assessment of cognitive load.

RQ4. How does LLM affect designers' user experience (UX) and acceptance of technology in video-based design (VBD) during search for inspiration?

RQ4 literature review rationale: This question derives from the broader discourse on user engagement and co-creation in the integration of technology and AI within creative workflows (Licklider, 1960; Suh et al., 2023). By focusing on user experience and technology acceptance, this research question aims to uncover the broader implications of using advanced AI tools like LLMs within the specific context of video-based design, particularly how these tools influence the overall experience and acceptance of new technologies in creative domains. Given the increasing reliance on AI tools like LLMs in various stages of the design process, understanding their impact on UX is crucial for optimizing their application and enhancing designers' creativity.

To quantitatively measure the effects of LLMs on UX in VBD, this study employs the User Experience Questionnaire (UEQ) (User Experience Questionnaire (UEQ), n.d.). Conversely, to measure acceptance of technology, the study employs a UTAUT questionnaire ("User Acceptance of Information Technology: Toward a Unified View on JSTOR," n.d.). We hypothesize that the inclusion of LLM will enhance UX, that it will be positively perceived as enhancing video-based design and that the technology will be accepted by participants.

3. Methodology

3.1. Introduction

After describing the research questions and framing the thesis the following section will provide an overview of the database and prototype, explain the development of the experiment, provide an overview of measured constructs and detail the procedure.

As mentioned, previous sections have identified a notable gap in the existing body of knowledge concerning how Large Language Models (LLMs) influence the ideation phase in design processes. This gap is not unexpected; the integration of LLMs with Video Learning Models (VLMs) represents a new technology that has not yet achieved broad acceptance or application.

The literature review revealed that inspiration within the design process is multifaceted, with video being a versatile tool primarily employed to analyze user behavior. Furthermore, the complexity of design tasks has been shown to be context-dependent, influenced by factors such as the designer's prior knowledge, task definition, experience, level of engagement, and the freedom to choose their exploratory tasks.

From this foundation, the research identified specific use cases for video in design, which have informed the criteria for video selection in this study. Despite the lack of satisfactory methods for categorizing videos by content complexity found in the literature, this study proposed a baseline categorization for this research.

Given the constraints of the available timeframe, an effective approach to testing was needed. The videos selected from the databases must be relevant to designers and must provide enough insight and context to facilitate idea generation. The selection process thus considers several critical factors that align with how designers traditionally utilize videos in the design process, ensuring that the selected content is optimally suited for evaluating the impact of LLMs on design ideation.

The system was deployed on a local computer within a university laboratory. The deployment of the system on a local computer within a university lab setting was done to bolster privacy and data security. This localized setup ensured that all interactions and conversation histories are stored and managed within a secure, controlled environment.

3.2. Prototype

The prototype employed was the context-injected generative AI. This prototype integrated two AI components: the BLIP model for video understanding and the GPT-4 model equipped with a Retrieval-Augmented Generation (RAG) mechanism for controlled information retrieval.

The prototype works by transforming input videos into individual frames. The BLIP (Bootstrapped Language-Image Pre-training) model, a vision-and-language representation learning system, was employed to analyze these frames. In it each frame is transcribed into text descriptions with timestamps capturing both the visual elements and their temporal dynamics.

Once the video was transcribed, the text descriptions were combined with a system generated prompt. This primed the GPT-4 model, which then processes queries submitted by participants. However, GPT-4, enhanced with the RAG mechanism, retrieved information by accessing a controlled repository of design books. This controlled access was employed to manage the balance between providing sufficient inspiration, avoiding design fixation by limiting overreliance on existing designs and providing precise and professional answers.

The RAG mechanism integrated retrieved information with the generative capabilities of GPT-4 to produce responses.

Figure 1 explains the working principle of the prototype

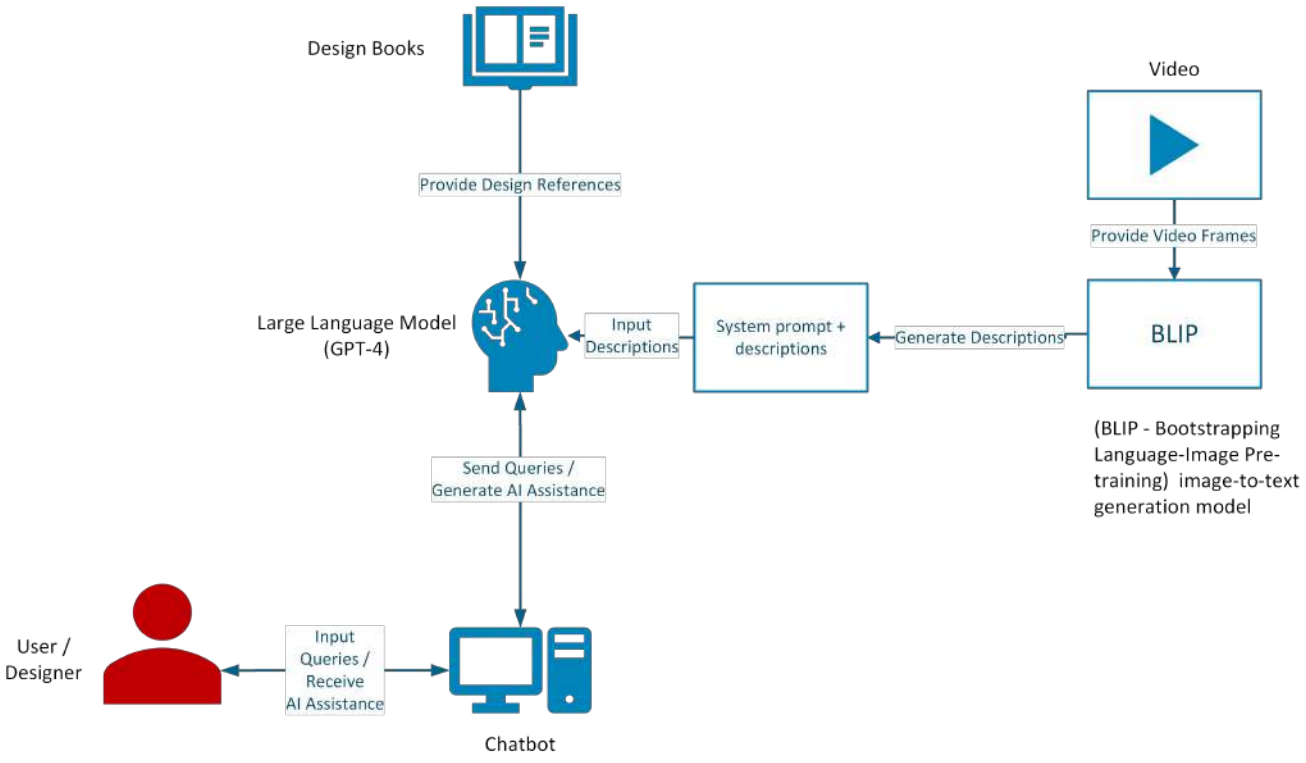
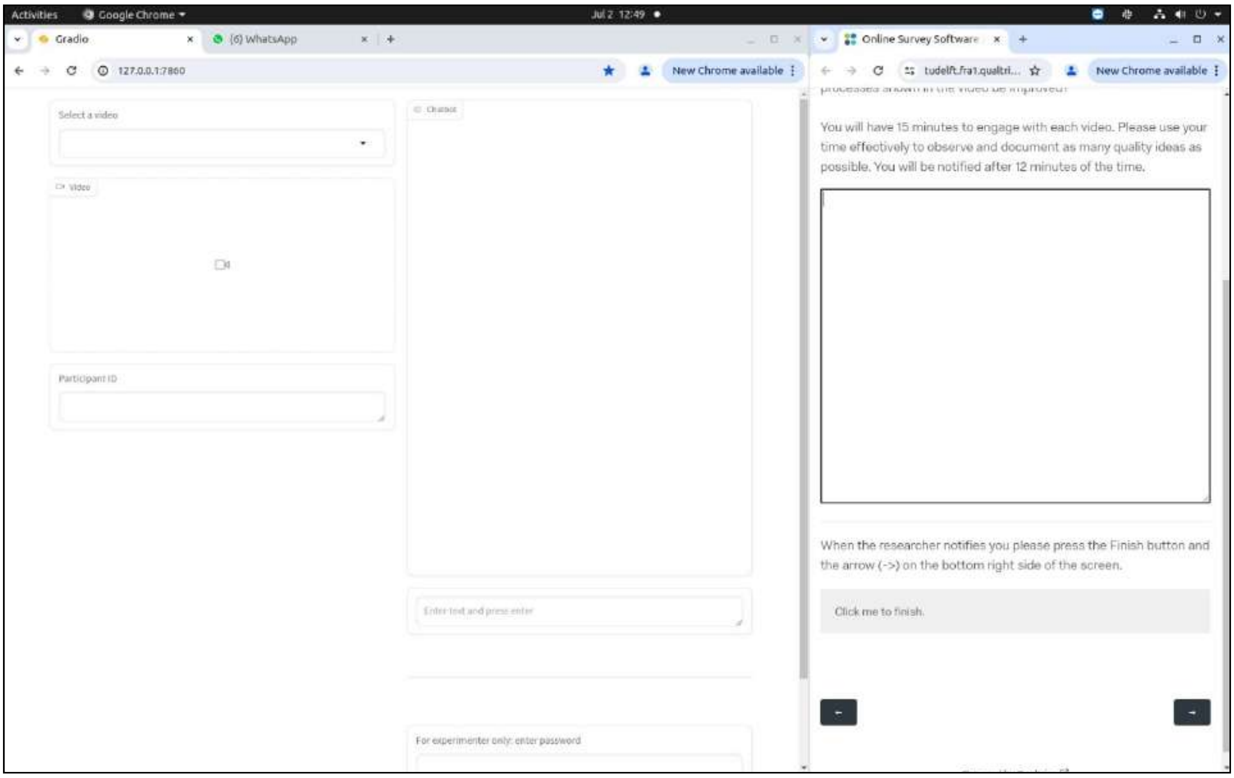


Figure 1: Prototype

3.3. User Interface

The interface had a video playback area on the left side of the screen. Directly below the video playback area, there was a field where facilitators could enter a previously determined participant ID to save chat history. Above the video playback area, a dropdown menu presents a list of preprocessed videos. This allowed facilitators to select from a range of four videos chosen for ideation session. This feature was designed to reduce the setup time for each session.

On the right side of the interface, a chatbot window facilitated direct interaction between designer participants and the AI model. Below the chat box, a "Save Chat History" field was available, secured by a password only known to facilitator. This feature allowed them to save, export, and review their conversations as JSON files for subsequent data analysis. Figure 2 shows the screenshot of the interface



processes shown in the video be improved?

You will have 15 minutes to engage with each video. Please use your time effectively to observe and document as many quality ideas as possible. You will be notified after 12 minutes of the time.

When the researcher notifies you please press the Finish button and the arrow (->) on the bottom right side of the screen.

Figure 2: Interface

3.4. Database

The study employed egocentric, first-person perspective videos from an Ego4D database. (Welcome to EGO4D! | Ego4D, n.d.). Ego4D database is a collaborative effort involving 15 universities and Meta AI. It is characterized by its collection of egocentric video recordings accompanied by annotations for machine learning models. It offers a dataset for understanding human interaction with objects and environments across a wide array of everyday activities. This dataset includes over 9000 videos, capturing 51 unique scenarios all recorded from a first-person perspective. It also encompasses diverse data types including audio, 3D poses, Inertial Measurement Unit (IMU) data, stereo vision, and multi-person activities captured from different viewpoints and gaze directions. Additionally, The dataset is annotated with summaries, narrations, and challenge annotations. These annotations detail the video content, specify actions, and highlight moments where the algorithm is expected to respond to queries related to past, present, and future events (Welcome to EGO4D! | Ego4D, n.d.).

This choice of this dataset was deliberate for several reasons – firstly to narrow the study scope and secondly to provide designers with a more nuanced exploration of user behavior regarding the context, interaction with objects and environments as opposed to also including videos filmed from exocentric perspective (from an external viewpoint).

The following are the reasons why solely egocentric videos were used:

Benefits considering the study scope:

1. Using a single type of video perspective eases the data collection and analysis process. This uniformity ensures that all data is comparable and analyzed under the same criteria, which can enhance the reliability of the findings and analysis.
2. Feasible within the timeframe and capabilities of the study by narrowing down the volume and variety of visual data that needs to be processed and analyzed

Benefits for designers:

1. The subjective nature of this perspective (egocentric) adds rigor and structure to design research (Höök et al., 2018) which are central for a deeper analysis of task execution, ergonomic factors, and user engagement.
2. This type of analysis has been found beneficial in sports environments and can be extrapolated to understand user interactions in designed spaces (Su et al. ,2017).
3. Multitask clustering of activities from wearable camera data provides valuable insights into user behaviors in different environments (Yan et al., 2015). This makes it appropriate for designing user-centered interfaces and products, as they reflect actual user interactions and experiences from the user's own viewpoint. This is especially important considering the everyday activities which the videos within the database are composed of. While exocentric videos can provide a broader view and context of user interaction within an environment, they often miss the nuanced, personal interactions captured by egocentric videos.
4. Egocentric videos minimize the bias introduced by an external observer's interpretation of the events. They provide more direct evidence of what the user is seeing and doing, reducing the layers of interpretation that might come with analyzing footage from an outside perspective.

3.5. Experiment development

This section of the thesis will explain the rationale behind the selection of videos for ideation session and outline the development of the procedure.

3.5.1. Video content selection

In selecting appropriate videos from the Ego4D database for this user testing research, the aim was to align closely with how designers utilize videos in the design process (Studying what people do, 2007.)

The alignment therefore narrowed the focus down to three analytical areas—usability studies, interaction analysis, and design ethnography—each contributing to understanding of user interactions and contextual influences. These areas were appropriate for the following reasons:

Usability Studies:
- showcases user interactions with products or services within their everyday environment

2. Interaction Analysis:
- showcases detailed sequences of activities
- user interaction with technology or tools is clear and evident
- captures the nuances of human-object interactions
- offers insights into the user experience and the ergonomic and cognitive demands of product interactions

3. Design Ethnography:
- portrays the cultural and social context of the activities featured
- captures scenarios that provide deep insights into user habits, rituals, and environmental factors that significantly influence their actions and decisions

To limit the testing session within one hour and to allow pre and post session activities, two specific categories of video content were prioritized: low and high complexity videos within the contexts of cooking and construction. These categories were selected because they are abundant in interactions and provide a diverse array of use cases that are pertinent for designers during the idea generation stage. Additionally, these categories are well-represented in the database, unlike other potentially suitable scenarios that were excluded due to the limited number of videos available, which restricts content complexity analysis.

3.5.2. Exclusion of Certain Design Video Uses

The use of videos for scenario-based design and participatory design was excluded from this research. These approaches involve creating envisioned future scenarios for presentations or engaging stakeholders in the design process, which are not directly tied to the search for inspiration as defined by the goals of this study.

3.5.3. Additional Video Selection criteria

The videos for this research needed to be rich in interactions and display a range of use cases while excluding repetitive activities. They also needed to provide insights into user habits, rituals, and the cultural or social context—key aspects that inform design decisions. The duration of the videos was kept short (3 minutes each), considering the total testing timeframe of maximum 15 minutes per video. This duration ensured that designers have sufficient time to engage with the video content, interact with the LLM, and generate ideas.

3.6. Participants

In the study, a total of 35 participants were recruited, adhering to the Human Research Ethics Committee (HREC) guidelines, which were obtained prior to the commencement of the research. The participants comprised a group of students from the Industrial Design Engineering faculty of TU Delft, including bachelor's, master's students, and PhD candidates. The recruitment process used convenience sampling.

The demographic average age of participants was 26.23 years (SD = 6.264). For gender, the distribution was 51.4% male and 48.6% female. The years of learning or experience in design among participants had a mean of 6.70 years (SD = 5.402).

3.7. Apparatus

The primary tools for data collection included Qualtrics for administering forms and capturing participant responses. Video recordings of the eye movements were captured using the incorporated recording feature of the Neon eyeglasses. Participants interacted with a desktop computer system within an office environment, which was equipped with a monitor of size 22 inch, a mouse, and a keyboard. To enable further enrichments of data gathered from eye tracking, four April tags were affixed to each corner of the monitor screen. The eye-tracking device used were Neon glasses by Pupil Labs. These glasses feature dual eye cameras for each eye to monitor eye movement. This includes tracking gaze direction, eye orientation, and blink detection, among other metrics. A wide-angle scene camera equipped on the glasses captures the frontal view from the wearer's perspective (Neon - Technical Specifications - Neon Eye Tracking Module and Frames, n.d.). The eye-tracking glasses were connected to an Android phone, which facilitated the storage of the data collected.

3.8. Measures

3.8.1. Objective Measures

3.8.1.1. Eye movement tracking

- Blink rate: The frequency at which blinks occur over time, indicating how often the eyes close and open.

- Blink duration: The length of time each blink lasts, showing how long the eyes remain closed during a blink.

- Saccade rate: The frequency of rapid, abrupt eye movements from one fixation point to another, highlighting how often the eyes jump between points of interest.

- Saccade duration: The time it takes for the eye to move from one point to another during a saccade, measuring the speed of these quick eye movements.

- Fixation rate: The frequency at which the eyes remain steadily focused on a single point, indicating how often the eyes stop to take in information.

- Fixation duration: The length of time the gaze remains on a single point before moving, showing how long information is processed from a specific location.

3.8.2. Subjective Measures

3.8.2.1. NASA TLX

The NASA Task Load Index (TLX) is a workload assessment tool designed to evaluate the workload experienced by operators working with various human-machine interface systems. NASA TLX utilizes a multi-dimensional rating procedure to derive an overall workload score from a weighted average of ratings on six subscales: Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration. This tool has been widely applied across different environments as well as in simulations and laboratory tests globally (The NASA TLX Tool: Task Load Index, n.d.-a)

3.8.2.2. Semi-structured Comparative Interview:

A methodology characterized by a set sequence of predetermined questions, which ensures consistency and comparability across interviews. For this study the primary focus was on collecting and analyzing non-numerical data (e.g., text) to understand concepts, thoughts, or experiences, while still allowing for their exploration (George, 2023).

In context of the study participants will reflect on:
- overall experience
- typical processes/methods for finding inspiration in design processes
- perceived benefits, drawbacks and significant considerations of using AI in design processes

3.8.2.3. Generated Ideas

Divergent thinking can be measured through the following three measures (Guilford, 1950)
Fluency – measured through the number of comprehensive ideas (portraying the purpose and functionality in sufficient detail to be understandable (Guilford, 1950). Therefore, within the same period, those who create a higher number of ideas have a higher probability of having higher creative output.

Flexibility – indicates the ability of participants to ideate in different categorical domains. During idea generation flexibility can be considered relevant as it reflects the importance of diverging into different domains to find a solution for the problem. Similarly, as fluency, higher flexibility can be related to increased likelihood of generating more creative ideas.

Originality - original idea is defined as an uncommon response to the design brief and was assessed by the statistical infrequency of each solution. Thus, originality is inversely correlated to the probability of being generated by the participants (Mednick, 1962).

3.8.2.4. User Experience Questionnaire (UEQ)

The User Experience Questionnaire (UEQ) is an assessment tool designed to measure the user experience of interactive products (User Experience Questionnaire (UEQ), n.d.; (PDF) User Experience Questionnaire Handbook Version 2, n.d.). Specifically, it assesses:

- Attractiveness: Overall impression of the product—whether users like or dislike it.
- Perspicuity: Ease of getting familiar with the product and learning how to use it.
- Efficiency: Users' ability to solve tasks without unnecessary effort and the product's responsiveness.
- Dependability: Whether users feel in control during interaction and perceive the product as secure and predictable.
- Stimulation: Excitement and motivation derived from using the product.
- Novelty: Creativity in design and its ability to catch users' interest.

3.8.2.5.Unified Theory of Acceptance and Use of Technology Questionnaire

The Unified Theory of Acceptance and Use of Technology (UTAUT) questionnaire assesses user acceptance of technology. It measures factors such as performance expectancy, effort expectancy, social influence, attitude towards using technology, social influence, self-efficacy, facilitating conditions, anxiety and behavioral intention to use the system ("User Acceptance of Information Technology: Toward a Unified View on JSTOR," n.d.).

As this research is focused on the early stages of understanding interaction with artificial intelligence and was conducted within an academic setting, some of the variables were left out as they were deemed not relevant. Specifically, these variables were included in the modified questionnaire: Performance expectancy, Effort Expectancy, Attitude toward using technology, Anxiety and Behavioral intention to use the system.

3.9. Procedure

3.9.2. Preparation and Main Experiment Session

Preparation: Before finalizing the procedure, two pilot tests were conducted prior to the main experiment session to refine the user testing procedure. These pilot tests aimed to determine the adequacy of the time allotted for participants to interact with the chatbot and simultaneously ideate, the number of videos participants would watch during the experiment session, and the optimal order of the questionnaires (demographics, NASA TLX, UTAUT, and UEQ). The pilot tests also sought to identify and resolve any unforeseen issues, such as confusing task descriptions, unclear questions, the necessity of breaks in the questionnaire, and ensuring all questions were formulated clearly and required responses.

To ensure a balanced exposure to the videos' complexity levels, a counterbalancing strategy was implemented. Four permutations of the video order from two categories were tracked using an Excel sheet. Participants were alternately assigned to the control and LLM groups in a sequenced manner—this approach was chosen to achieve an equal distribution of participants across groups while minimizing any bias that could arise from testing one group before the other.

After the preparatory steps were addressed, the experiment procedure was finalized. For the visual overview of the procedure see Figure 3.

Experiment Session: Upon their scheduled arrival at the laboratory office, facilitated by a Calendly reservation system (Free Online Appointment Scheduling Software | Calendly, n.d.) and confirmed via an automatic email notification, participants were greeted with a brief introduction and welcoming. Before proceeding with the experimental tasks, they were presented with consent forms, which they were allowed the time to read and sign digitally. Consent forms were modified based on the template provided by the The Human Research Ethics Committee (HREC) at TU Delft to protect participants from physical, emotional, and data privacy risks, ensuring that the research does not cause undue harm (Human Research Ethics, n.d.).

Following the initial orientation and consent procedures, participants sat in front of the desktop and did a 2-minute introduction session tailored to their assigned group. Those in the LLM group watched a short introductory video familiarizing them with the intended usage, the interface, expectations and capabilities of the interface. This video was to ensure participants felt comfortable with the LLM, the study's activities, apparatus and the concept of video-based design before proceeding. The control group, while not interacting with the LLM, had the opportunity to read the task description while asking questions and familiarizing themselves with the study's format and expectations. Participants were seated in front of the desktop and wearing all necessary equipment, and a design brief could be read on the screen during the ideation session. After participants indicated they understood it or asked additional questions, a time counter was initiated on Qualtrics. Similarly, recording of Neon eyeglasses was initiated.

The design brief provided to both groups was consistent but differed slightly to account for the use of the LLM in the experimental group. The LLM group was instructed as follows:

> *"You will be shown two videos. Your task is to analyze the videos and pinpoint processes or methods that could be enhanced. Focus on the activities and consider alternative tools, interactions, or contextual improvements. Generate and write out as many ideas as possible. You are encouraged to think out loud.*
>
> *Please use the provided chatbot to assist you. This tool offers insights and suggests improvements based on the video content. Type your questions or thoughts into the chatbot and use its responses to enhance your ideation. For example, ask, 'How can the process shown in the video be improved?'*
>
> *You will have 15 minutes to engage with each video. Please use your time effectively and document as many ideas as possible. Please note that videos do not have sound. You will be notified after 12 minutes of the time.*
>
> *When you are ready to proceed press the Start button and the arrow (->) on the bottom right side of the screen."*

Conversely, the control group received a similar brief without the mention of AI assistance:

> *"You will be shown two videos. Your task is to analyze the videos and pinpoint processes or methods that could be enhanced. Focus on the activities and consider alternative tools, interactions, or contextual improvements. Generate and write out as many ideas as possible. You are encouraged to think out loud. For example, ask yourself, 'How can the processes shown in the video be improved?'*
>
> *You will have 15 minutes to engage with each video. Please use your time effectively to observe and document as many ideas as possible. Please note that videos do not have sound. You will be notified after 12 minutes of the time.*
>
> *When you are ready to proceed press the Start button and the arrow (->) on the bottom right side of the screen."*

Participants were alerted at the twelve-minute mark of each session that they had three minutes remaining, ensuring they were aware of the time constraints and could prepare to conclude their current task. This cycle was repeated for both videos, after which the Qualtrics time counter and eye tracking recording was stopped, and participants were asked to evaluate aspects of their experience using NASA-TLX, UEQ and UTAUT questionnaires.

The study's last part involved a short, structured comparative interview where audio was recorded using meeting transcription of Microsoft Teams. Participants reflected on their overall experience and their typical processes for finding inspiration. Both groups engaged in discussions on their views about implementation of AI in design processes. After all the questions were asked the audio recording was stopped and participants were asked to fill out the payment tracking sheet after which they received gift cards.

3.9.3. Post Session

To express gratitude for their participation, each participant received a gift card. Personal details such as names and email addresses were collected on a sheet of paper in case of university financial audit purposes.

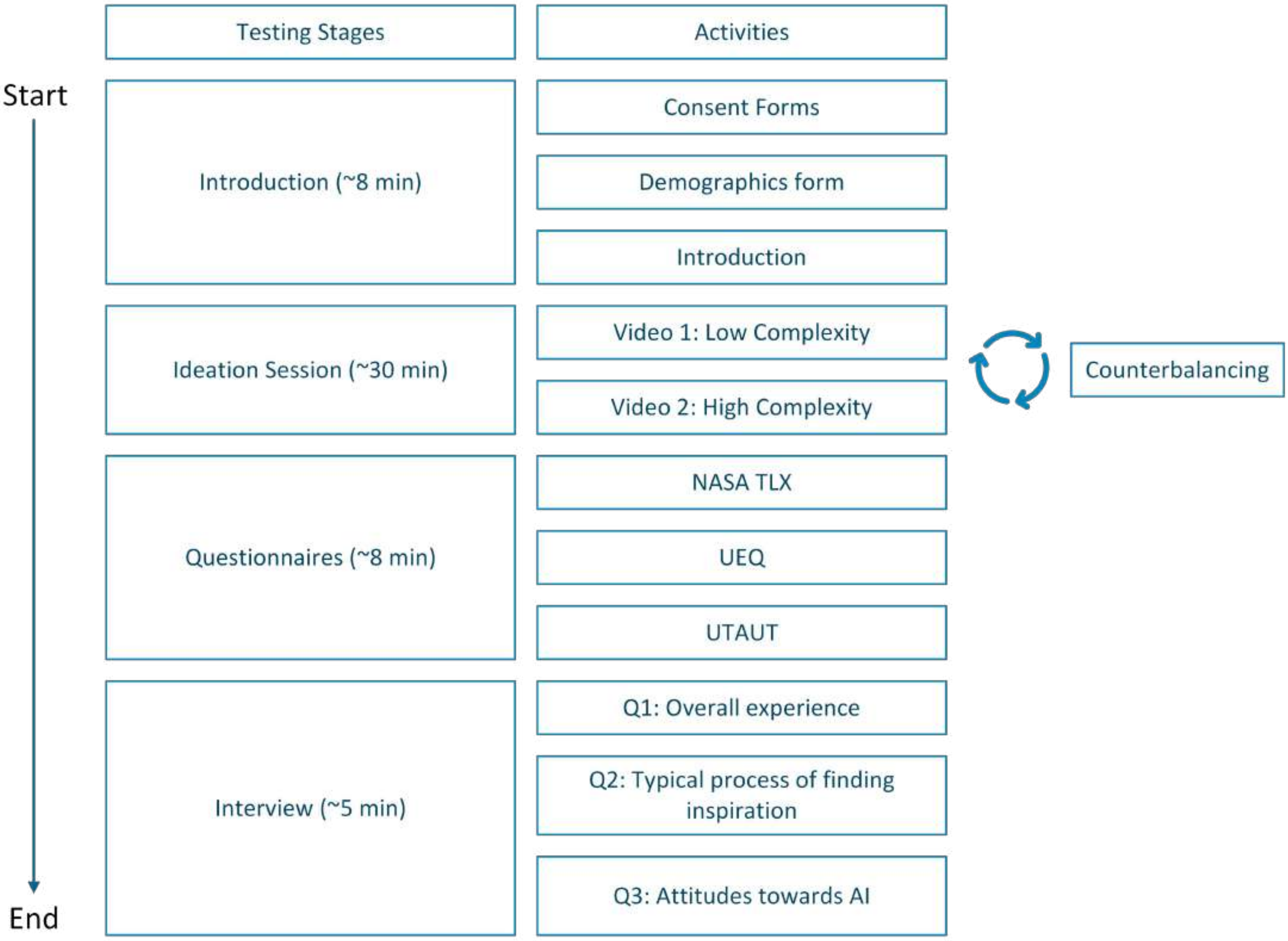


Figure 3: Experiment procedure

4. Data Analysis

4.1. Objective

As the study's main objective was to construct and develop the framework surrounding usage and validation of the tested system, data analysis was explored in a preliminary manner. Therefore, the objective of inferential analysis of data was to present the answers to the four research questions at large while focusing on qualitative data and to discern whether significant statistical differences exist between the experimental group, which interacted with the language model/chatbot, and the control group, which did not.

4.2. Overview

4.2.1. NASA Task Load Index

The NASA TLX metric was analyzed to evaluate the cognitive workload differences between the two study groups. The primary questions addressed are whether the LLM influences the overall workload score and whether such an influence is statistically significant.

4.2.2. User Experience Questionnaire (UEQ)

Using the UEQ, this study assessed the quality of user experience between the experimental and control groups. The exploration focused on whether the chatbot enhances or detracts from aspects of user experience such as attractiveness, perspicuity, efficiency, dependability, stimulation, and novelty.

4.2.2. Divergent Thinking

Divergent thinking was analyzed through its three main components: fluency, flexibility, and originality. This segment of the analysis seeks to identify whether any aspect of divergent thinking shows significant difference between the groups and whether the use of an LLM notably enhances or diminishes these creative traits. Data for the components will was analyzed manually meaning that it might be biased. This analysis also included a development of measurement framework of ideas. Specifically, based on the dataset, it defined ideas to be counted, assessed them on originality and created ideation domains into which ideas will be categorized.

4.2.3. Interviews

Thematic analysis of interview data explored common themes and patterns mentioned by participants, focusing on their experiences, view of artificial intelligence and general process of finding inspiration. The analysis compared between the experimental and control group.

4.2.2. UTAUT

Dataset was checked for normality between each group, checked for internal reliability within measured constructs using Cronbach's alpha, and corresponding test was performed based on normality.

4.3. Data Description

4.3.1. Data Source

The collection methods for this study encompassed both qualitative and quantitative primary data. Survey data were collected using Qualtrics, providing structured responses for analysis. Interviews were conducted and subsequently transcribed via Microsoft Teams, ensuring accurate capture of verbal data. Although data measured with Neon eye tracking glasses were gathered, it is not within the scope of this report and thus is not included in the analysis.

4.3.2. Variable Description

Divergent thinking:
Measured constructs:

- Fluency: Number of ideas.
- Flexibility: Number of ideation domains and subdomains the ideas cover. Counted separately. Subdomains were introduced to offer a greater level of detail and nuance
- Originality: Statistical infrequency of ideas. Measured on a 7-point Likert scale with 7 being highly original and 1 not being original.

NASA TLX: Workload assessment. Measuring only the overall workload score.
UTAUT: Performance Expectancy, Effort Expectancy, Attitude, Anxiety, Behavioral Intention (all measured on a 5-point Likert scale).
UEQ: User experience ratings on a 7-point bipolar scale.

4.4. Data Preparation

4.4.1. Renaming and Reorganizing Variables

The variables were renamed in SPSS and their order was reorganized from the original Qualtrics output. This was done to improve clarity and ease of interpretation during the analysis process.

4.4.2. Creation of New Variables

New variables were created to allow for the measurement of adjusted ratings of the NASA TLX score and the overall workload score. Additional variables were introduced to differentiate which participants watched videos according to theme (cooking and construction) and complexity levels (low and high). This differentiation was performed to ease the later analysis process.

4.4.3. Analysis Tehniques

The analysis was conducted using a combination of SPSS, Atlas.ti, Excel spreadsheets, and Python programming language.

- SPSS: SPSS was utilized for its statistical analysis capabilities. It was employed for calculating descriptive statistics, making graphs, and conducting hypothesis tests. The final workload scores, derived from NASA TLX comparison cards, were manually transferred to SPSS for further analysis.
- Atlas.ti: This software was employed to handle and code textual data, helping to identify patterns and themes within the qualitative responses.
- Python: Python programming language was used for weight counting of NASA TLX comparison cards. It enabled a more accurate processing of weights, which were then manually transferred to SPSS for subsequent analysis.
- UEQ Data Analysis: The User Experience Questionnaire (UEQ) was analyzed using the UEQ_Data_Analysis Excel spreadsheet (User Experience Questionnaire (UEQ), n.d.)

4.4.3.1. Divergent thinking

As mentioned beforehand, to explore the impact of video complexity and LLM assistance on divergent thinking, participants were exposed to two contextually distinct video scenarios, each presented in both low and high complexity variants. The analysis aimed to assess the creative outputs in terms of ideas across different complexity variables and compare the experimental (LLM) and control (no-LLM) groups for statistically significant differences in divergent thinking.

It is important to mention that although there might be statistically significant differences between the scenarios of cooking and construction, they were not considered in this study. Participants from both groups were exposed equally to the same sets of videos as counterbalancing strategy was performed.

Establishing the Definition of Ideas
To count as an idea, a suggestion within the text needed to offer a distinct solution to a problem or improvement. Ideas starting with identifying a problem or issue followed by a proposed solution or improvement were also considered. Mentions that did not offer clear indications of their functions and purpose were disregarded. Each idea was given a +1 score to count fluency.

Establishing Main Domains and Sub-Domains to Assess Flexibility
Ideas from both groups were clustered into domains and subdomains based on the entire dataset, irrespective of the scenario in the video or complexity, to ensure that each group is measured against the same criteria. This approach reduced variability from using different frameworks for different groups and allowed for direct comparison of data across groups where the difference is only in one condition.

While watching the videos, designers tended to engage in two main activities: identifying issues and offering design suggestions. The sequence of these activities varied and did not follow a strict chronological order during the ideation sessions while observing the two videos. In terms of identifying issues, designers identified issues either about the objects the individuals in the videos interacted with or the context they were situated in. Simultaneously, while identifiying issues they provided design solutions, for example, if a person was cutting vegetables in a cluttered kitchen, designers suggested improvements both for the knife being used and for organizing the kitchen to reduce clutter. These activities were categorized into two domains: context design and object design.

In addition to these design solution categories, there was a distinct category where designers offered solutions by integrating advanced, task-automating technologies. This was often speculative or conceptual in nature, proposing somewhat unfeasible technological solutions. For instance, if a designer observed difficulties in cutting vegetables, they might suggest improving the knife by altering its handle shape or cutting material, or they could propose designing an augmented reality (AR) cutting machine. Due to the prevalence of these speculative design ideas in the ideation outputs, a third domain focused on technology integration was created. This domain encompassed design suggestions involving futuristic or theoretical technologies aimed at automating tasks and significantly enhancing the objects or contexts observed in the videos.

After the domains were established, subdomains were created based on the clusters observed in the database. Generally, these subdomain clusters tended to focus on key aspects such as safety, management, usage, accessibility, and efficiency.

Five distinct patterns are mentioned below:

- Interaction issues - defining issues in interaction with the object.
- Context issues - defining issues of surrounding objects in the environment, such as kitchen elements, counters, storage, shelves, etc.
- Object design - Ideas focusing on the improvement and suggestions for the used tools.
- Speculative design - suggestions for the integration of new somewhat unfeasible technologies with used objects or surrounding objects.
- Context design - design suggestion for the context in which actions were taken.

Participants were given a +1 score for each domain and subdomain they covered. Scores for domains and subdomains were separated.

Establishing Criteria to Assess Originality

Each participant was given a single originality score on a 7-point Likert scale for each video watched. This approach reduced the complexity and time needed to evaluate each response individually, allowing for quicker analysis. This method assessed the overall originality of the entire output rather than individual ideas, as the overall creative level of the response is more relevant than the particulars of each idea in the context of this study. This provided a straightforward numerical value that can be compared across all participants.

Refer to Figure 4 for a list of domains and subdomains identified in the dataset

4.4.3.2. UEQ

Excel data analysis tool was used. (User Experience Questionnaire (UEQ), n.d.)

4.4.3.3. NASA TLX

The scores from the NASA TLX sources of workload comparison cards were counted using Python code, which can be seen in Appendix D. These scores were then transferred to SPSS, where they were multiplied by raw ratings. The new adjusted ratings were subsequently summed up a

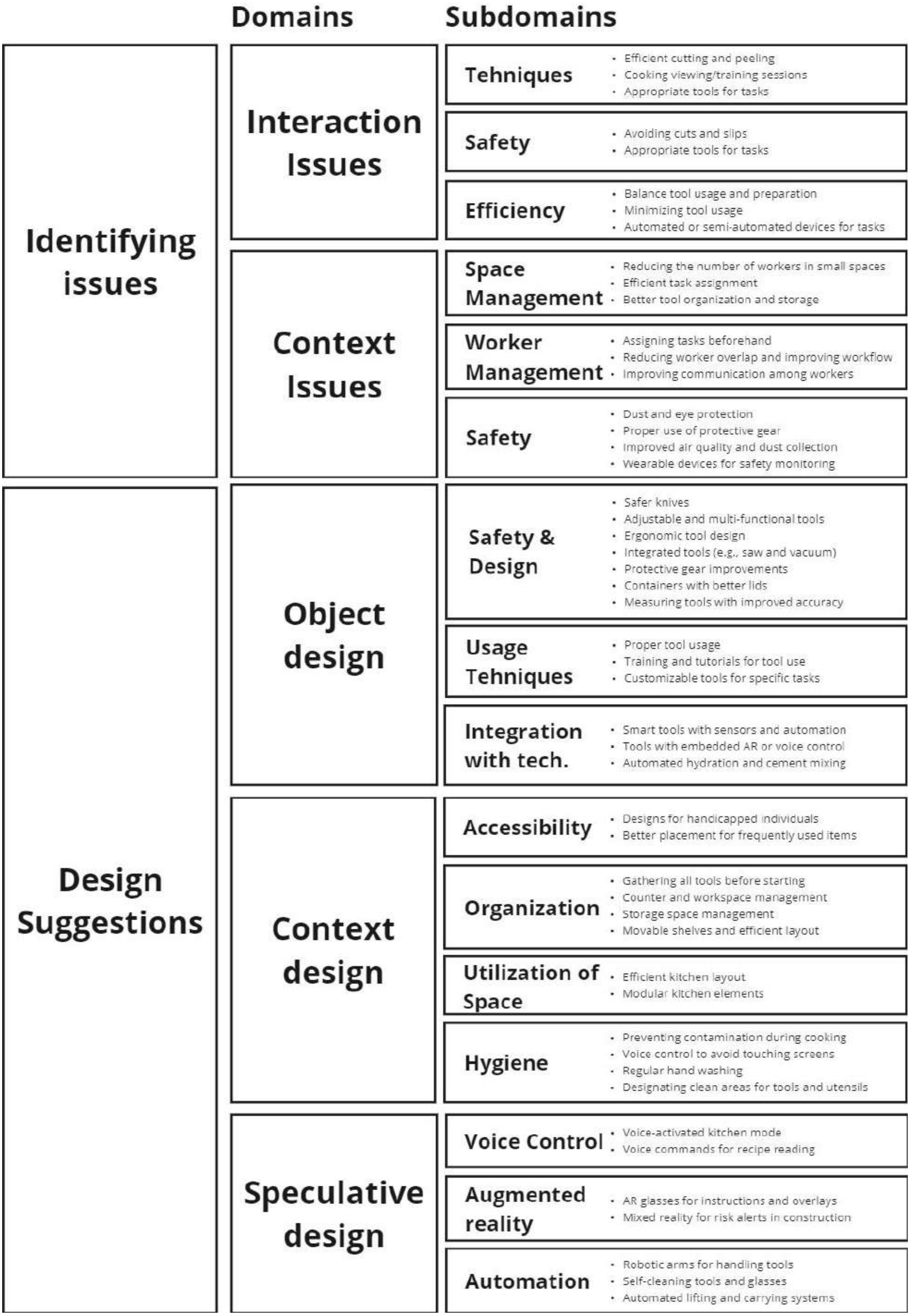
4.4.3.4. UTAUT

To ensure the reliability and internal consistency of the scales used in the questionnaire, Cronbach's alpha was employed. This metric measured how closely related the set of items are as a group, in the context of the project for constructs such as Performance Expectancy, Effort Expectancy, Attitude Towards Using Technology, Anxiety, and Behavioral Intention to Use the System. A Cronbach's alpha value greater than 0.7 was considered indicative of relatively high internal consistency, suggesting that the items consistently measured the same underlying construct. Each of the five chosen UTAUT constructs was used to create a new variable, representing the average value of the Likert scale questions associated with them. These means were then checked for internal reliability and normality, and the corresponding tests were performed accordingly. The dataset was examined for normality between each group and for internal reliability within the measured constructs using Cronbach's alpha, followed by appropriate statistical tests based on the normality results.

4.4.3.4. Thematic analysis of interviews

Qualitative data from interviews conducted under two conditions were analyzed using Atlas.ti software. As the interviews yielded a significant number of detailed patterns, they were subsequently grouped thematically into primary themes and subthemes for each question for both groups. Moreover, since only the first question was directly related to the experiment experience and was influenced by the group distinction, it was left separated within themes and subthemes while question two and three were grouped. Contrary to this, the detailed patterns were left separated for each questions for all the groups for a more thorough understanding, This approach was done to provide both a clearer, more focused, and organized overview of the data, making it easier to communicate the results, while at the same time enabling a more thorough overview. By understanding the broader context and significance of the data, the themes and subthemes helped to combine viewpoints and experiences. Additionally, these themes facilitated the formulation of future work recommendations.

Figure 6 represents the primary themes and subthemes.



	Domains	Subdomains	
Identifying issues	Interaction Issues	Tehniques	• Efficient cutting and peeling • Cooking viewing/training sessions • Appropriate tools for tasks
		Safety	• Avoiding cuts and slips • Appropriate tools for tasks
		Efficiency	• Balance tool usage and preparation • Minimizing tool usage • Automated or semi-automated devices for tasks
	Context Issues	Space Management	• Reducing the number of workers in small spaces • Efficient task assignment • Better tool organization and storage
		Worker Management	• Assigning tasks beforehand • Reducing worker overlap and improving workflow • Improving communication among workers
		Safety	• Dust and eye protection • Proper use of protective gear • Improved air quality and dust collection • Wearable devices for safety monitoring
Design Suggestions	Object design	Safety & Design	• Safer knives • Adjustable and multi-functional tools • Ergonomic tool design • Integrated tools (e.g., saw and vacuum) • Protective gear improvements • Containers with better lids • Measuring tools with improved accuracy
		Usage Tehniques	• Proper tool usage • Training and tutorials for tool use • Customizable tools for specific tasks
		Integration with tech.	• Smart tools with sensors and automation • Tools with embedded AR or voice control • Automated hydration and cement mixing
	Context design	Accessibility	• Designs for handicapped individuals • Better placement for frequently used items
		Organization	• Gathering all tools before starting • Counter and workspace management • Storage space management • Movable shelves and efficient layout
		Utilization of Space	• Efficient kitchen layout • Modular kitchen elements
		Hygiene	• Preventing contamination during cooking • Voice control to avoid touching screens • Regular hand washing • Designating clean areas for tools and utensils
	Speculative design	Voice Control	• Voice-activated kitchen mode • Voice commands for recipe reading
		Augmented reality	• AR glasses for instructions and overlays • Mixed reality for risk alerts in construction
		Automation	• Robotic arms for handling tools • Self-cleaning tools and glasses • Automated lifting and carrying systems

Figure 4: Flexibility - domains and subdomains

4.5. Results

4.5.1. Descriptive Statistics

The study involved a total of 35 participants. The demographic variables included age, gender, years of learning or experience in design, familiarity with Language Learning Models (LLMs), frequency of LLM usage, English proficiency, the highest degree or educational level attained, and familiarity with video-based design.

The average age of participants was 26.23 years (Figure 5 shows the age distribution graph), with a median of 24 years and a mode of 24 years. The standard deviation was 6.264, with a variance of 39.240 and a range of 37 years (minimum: 23, maximum: 60). The mean years of design experience was 6.70 years, with a median of 6 years and a mode of 6 years. The standard deviation for design experience was 5.402, with a variance of 29.179 and a range of 34 years (minimum: 1, maximum: 35).

Regarding gender, the distribution was 51.4% male and 48.6% female. In terms of educational background, 2.9% had a bachelor's degree, 31.4% had a master's degree, 62.9% were currently enrolled in a master's program, and 2.9% were currently enrolled in a doctorate or professional degree program.

For English proficiency, 2.9% were not proficient, 11.4% were somewhat proficient, 28.6% were moderately proficient, 37.1% were very proficient, and 20.0% were native speakers or extremely proficient. Familiarity with video-based design showed that 42.9% were not familiar at all, 20.0% were slightly familiar, 25.7% were moderately familiar, 5.7% were very familiar, and 5.7% were extremely familiar.

Familiarity with the brainwriting ideation method indicated that 5.7% were not familiar at all, 5.7% were slightly familiar, 25.7% were moderately familiar, 48.6% were very familiar, and 14.3% were extremely familiar. The frequency of LLM usage revealed that 5.7% rarely used LLMs, 20.0% occasionally used LLMs, 37.1% regularly used LLMs, and 37.1% frequently used LLMs.

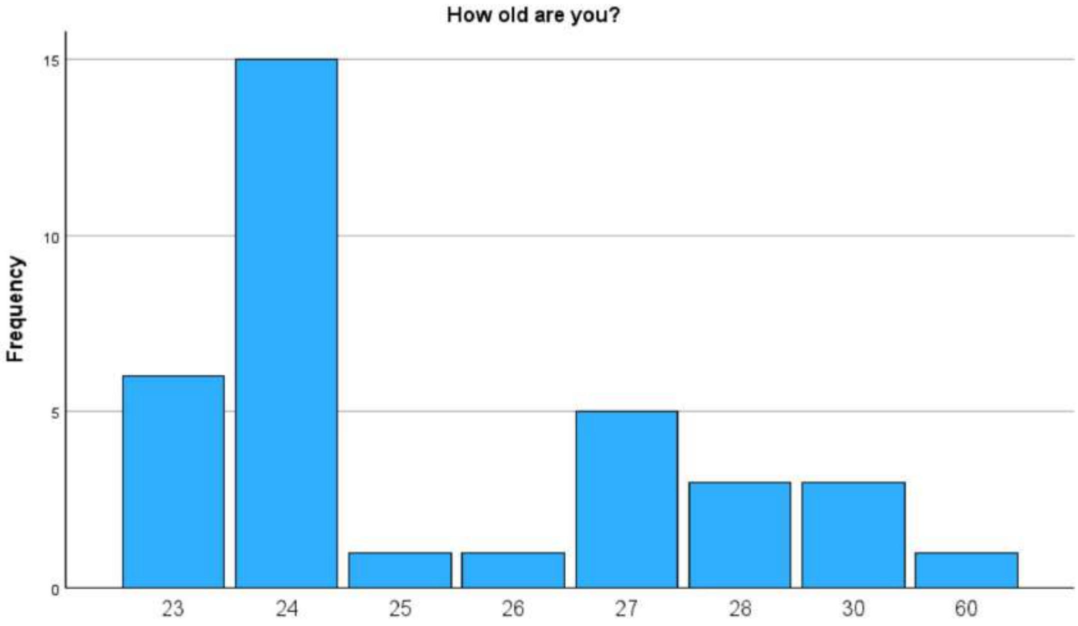
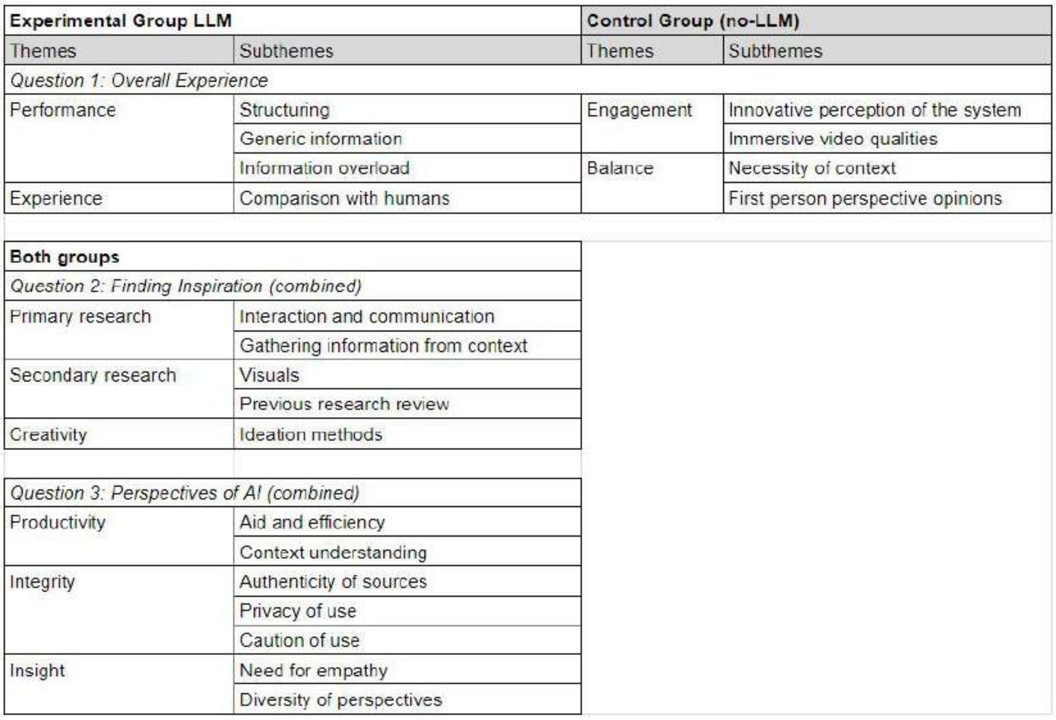


Figure 5: Age distribution

4.5.2. Qualitative Results

4.5.2.1. Interviews

High level themes among the interviewees:



Experimental Group LLM		**Control Group (no-LLM)**	
Themes	Subthemes	Themes	Subthemes
Question 1: Overall Experience			
Performance	Structuring	Engagement	Innovative perception of the system
	Generic information		Immersive video qualities
	Information overload	Balance	Necessity of context
Experience	Comparison with humans		First person perspective opinions

Both groups	
Question 2: Finding Inspiration (combined)	
Primary research	Interaction and communication
	Gathering information from context
Secondary research	Visuals
	Previous research review
Creativity	Ideation methods
Question 3: Perspectives of AI (combined)	
Productivity	Aid and efficiency
	Context understanding
Integrity	Authenticity of sources
	Privacy of use
	Caution of use
Insight	Need for empathy
	Diversity of perspectives

Figure 6: Interview themes and subthemes

The graph outlines the themes and subthemes derived from the experimental and control groups' responses to three key questions. For Question 1, focused on overall experience, the experimental group (LLM) concentrated on the chatbot's performance, with prominent subthemes including structuring of information, irrelevant information, and information overload. Participants also compared their interactions with the chatbot to interactions with human designers. The control group emphasized engagement, with subthemes such as perceiving the system innovatively and acknowledging immersive video qualities. They also focused on balance, highlighting the importance of understanding the context during ideation and the mixed impact of first-person perspectives, which provide direct insights but can limit overall context understanding.

In the combined analysis of both groups for Question 2, which addressed the process of finding inspiration, primary research emerged as a key theme. Participants highlighted the importance of interaction and communication with others, and gathering information from context and surroundings. Secondary research was also important, with subthemes including searching for online visuals and reviewing previous research. Creativity was encouraged through various ideation methods, such as brainstorming.

For Question 3, which explored perspectives on AI, participants frequently discussed how AI can enhance or hinder productivity. Subthemes included chatbots providing aid and efficiency, and the necessity of understanding context. Integrity was another key theme, with participants frequently mentioning the need for the authenticity of sources, privacy of use, and caution of use. Insight was also highlighted, with a focus on the need for empathy and diversity of perspectives.

4.5.2. Qualitative Results

4.5.2.1. Interviews

Detailed patterns among the interviewees:

4.5.2.1.1. Experimental Group (LLM)

Question 1: Overall Experience During the Session

- Participants generally found that LLMs could enhance the creative process by providing quick access to a wide range of ideas and information. They found the use of AI both intriguing in terms of its capabilities and challenging in terms of the chatbot fixating on elements interpreted in the video not important in the overall context of the video. For example, several participants mentioned that the chatbot tended to focus on the phone which appeared for a few seconds in the videos, while the overall activities performed in the video were related to cooking (cutting vegetables, plating food, washing dishes etc.)
- Some participants appreciated the potential of the chatbot for quickly generating a large volume of ideas. One participant highlighted the benefit of using the chatbot for "supporting ideation and doing summarization," indicating that it could be useful for initial brainstorming phases. However, the overall sentiment leaned towards chatbot's limitations in understanding context and providing relevant, accurate information. Participants generally found that LLMs could enhance the creative process by providing quick access to a wide range of ideas and information.
- Mentions of the AI chatbot providing overwhelming amount of information which made it difficult to process it during ideation
- Participants tended to compare interactions between AI as opposed to humans - human discussions deemed to provide more valuable feedback compared to the AI chatbot as participants appreciated the value of human interaction.

Question 2: Typical Process of Finding Inspiration When Generating Ideas

- Participants shared diverse methods for finding inspiration, emphasizing the importance of human interaction and personal experiences. Many participants mentioned that they typically draw inspiration from observing daily life, analyzing problems, and engaging in discussions with others. For example, one participant finds inspiration by "empathizing and putting oneself into the shoes of the person the design is for," suggesting an empathetic and user-centered approach to design.
- Many participants emphasized the importance of initial research, including reviewing existing works in their field, gathering information from online sources and observing surroundings and objects
- Mentions of importance of contexts and engagement for inspiration search; some mentioned finding inspiration spontaneously while most viewed inspiration as an integral part of ideation.
- The use of visual content, particularly videos, was appreciated for its ability to provide rich contextual information.
- Participants also mentioned using AI tools as a co-pilot to generate additional ideas; generally seen as secondary, with the AI providing a list of structuring help which the designers could then refine and expand upon.

Question 3: Integrating AI Chatbots into Creative Processes

- Mentions of AI offering a broader perspective by summarizing vast amounts of information and highlighting trends or patterns that might not be immediately obvious to human designers; providing a more comprehensive view of user experiences.
- Several participants pointed out that AI could inadvertently lead to design fixation or biased outcomes if not used carefully. The risk of AI providing misleading information was a concern to some, with one participant describing the chatbot as sometimes giving "misleading information" or hyper-fixating on irrelevant details in the video. Because of this they emphasized the process of validation and reflection. They also mentioned the risk of over-reliance on AI, which might lead to shortcuts in the design process
- Emphasizing the human element of creativity and empathy. Theme of ensuring that AI complements rather than replaces human creativity
- Some participants viewed AI as a valuable tool for handling repetitive and mundane tasks, thus freeing up designers to focus on more creative aspects. One participant described AI as being useful for "cutting down on mundane work," allowing designers to concentrate on higher-level ideation and decision-making.

4.5.2.1.2. Control Group (no-LLM)

Question 1: Overall Experience During the Session

- Many participants found the session innovative and helpful for taking notes and gathering insights. Mentions of the ease of taking notes while watching videos, although some preferred to ideate on paper.
- Several participants noted the difficulty of keeping up with the sometimes-overwhelming content of the videos. Some found shakiness and constant movement of the first-person perspective videos were particularly problematic, causing dizziness and difficulty in focusing on the task.
- The first-person perspective was appreciated for its immersive experience of user behaviour
- The process of watching videos and taking notes was seen as straightforward and helpful in providing context.

Question 2: Typical Process of Finding Inspiration When Generating Ideas

- Many participants mentioned starting by conducting extensive research, consulting with experts, and looking at end-user experiences. This includes watching relevant videos, reading articles, and checking user feedback to gather a broad understanding of the context and needs.
- Online resources like Pinterest/Behance and other design platforms are used to gather visual inspiration and see what is trending in the industry.
- Some participants prefer to observe real-life scenarios relevant to their projects, to help them understand user behavior and needs directly from the environment where the product or design will be used.
- Mentions of engagement in brainstorming sessions, often iterating on their ideas with feedback from peers or further research.

Question 3: Integrating AI Chatbots into Creative Processes

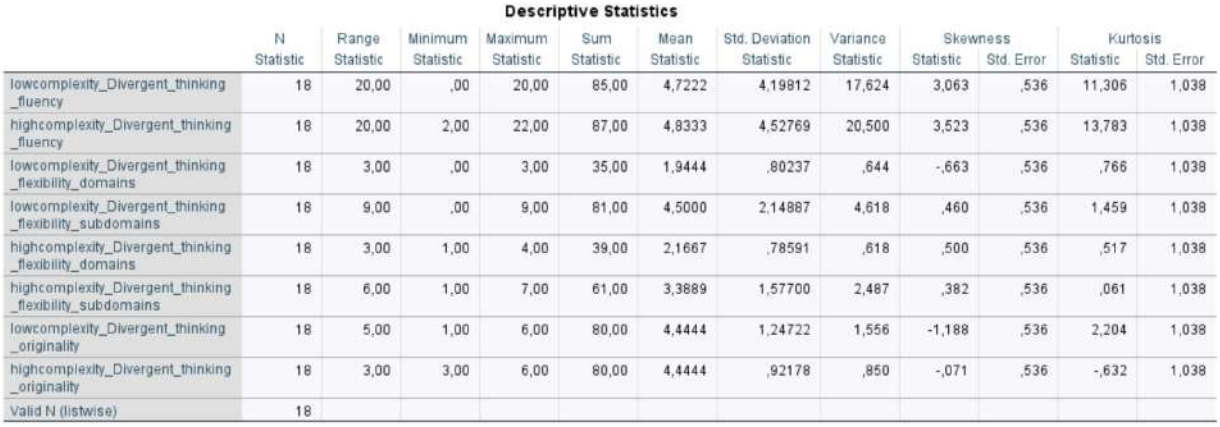
- Perceptions of AI chatbots as being able to significantly enhance efficiency by automating repetitive tasks, providing quick access to information, and offering suggestions
- Several participants mentioned that AI could assist in the design process by providing comparisons, quantitative analysis, and aiding in the search for relevant content. However, they emphasized the importance of designers making the final decisions to ensure creativity and authenticity in the design outputs. Concerns about AI providing generic answers and the potential loss of human uniqueness in the design process were noted.
- Mentions of importance of contextual understanding and consideration is the AI's ability to understand the specific context and needs of the designer.
- Mentions of ensuring that AI respects privacy and is used responsibly.

4.5.3. Quantitative Results

4.5.3.1. Divergent Thinking

4.5.3.1.1. Comparison of descriptives between groups

4.5.3.1.1.1. Experimental Group (LLM)

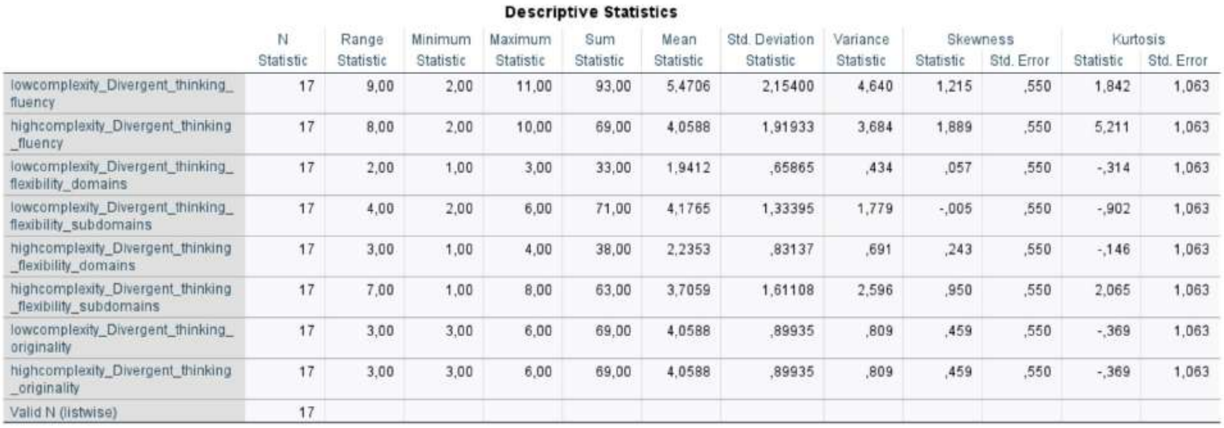


Descriptive Statistics

	N	Range	Minimum	Maximum	Sum	Mean	Std. Deviation	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
lowcomplexity_Divergent_thinking_fluency	18	20,00	,00	20,00	85,00	4,7222	4,19812	17,624	3,063	,536	11,306	1,038
highcomplexity_Divergent_thinking_fluency	18	20,00	2,00	22,00	87,00	4,8333	4,52769	20,500	3,523	,536	13,783	1,038
lowcomplexity_Divergent_thinking_flexibility_domains	18	3,00	,00	3,00	35,00	1,9444	,80237	,644	-,663	,536	,766	1,038
lowcomplexity_Divergent_thinking_flexibility_subdomains	18	9,00	,00	9,00	81,00	4,5000	2,14887	4,618	,460	,536	1,459	1,038
highcomplexity_Divergent_thinking_flexibility_domains	18	3,00	1,00	4,00	39,00	2,1667	,78591	,618	,500	,536	,517	1,038
highcomplexity_Divergent_thinking_flexibility_subdomains	18	6,00	1,00	7,00	61,00	3,3889	1,57700	2,487	,382	,536	,061	1,038
lowcomplexity_Divergent_thinking_originality	18	5,00	1,00	6,00	80,00	4,4444	1,24722	1,556	-1,188	,536	2,204	1,038
highcomplexity_Divergent_thinking_originality	18	3,00	3,00	6,00	80,00	4,4444	,92178	,850	-,071	,536	-,632	1,038
Valid N (listwise)	18											

Figure 7: Divergent thinking descriptives - experimental group

4.5.3.1.1.2. Control Group (no-LLM)



Descriptive Statistics

	N	Range	Minimum	Maximum	Sum	Mean	Std. Deviation	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
lowcomplexity_Divergent_thinking_fluency	17	9,00	2,00	11,00	93,00	5,4706	2,15400	4,640	1,215	,550	1,842	1,063
highcomplexity_Divergent_thinking_fluency	17	8,00	2,00	10,00	69,00	4,0588	1,91933	3,684	1,889	,550	5,211	1,063
lowcomplexity_Divergent_thinking_flexibility_domains	17	2,00	1,00	3,00	33,00	1,9412	,65865	,434	,057	,550	-,314	1,063
lowcomplexity_Divergent_thinking_flexibility_subdomains	17	4,00	2,00	6,00	71,00	4,1765	1,33395	1,779	-,005	,550	-,902	1,063
highcomplexity_Divergent_thinking_flexibility_domains	17	3,00	1,00	4,00	38,00	2,2353	,83137	,691	,243	,550	-,146	1,063
highcomplexity_Divergent_thinking_flexibility_subdomains	17	7,00	1,00	8,00	63,00	3,7059	1,61108	2,596	,950	,550	2,065	1,063
lowcomplexity_Divergent_thinking_originality	17	3,00	3,00	6,00	69,00	4,0588	,89935	,809	,459	,550	-,369	1,063
highcomplexity_Divergent_thinking_originality	17	3,00	3,00	6,00	69,00	4,0588	,89935	,809	,459	,550	-,369	1,063
Valid N (listwise)	17											

Figure 8: Divergent thinking descriptives - control group

Fluency:

For low complexity fluency, the control group has a higher mean (5.47 vs. 4.72) and lower skewness (1.22 vs. 3.06), indicating a more balanced distribution with less extreme values. For high complexity fluency, the experimental group has a slightly higher mean (4.83 vs. 4.06) but also higher skewness (3.52 vs. 1.89), suggesting a more positively skewed distribution.

Flexibility (Domains and Subdomains):

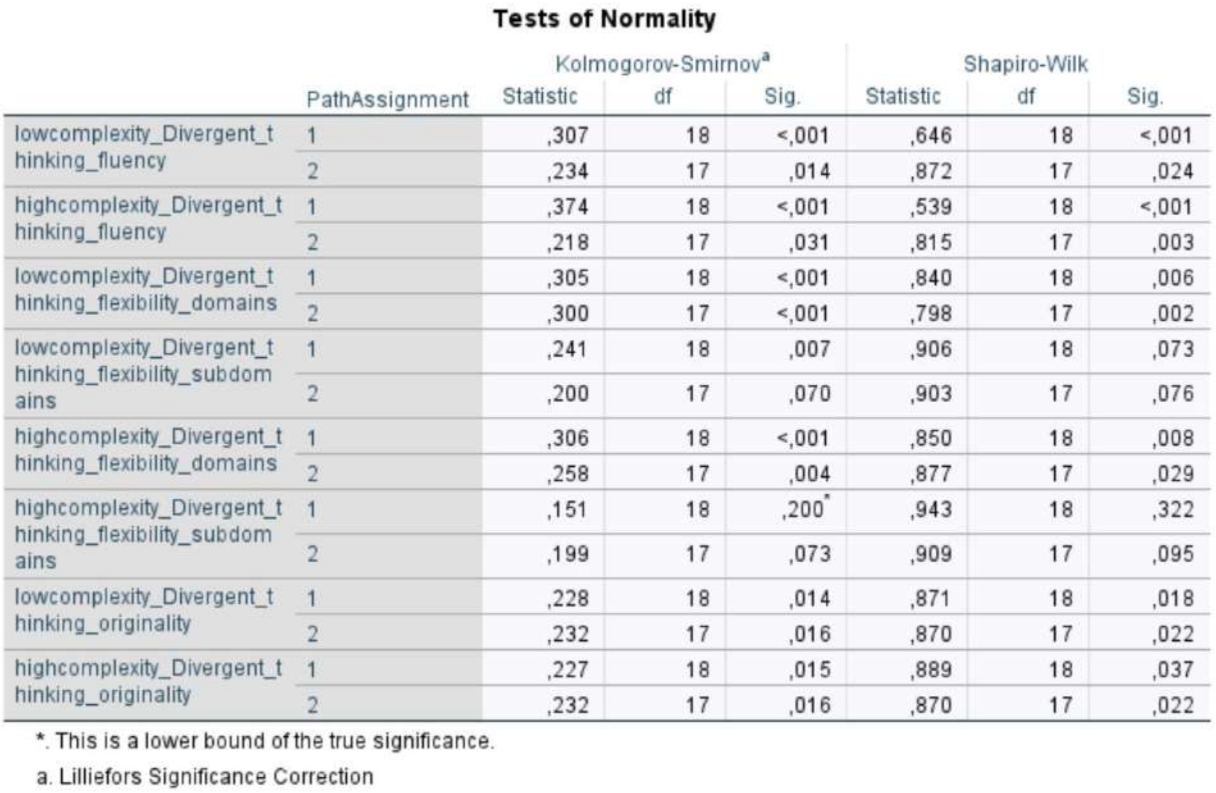
For low complexity flexibility (domains), both groups have similar means (1.94) and similar skewness, indicating very similar distributions. For low complexity flexibility (subdomains), the experimental group shows a higher mean (4.50 vs. 4.18) and higher variability (SD = 2.15 vs. 1.33), with a slight positive skew. For high complexity flexibility (domains), the control group has a slightly higher mean (2.24 vs. 2.17) with a slightly lower skewness. For high complexity flexibility (subdomains), the control group has a higher mean (3.71 vs. 3.39) and higher skewness (0.95 vs. 0.38), indicating more spread in the scores.

Originality:

For both low and high complexity originality, the experimental group has slightly higher means (4.44 vs. 4.06), but lower skewness for low complexity originality, suggesting less extreme values. The control group has higher kurtosis for low complexity originality, indicating a more peaked distribution.

The descriptive statistics highlight that the experimental group generally exhibits higher means for most measures of divergent thinking, especially in high complexity fluency and originality measures. However, the control group shows less skewness in fluency measures, indicating more balanced distributions.

4.5.3.1.2. Normality Test



Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	PathAssignment	Statistic	df	Sig.	Statistic	df	Sig.
lowcomplexity_Divergent_thinking_fluency	1	,307	18	<,001	,646	18	<,001
	2	,234	17	,014	,872	17	,024
highcomplexity_Divergent_thinking_fluency	1	,374	18	<,001	,539	18	<,001
	2	,218	17	,031	,815	17	,003
lowcomplexity_Divergent_thinking_flexibility_domains	1	,305	18	<,001	,840	18	,006
	2	,300	17	<,001	,798	17	,002
lowcomplexity_Divergent_thinking_flexibility_subdomains	1	,241	18	,007	,906	18	,073
	2	,200	17	,070	,903	17	,076
highcomplexity_Divergent_thinking_flexibility_domains	1	,306	18	<,001	,850	18	,008
	2	,258	17	,004	,877	17	,029
highcomplexity_Divergent_thinking_flexibility_subdomains	1	,151	18	,200*	,943	18	,322
	2	,199	17	,073	,909	17	,095
lowcomplexity_Divergent_thinking_originality	1	,228	18	,014	,871	18	,018
	2	,232	17	,016	,870	17	,022
highcomplexity_Divergent_thinking_originality	1	,227	18	,015	,889	18	,037
	2	,232	17	,016	,870	17	,022

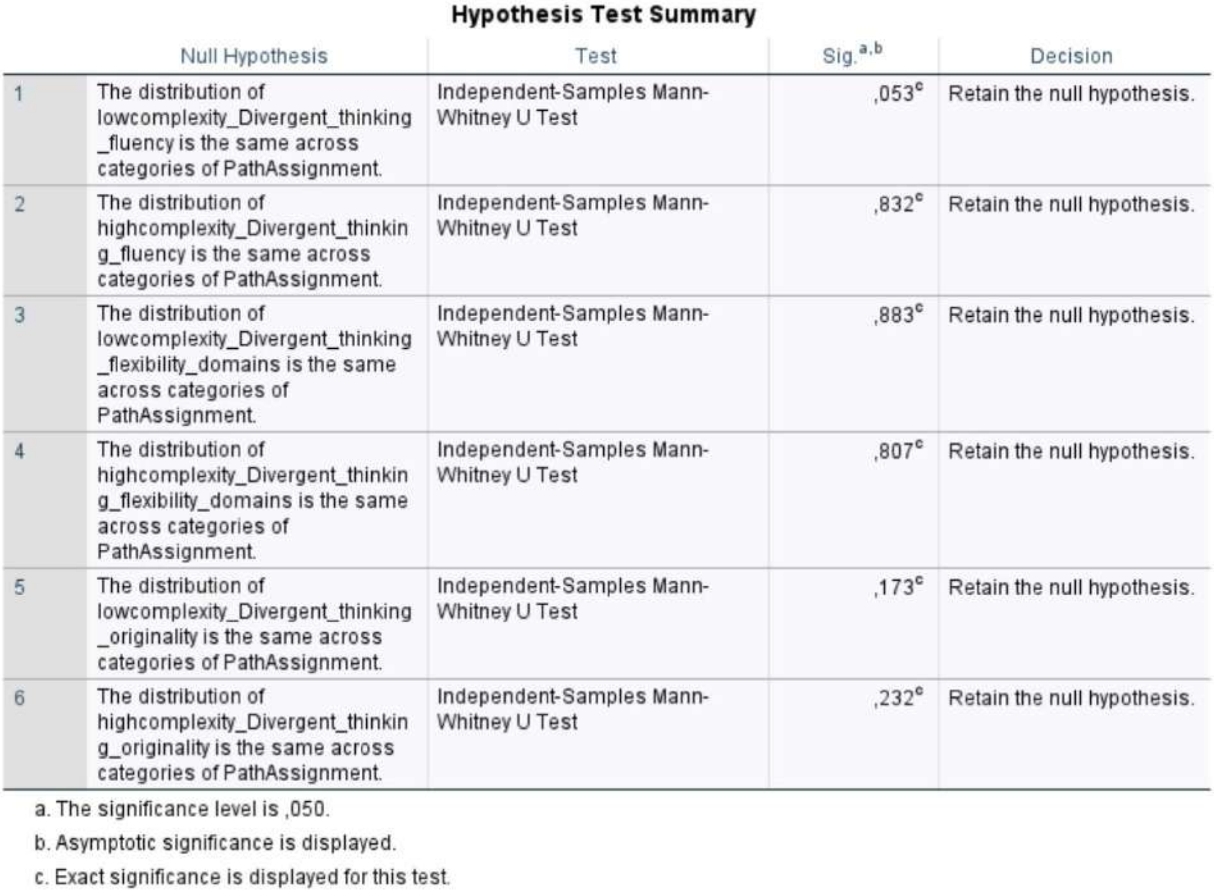
*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Figure 9: Divergent thinking - normality test

The Shapiro-Wilk test (Figure 9) results indicate that the data is normally distributed under the following conditions: low complexity subdomains in the experimental group (p-value = 0.073), low complexity subdomains in the control group (p-value = 0.076), high complexity subdomains in the control group (p-value = 0.095), and high complexity subdomains in the experimental group (p-value = 0.322). Each of these p-values exceeds the 0.05 threshold, confirming normality for these specific conditions. In other cases, data is not normally distributed.

4.5.3.1.3. Divergent thinking: Mann Whitney U Test



Hypothesis Test Summary

	Null Hypothesis	Test	Sig.[a,b]	Decision
1	The distribution of lowcomplexity_Divergent_thinking_fluency is the same across categories of PathAssignment.	Independent-Samples Mann-Whitney U Test	,053[c]	Retain the null hypothesis.
2	The distribution of highcomplexity_Divergent_thinking_fluency is the same across categories of PathAssignment.	Independent-Samples Mann-Whitney U Test	,832[c]	Retain the null hypothesis.
3	The distribution of lowcomplexity_Divergent_thinking_flexibility_domains is the same across categories of PathAssignment.	Independent-Samples Mann-Whitney U Test	,883[c]	Retain the null hypothesis.
4	The distribution of highcomplexity_Divergent_thinking_flexibility_domains is the same across categories of PathAssignment.	Independent-Samples Mann-Whitney U Test	,807[c]	Retain the null hypothesis.
5	The distribution of lowcomplexity_Divergent_thinking_originality is the same across categories of PathAssignment.	Independent-Samples Mann-Whitney U Test	,173[c]	Retain the null hypothesis.
6	The distribution of highcomplexity_Divergent_thinking_originality is the same across categories of PathAssignment.	Independent-Samples Mann-Whitney U Test	,232[c]	Retain the null hypothesis.

a. The significance level is ,050.

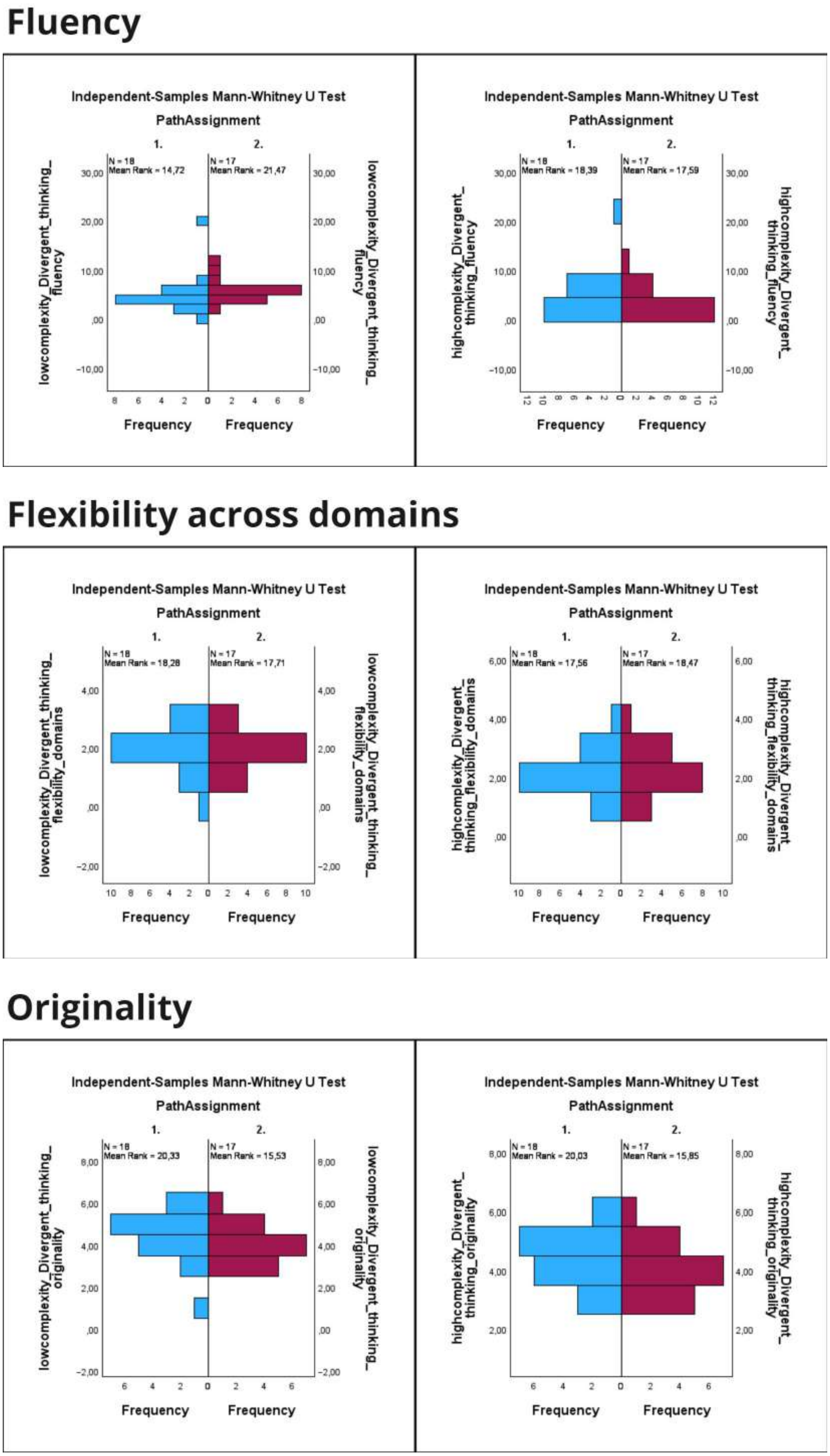
b. Asymptotic significance is displayed.

c. Exact significance is displayed for this test.

Figure 10: Fluency, flexibility across domains and originality Mann Whitney U test results

The results of the hypothesis tests, summarized in the table (Figure 10), show that the distributions of various measures of divergent thinking are the same across the experimental and control groups. Specifically, the Independent-Samples Mann-Whitney U Test was used to compare the groups, and in all cases, the p-values were greater than 0.05, leading to the retention of the null hypothesis in each case. This means that there are no significant differences between the experimental and control groups in terms of low complexity fluency (p=0.053), high complexity fluency (p=0.832), low complexity flexibility at the domain level (p=0.883), high complexity flexibility at the domain level p=0.807), low complexity originality (=0.173), and high complexity originality (p=0.232). These findings indicate that the measures of divergent thinking do not differ significantly between the two groups.

Figure 11: Fluency, flexibility across domains and originality Mann Whitney U test results



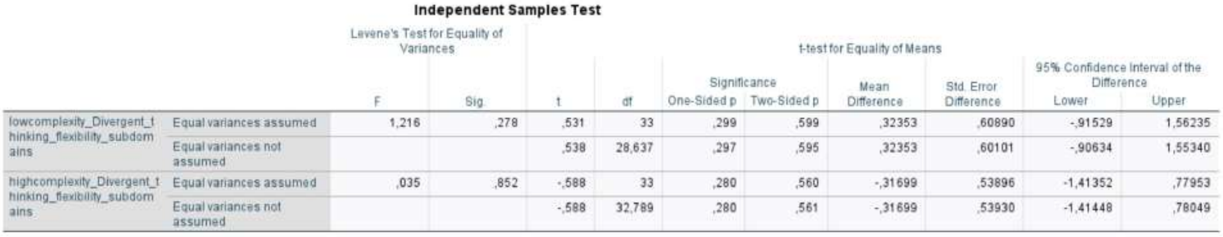
Fluency

Flexibility across domains

Originality

The frequency distributions (Figure 11) provide a visual comparison of divergent thinking measures between the experimental and control groups. For low complexity fluency, the control group shows a higher concentration of higher scores with a mean rank of 21.47 compared to 14.72 for the experimental group. In high complexity fluency, the distributions are similar, with mean ranks of 18.39 for the experimental group and 17.59 for the control group. For low complexity flexibility at the domain level, the distributions are balanced with mean ranks of 18.28 for the experimental group and 17.71 for the control group, while high complexity flexibility at the domain level also shows similar distributions with mean ranks of 17.56 and 18.47, respectively. In low complexity originality, the experimental group shows higher scores with a mean rank of 20.33 compared to 15.53 for the control group, and similarly, in high complexity originality, the experimental group has higher scores with a mean rank of 20.03 compared to 15.85 for the control group. These visual comparisons suggest some differences between the groups, with the control group showing higher scores in fluency measures and the experimental group showing higher scores in flexibility and originality measures, although the hypothesis tests found no significant differences.

4.5.3.1.4. Divergent thinking: Independent Samples t-Test



Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means							
						Significance				95% Confidence Interval of the Difference	
		F	Sig.	t	df	One-Sided p	Two-Sided p	Mean Difference	Std. Error Difference	Lower	Upper
lowcomplexity_Divergent_thinking_flexibility_subdomains	Equal variances assumed	1,216	,278	,531	33	,299	,599	,32353	,60890	-,91529	1,56235
	Equal variances not assumed			,538	28,637	,297	,595	,32353	,60101	-,90634	1,55340
highcomplexity_Divergent_thinking_flexibility_subdomains	Equal variances assumed	,035	,852	-,588	33	,280	,560	-,31699	,53896	-1,41352	,77953
	Equal variances not assumed			-,588	32,789	,280	,561	-,31699	,53930	-1,41448	,78049

Figure 12: Divergent thinking - independent samples t-test results

The independent samples test results (Figure 12) for low and high complexity divergent thinking flexibility subdomains indicate no significant differences between the experimental and control groups. For low complexity, the equal variances assumed test yielded t(33) = 0.531, p = 0.599, and the equal variances not assumed test yielded t(28.637) = 0.538, p = 0.595, with both tests showing non-significant mean differences of 0.32353 (95% CI: -0.91529 to 1.56235 and -0.90634 to 1.55340, respectively). For high complexity, the equal variances assumed test yielded t(33) = -0.588, p = 0.560, and the equal variances not assumed test yielded t(32.789) = -0.588, p = 0.561, with non-significant mean differences of -0.31699 (95% CI: -1.41352 to 0.77953 and -1.41448 to 0.78049, respectively). These results suggest that there are no statistically significant differences in the means for either complexity level, indicating that the condition did not substantially impact thinking flexibility subdomains.

4.5.3.2. UEQ

Each construct's internal consistency was evaluated using Cronbach's Alpha and Guttman's Lambda 2 coefficients to ensure the reliability of the questionnaire items.

Based on the provided data analysis tools by UEQ, values of the six UEQ scales between -0.8 and 0.8 represent a neutral evaluation of the corresponding scale. Values greater than 0.8 represent a positive evaluation, and values less than -0.8 represent a negative evaluation.

The scales of the UEQ are grouped into pragmatic quality (Perspicuity, Efficiency, Dependability) and hedonic quality (Stimulation, Originality). Pragmatic quality describes task-related quality aspects, while hedonic quality describes non-task-related quality aspects.

4.5.3.2.1. Determining Precision and Reliability

To understand how many participants are required to achieve a certain precision in the measurement of the scale means, we considered their error probabilities based on standard deviations of the six scales.

LLM Group Analysis
For the LLM group with 18 participants, the standard deviations of the scales from the User Experience Questionnaire suggest moderate variability in responses. Specifically, with 18 participants, we can achieve a precision of 0.5 with an error probability of 0.1 for most scales. This means the true mean is expected to be within ±0.5 of the sample mean with 90% confidence. However, for scales with higher variability, such as Novelty standing at 19, achieving this precision is with slightly less confidence.

Control Group Analysis
For the Control group with 17 participants, the standard deviations indicate that we can achieve a precision of 0.5 with an error probability of 0.1 for most scales. However, the efficiency and novelty scales are slightly below the threshold. This suggests that for these scales, the true mean might be less confidently within ±0.5 of the sample mean with 90% confidence. For the other scales, we can be confident in the precision of the results.

Experimental Group Reliability

- Attractiveness: Cronbach's Alpha: 0.90, Lambda2: 0.90
- Perspicuity: Cronbach's Alpha: 0.69, Lambda2: 0.73
- Efficiency: Cronbach's Alpha: 0.73, Lambda2: 0.76
- Dependability: Cronbach's Alpha: 0.30, Lambda2: 0.49
- Stimulation: Cronbach's Alpha: 0.69, Lambda2: 0.68
- Novelty: Cronbach's Alpha: 0.78, Lambda2: 0.79

Control Group Reliability

- Attractiveness: Cronbach's Alpha: 0.90, Lambda2: 0.90
- Perspicuity: Cronbach's Alpha: 0.84, Lambda2: 0.85
- Efficiency: Cronbach's Alpha: 0.82, Lambda2: 0.82
- Dependability: Cronbach's Alpha: 0.61, Lambda2: 0.63
- Stimulation: Cronbach's Alpha: 0.84, Lambda2: 0.83
- Novelty: Cronbach's Alpha: 0.85, Lambda2: 0.83

4.5.3.2.2. Experimental Group (LLM) Results

Overall UEQ Scales (Figure 13)

- Attractiveness: Mean = 0.676, Variance = 1.12
- Perspicuity: Mean = 1.375, Variance = 1.16
- Efficiency: Mean = 0.972, Variance = 1.22
- Dependability: Mean = 0.542, Variance = 0.89
- Stimulation: Mean = 0.778, Variance = 0.74
- Novelty: Mean = 0.625, Variance = 1.77

Grouped UEQ scales (Figure 14)

- Pragmatic Quality: Mean = 1.01
- Hedonic Quality: Mean = 0.75
- Attractiveness: Mean = 0.73

Interpretation of Results

The attractiveness score, with a mean value of 0.676, indicates a generally positive user experience, suggesting that users found the system pleasant to use, though there is room for improvement to make it more enjoyable. The perspicuity score, with a high mean value of 1.375, indicates that users found the system very easy to understand and use, likely due to its clear and intuitive design. A mean efficiency score of 0.972 signifies that users found the system efficient in helping them achieve their tasks, with the integration of video viewing, chatbot interaction, and idea documentation likely easy to use while enhancing productivity. The dependability score, with a mean value of 0.542, suggests mixed feelings about the system's reliability, indicating that users may have encountered occasional issues with the chatbot's performance or system stability, and highlighting the need for improvements in these areas to enhance trust. The stimulation score, with a mean value of 0.778, suggests that users found the system somewhat engaging and motivating, with interactive elements and innovative features contributing to an enjoyable experience, though further enhancements could increase engagement. Finally, the novelty score, with a mean value of 0.625, indicates that users found the system moderately innovative, with the combination of video analysis and AI interaction perceived as fresh, but with potential for introducing more unique features to elevate the sense of novelty.

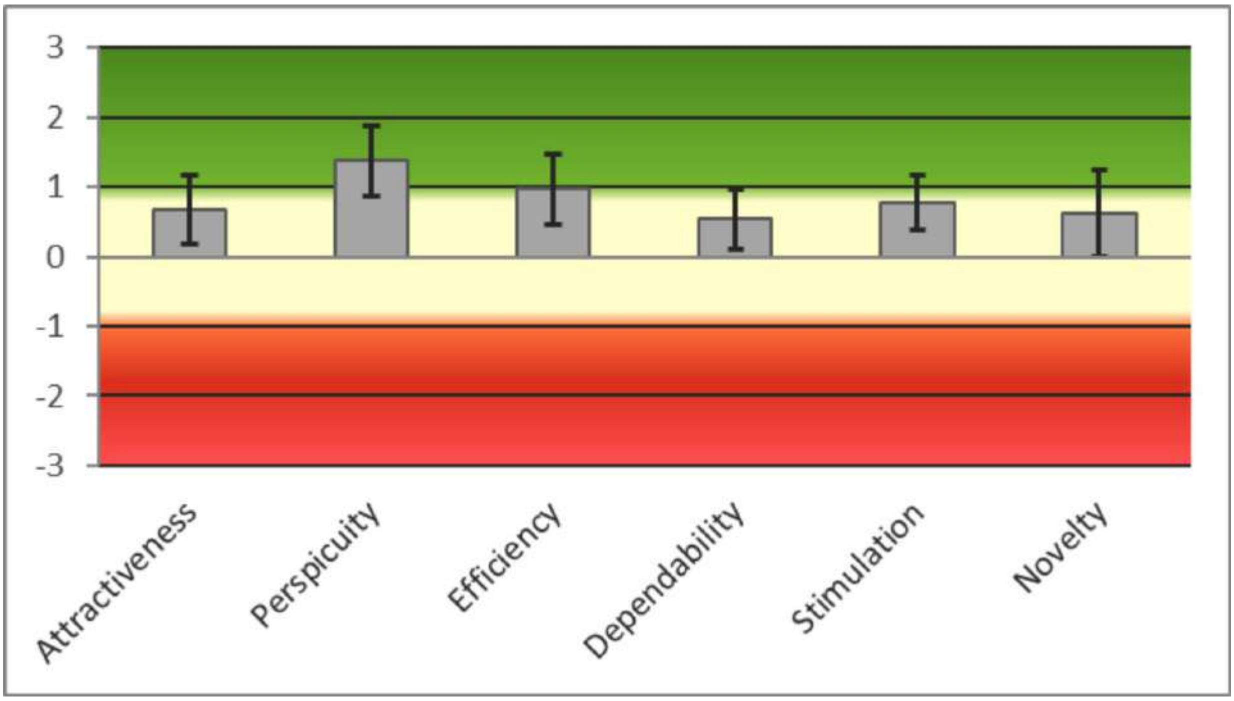


Figure 13: Experimental Group (LLM) UEQ results-six subscales

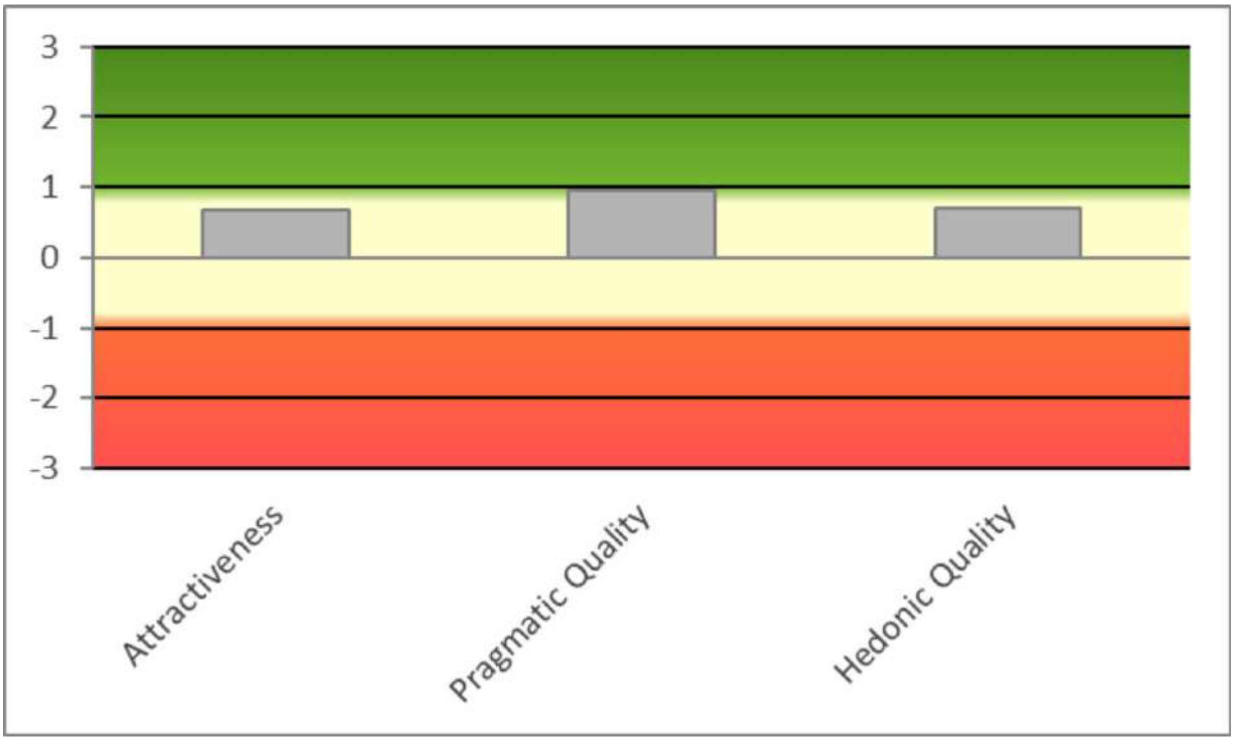


Figure 14: Experimental Group (LLM) UEQ results-grouped scales

4.5.3.2.2. Control Group (no-LLM) Results

Overall UEQ Scales (Figure 15)

- Attractiveness: Mean = 0.833, Variance = 1.16
- Perspicuity: Mean = 1.632, Variance = 1.38
- Efficiency: Mean = 1.000, Variance = 1.77
- Dependability: Mean = 0.676, Variance = 1.22
- Stimulation: Mean = 0.471, Variance = 1.34
- Novelty: Mean = 0.221, Variance = 1.82

Grouped UEQ scales (Figure 16)

- Pragmatic Quality: Mean = 0.83
- Hedonic Quality: Mean = 1.10
- Attractiveness: Mean = 0.35

Interpretation of Results

The User Experience Questionnaire (UEQ) scales show that the video watching and ideas taking system received an attractiveness mean score of 0.833 with a variance of 1.16, indicating a generally positive user experience. Perspicuity scored highest with a mean of 1.632 and a variance of 1.38, reflecting that users found the system very easy to understand and use. Efficiency also scored positively, with a mean of 1.000 and a variance of 1.77, suggesting that the system effectively helps users achieve their tasks. Dependability, with a mean score of 0.676 and a variance of 1.22, shows mixed feelings about the system's reliability. The stimulation score of 0.471 with a variance of 1.34 suggests that the system is somewhat engaging, while the novelty score of 0.221 and a variance of 1.82 indicates that users found the system only slightly innovative.

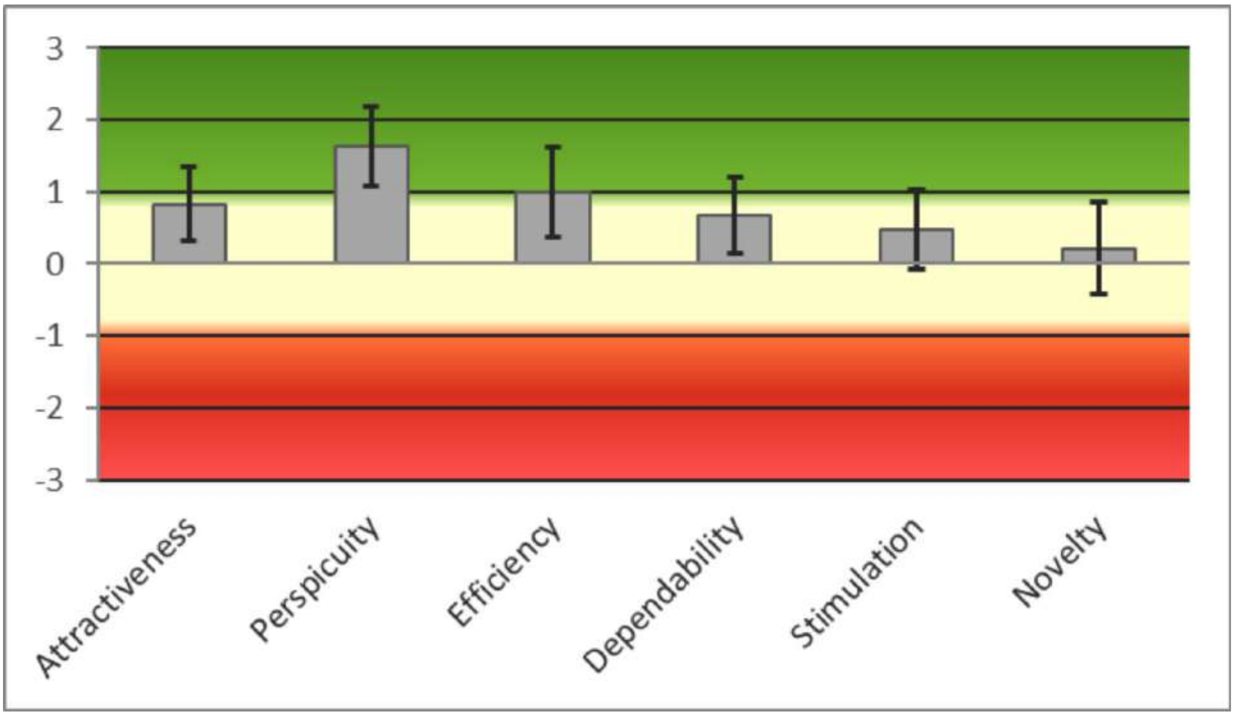


Figure 15: Control Group (LLM) UEQ results-six subscales

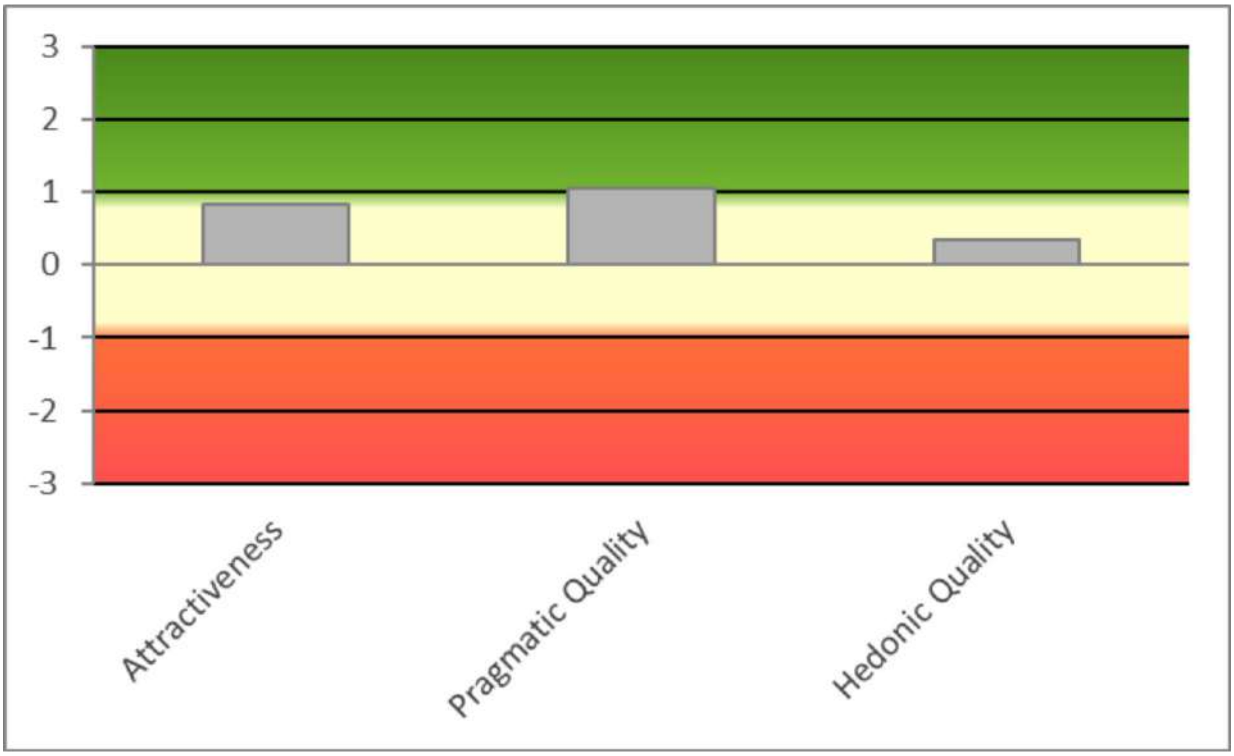


Figure 16: Control Group (no-LLM) UEQ results-grouped scales

4.5.3.3. NASA TLX

Shapiro-Wilk test of normality was conducted. Shapiro-Wilk test showed a statistic of 0.960 with 18 degrees of freedom and a significance level of 0.603. Both tests show non-significant results (p > 0.05), indicating that the distribution of NASA TLX workload scores does not significantly deviate from normality. For the control group, the test showed a statistic of 0.939 with 17 degrees of freedom and a significance level of 0.305. Tests among two groups indicated that the distribution of NASA TLX workload scores does not significantly deviate from normality.

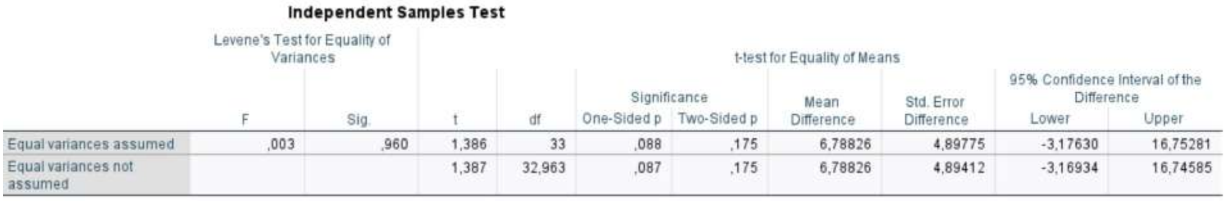
4.5.3.3.1. NASA TLX Experimental Group (LLM) Descriptives

The NASA TLX workload scores of the LLM group (N=18) reveal that the mean workload score is 49.9996 with a standard error of 3.45482. The mean score of approximately 50 out of 100 indicates a moderate perceived workload among participants. The 95% confidence interval for the mean ranges from 42.7106 to 57.2887. The 5% trimmed mean is slightly lower at 49.7848, while the median is higher at 53.3333, which indicates that more than half of the scores are above the average, pointing towards a slightly positive skew in the data distribution. The variance is 214.844, and the standard deviation is 14.65755, indicating a relatively wide variation among participants' scores. The minimum score recorded is 26.47, and the maximum is 77.40, resulting in a range of 50.93. The interquartile range of middle 50% of data is 24.93. The skewness is close to zero at 0.020, with a standard error of 0.536, indicating a nearly symmetric distribution. The kurtosis is -0.742 with a standard error of 1.038, suggesting a distribution that is slightly flatter than normal.

4.5.3.3.2. NASA TLX Control Group (no-LLM) Descriptives

The NASA TLX workload scores of the control group, consisting of 17 participants, indicate a mean workload score of 43.2114 with a standard error of 3.46651. The 95% confidence interval for the mean ranges from 35.8627 to 50.5600, suggesting a moderate level of workload. The 5% trimmed mean is slightly higher at 43.7571, and the median score is 43.7333, indicating that the distribution of scores is fairly symmetric. The variance is 204.283, and the standard deviation is 14.29278, showing a considerable spread in the scores which indicates diverse perceptions of workload within the group. The minimum score recorded is 14.60, and the maximum is 62.00, resulting in a range of 47.40. The interquartile range is 23.73. These show that while most data points are clustered (as shown by the IQR), there are still wide variations in the extremes (as shown by the total range). The skewness is -0.276 with a standard error of 0.550, indicating a slight negative skew in the distribution, while the kurtosis is -0.904 with a standard error of 1.063, suggesting a distribution that is slightly flatter than normal.

4.5.3.3.3. NASA TLX: Independent samples t-Test



Independent Samples Test

	Levene's Test for Equality of Variances		t-test for Equality of Means							
					Significance		Mean	Std. Error	95% Confidence Interval of the Difference	
	F	Sig.	t	df	One-Sided p	Two-Sided p	Difference	Difference	Lower	Upper
Equal variances assumed	,003	,960	1,386	33	,088	,175	6,78826	4,89775	-3,17630	16,75281
Equal variances not assumed			1,387	32,963	,087	,175	6,78826	4,89412	-3,16934	16,74585

Figure 17: NASA TLX t-test results

An independent samples t-test was conducted to compare the workload scores between the two groups (Figure 17) . Levene's test for equality of variances showed no significant difference (F = 0.003, p = 0.960), indicating that the variances are equal. The t-test for equality of means revealed that there was no statistically significant difference in the workload scores between the two groups (t(33) = 1.386, p = 0.175) with a mean difference of 6.78826 and a 95% confidence interval ranging from -3.17630 to 16.75281.

4.5.3.4. UTAUT

The normality tests indicate that the variables Effort expectancy (both groups - p=0.066; p=0.408), Attitude toward using technology (both groups - p=0.054; p=0.233), and Anxiety (both groups - p=0.284; p=0.118) approximately follow a normal distribution. However, Performance expectancy (both groups - p=0.006; p=0.015) does not follow normal distribution, while Behavioral intention (control group - p=0.002) (experimental group - p-value=0.517) is divided.

For the experimental group, performance expectancy, measured by four items, had a Cronbach's alpha of 0.717, indicating acceptable internal consistency. Effort expectancy, also measured by four items, yielded a Cronbach's alpha of 0.645, suggesting moderate reliability. The construct of attitude toward using technology, with four items, demonstrated a high reliability with a Cronbach's alpha of 0.858. Anxiety, assessed through four items, showed good internal consistency with a Cronbach's alpha of 0.785.

Finally, behavioral intention, evaluated with three items, achieved a high reliability score of 0.859. These reliability scores indicate that the scales used in the questionnaire for Group A are generally reliable and suitable for assessing the constructs within the UTAUT model.

Similarly, for the control group, performance expectancy, measured by four items, had a Cronbach's alpha of 0.717, indicating acceptable internal consistency. Effort expectancy, also measured by four items, yielded a Cronbach's alpha of 0.645, suggesting moderate reliability. The construct of attitude toward using technology, with four items, demonstrated a high reliability with a Cronbach's alpha of 0.858. Anxiety, assessed through four items, showed good internal consistency with a Cronbach's alpha of 0.785. Finally, behavioral intention, evaluated with three items, achieved a high reliability score of 0.859. These reliability scores indicate that the scales used in the questionnaire for the control group are generally reliable and suitable for assessing the constructs within the UTAUT model.

4.5.3.4.1. UTAUT Independent Samples t-Test

The independent samples t-test results indicated no significant differences in means between groups for Effort Expectancy (p=0.110), Attitude Toward Using Technology (p=0.327), and Anxiety (p=0.268), as all two-sided p-values are greater than 0.05. Levene's tests confirmed the assumption of equal variances in all cases. The effect size estimates (Cohen's d, Hedges' correction, Glass's delta) for Effort Expectancy, Attitude Toward Using Technology, and Anxiety also showed relatively small effects, with confidence intervals crossing zero, further supporting the conclusion that the differences between groups for these constructs are not statistically significant.

4.5.3.4.1. Mann-Whitney U Test

A Mann-Whitney U test was conducted to evaluate differences in performance expectancy and behavioral intention across the experimental and control groups. The null hypothesis stated that the distribution of UTAUT_mean_Performance_expectancy is the same across groups (Figure 18). The test results indicated a significance level of 0.909, leading to the retention of the null hypothesis, suggesting no significant difference in performance expectancy between the groups.

For behavioral intention (Figure 19), the null hypothesis proposed that the distribution of UTAUT_mean_Behavioral_intention is the same across groups. The test results showed a significance level of 0.660, also leading to the retention of the null hypothesis, indicating no significant difference in behavioral intention between the groups.

Performance Expectancy

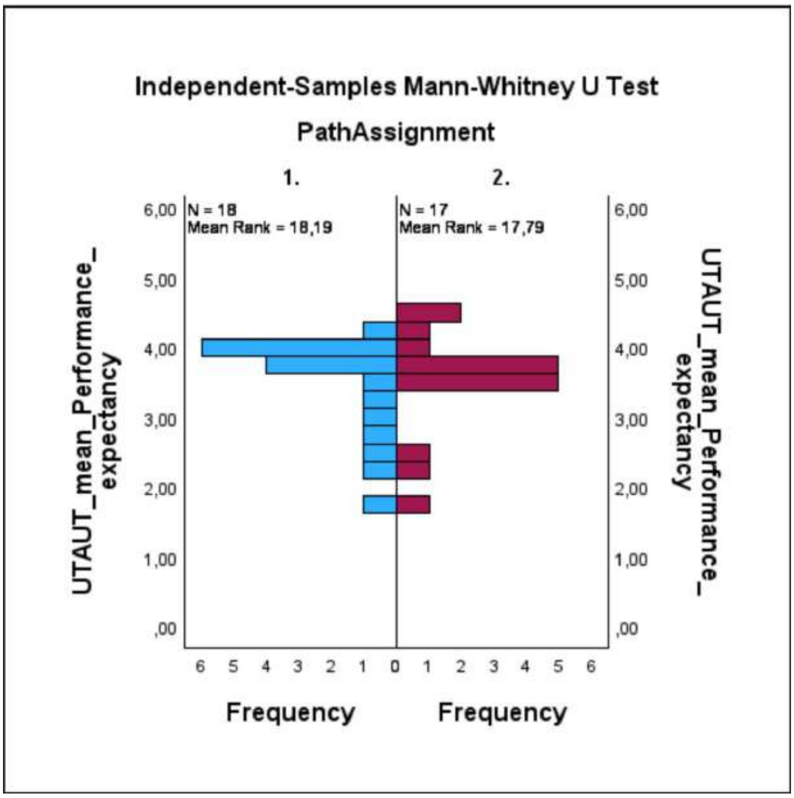


Figure 18: UTAUT Performance expectancy results

Behavioral Intention

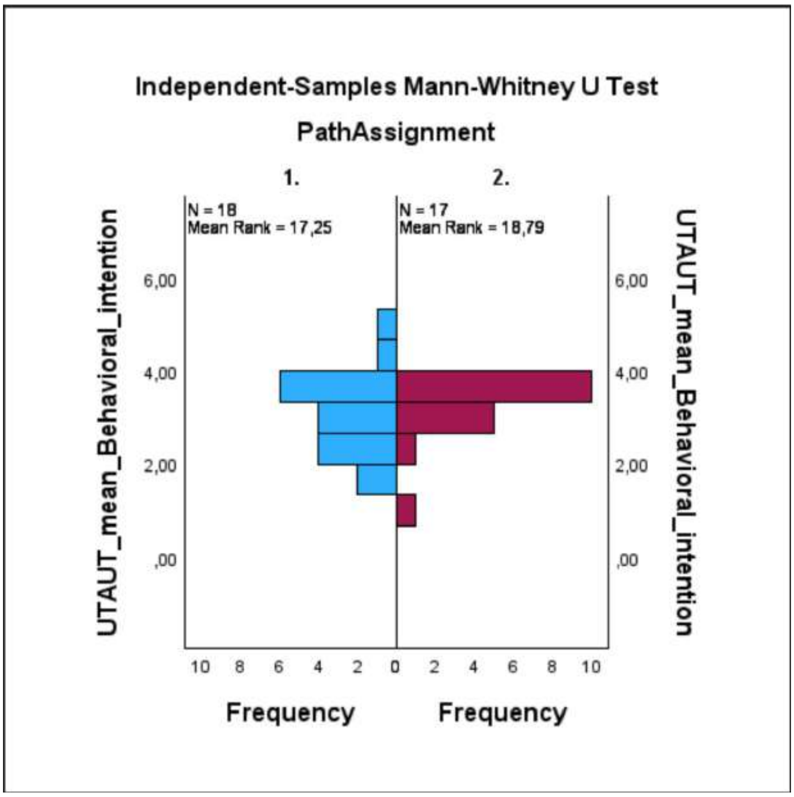


Figure 19: UTAUT Behavioral intention results

5. Discussion

The project aimed to rigorously experiment, test, and validate using two examples of video-based design cases to explore if the chatbot on ideation. However, the main goal was not detailed data analysis. During the preliminary data analysis, the user experience analysis, focusing on the chatbot’s capabilities to expand ideation, acceptance of technology, and workload, returned non-significant results. Users found the chatbot somewhat untrustworthy and unreliable, making it difficult to draw definite conclusions about its efficacy from the preliminary findings. Therefore, this discussion focuses on qualitative insights from interviews, framing the discussion around the needs, wants, and requirements of human-AI interaction.

Participants' attitudes towards AI and their experiences during the experiment revealed several implications. AI can potentially support ideation by generating a broad range of ideas and handling repetitive tasks, but it must provide more contextually relevant and accurate information. This aligns with Suh et al. (2023), who argue that AI tools should complement human creativity rather than replace it. Ensuring AI tools offer relevant, context-aware assistance while allowing designers to retain control is crucial for their successful integration into the design workflow, echoing the findings of Ding & Chan (2023). Otherwise, AI might introduce unnecessary and unproductive workload. Participants noted that while AI efficiently structured information, it often provided irrelevant data, leading to information overload. Future AI developments should enhance contextual understanding and relevance filtering to mitigate this overload, striking a better balance between AI efficiency and human cognitive processes.

Background research and this study found that designers generally have mixed to optimistic opinions about AI in design. Observations and qualitative analysis supported this, suggesting cautious optimism about AI's potential to enhance ideation, despite concerns about biases, privacy issues, and the need for transparent and accountable AI use. Engagement and balance were crucial, with participants appreciating innovation and immersive nature of egocentric videos, but stressing the necessity of understanding context. AI systems might need to be designed to toggle between detailed immersion and broader context views, providing a comprehensive understanding without losing nuance.

A recurring theme was that AI is seen as a teammate or co-pilot, helping with manual repetitive tasks or organizing information. Designers expressed a desire to retain decision-making control in the design process, avoiding a blur of creativity lines. AI should support without overshadowing human input, akin to Alfred supporting Batman —knowledgeable and helpful but not the primary decision-maker. Current AI lacks the capacity to align with designers' preferences beyond general aggregated information, raising questions about how AI can collaborate without replacing designers' thought processes. Over-reliance on AI could lead to lower-quality solutions when designers are tired or overworked. Ensuring designers remain independent in their thought processes is essential, potentially through co-creation, where AI aids but does not dominate creativity.

The risks of cognitive complacency or de-skilling due to reliance on AI are significant. Designers must remain critical actors in the process, making key decisions and prioritizing tasks. Continuous learning and skill development are necessary for effective AI collaboration, as noted by Ding & Chan (2023).

In conclusion, while designers view AI as offereing potential to support and enhance the design process, its implementation must preserve and support human creativity and control. Participants highlighted the significance of efficiently structured information, yet AI tools often provided irrelevant data, leading to information overload. Future developments should focus on enhancing contextual understanding to balance AI efficiency with human intuition. Engagement and balance emerged as significant, with a need for improved AI capabilities that deliver a complete understanding without losing nuance. Inspiration processes relied on primary research through human interaction and environmental observations, supported by secondary research like online visuals and prior work reviews. Future AI tools could better support these approaches by simulating human interaction and providing richer, context-aware content. Participants viewed AI as a productivity enhancer but emphasized context accuracy. Concerns about integrity—authenticity, privacy, and ethical use—underscore the need for refined ethical standards in AI systems. Emphasizing empathy and diverse perspectives indicates a future where AI must evolve to include emotional intelligence and multi-perspective analysis. Thus, AI integration in design ideation holds promising potential, provided future developments address these complex needs, aligning AI's capabilities with human intuition, empathy, and ethical integrity.

5.1. Study Limitations

This research faced several limitations that could impact the effectiveness and applicability of the findings:

1. The drawback of AI to accurately interpret and apply vast and nuanced terminology from design theory and methodology was lowered by incorporating an introductory training session for participants on how to effectively interact with the AI, providing them opportunities to ask questions and understand limitations of the AI. Furthermore, the AI was supplemented by a database of design books.
2. Risk of potential biases inherent in the applied model or during the analysis of results, which could skew the outcomes of the study. One of the models used was a GPT-4 which has a broad, diverse and representative dataset. This minimized the AI's inherent biases to not significantly affect the validity and reliability of the research outcomes. Moreover, the qualitative nature of the textual inputs and outputs allowed for a more detailed interpretation of the AI's outputs, where human judgment plays a primary role in contextualizing and understanding the data.
3. Current capabilities of applied AI model may not be sufficiently developed to handle the complex demands of design ideation, which could lead to less effective or misleading insights. To enhance and streamline ideation to capabilities of AI, iterative refinement, exploration of design ideation methods and continuous integration of feedback from two pilot sessions were employed.
4. The methodology chosen for exploring design space may not be optimal, as there are multiple unexplored ideation methods which designers employ either in parallel or linearly during the design process. Therefore, the ideation method chosen for the study was the one deemed most appropriate within the limits of a desktop screen.

5. Unforeseen participant variables could introduce further complexities into results, limiting the generalizability and applicability of the research findings. To counter this, the experiment was carried out and controlled by rigorous procedure, controlled environment and targeted sample.

5.2. Future Work

To enhance contextual understanding and interpretation, further research is needed to train models to provide more contextually aware insights for designers during the ideation phase. This involves exploring the application of Large Language Models (LLMs) and Vision-Language Models (VLMs) to boost creativity and inspiration in ideation. Studies should focus on how AI tools can support designers in the early stages of ideation, including brainstorming and conceptualization, and examine if these tools can facilitate both divergent (idea generation) and convergent (idea refinement) thinking processes in design.

In enhancing co-creation and collaboration, frameworks should be established that define co-creation processes between humans and AI, ensuring that AI acts as an augmentative tool supporting human creativity. This includes setting guidelines and protocols for human-AI collaboration, covering roles, decision-making processes, and creative control. Additionally, there is a need to educate designers on the capabilities and limitations of AI tools to enhance their ability to interact with AI models and maximize creative outcomes.

Regarding limitations and biases, transparency and accountability should be ensured in AI-assisted ideation to build trust among designers. This involves implementing transparent documentation and reporting systems for AI decisions and suggestions. Similarly, designer education should focus on informing designers about the capabilities and limitations of AI tools, enhancing their interaction with AI models to maximize creativity.

6. References

1. Alghamdi, E., Velloso, E., & Gruba, P. (2021). AUVANA: An Automated Video Analysis Tool for Visual Complexity. https://doi.org/10.31219/osf.io/kj9hx
2. Asadi, A. R. (2023). LLMs in Design Thinking: Autoethnographic Insights and Design Implications. https://doi.org/10.1145/3631991.3631999
3. Assari, S. M., Zamir, A. R., & Shah, M. (2014). Video Classification Using Semantic Concept Co-occurrences. https://doi.org/10.1109/cvpr.2014.324
4. Batalas, N., Bruikman, H., Van Drunen, A., Huang, H., Turzynska, D., Vakili, V., Voynarovskaya, N., & Markopoulos, P. (2012). On the Use of Video Prototyping in Designing Ambient User Experiences. In Lecture notes in computer science (pp. 403–408). https://doi.org/10.1007/978-3-642-34898-3_34
5. Borges, A., Zatt, B., Porto, M., & Correa, G. (2024). A systematic literature review on video transcoding acceleration: challenges, solutions, and trends. Multimedia Tools and Applications. https://doi.org/10.1007/s11042-023-17862-w
6. Brophy, D. R. (2001). Comparing the attributes, activities, and performance of divergent, convergent, and combination thinkers. Creativity Research Journal, 13(3–4), 439–455. https://doi.org/10.1207/s15326934crj1334_20
7. Chang, N. S., Deufemia, V., Polese, G., & Vacca, M. (2007). A normalization framework for multimedia databases. IEEE Transactions on Knowledge and Data Engineering, 19(12), 1666–1679. https://doi.org/10.1109/tkde.2007.190651
8. Chen, W. (2015). Exploring the learning problems and resource usage of undergraduate industrial design students in design studio courses. International Journal of Technology and Design Education, 26(3), 461–487. https://doi.org/10.1007/s10798-015-9315-2
9. Cross, N. (2004). Expertise in design: an overview. Design Studies, 25(5), 427–441. https://doi.org/10.1016/j.destud.2004.06.002
10. Dam, R. F., & Siang, T. Y. (2024, July 6). Learn how to use the best ideation methods: brainstorming, braindumping, brainwriting, and brainwalking. The Interaction Design Foundation. https://www.interaction-design.org/literature/article/learn-how-to-use-the-best-ideation-methods-brainstorming-braindumping-brainwriting-and-brainwalking
11. Damnjanovic, I., & Trow, I. (2023). Determining Video Complexity to optimise Video Quality Assessment. https://doi.org/10.1145/3588444.3591013
12. Dang, H., Benharrak, K., Lehmann, F., & Buschek, D. (2022a). Beyond Text Generation: Supporting Writers with Continuous Automatic Text Summaries. https://doi.org/10.1145/3526113.3545672
13. Dang, H., Benharrak, K., Lehmann, F., & Buschek, D. (2022b, October 28). Beyond Text Generation: Supporting Writers with Continuous Automatic Text Summaries. https://doi.org/10.1145/3526113.3545672
14. Dazkir, S. S., Mower, J. M., Reddy-Best, K., & Pedersen, E. L. (2013). An exploration of design students' inspiration process. ResearchGate. https://www.researchgate.net/publication/260124131_An_exploration_of_design_students'_inspiration_process
15. Dewitte, S., Pandelaere, M., Briers, B., & Warlop, L. (2005). Cognitive Load has Negative After Effects on Consumer Decision Making. Social Science Research Network. https://doi.org/10.2139/ssrn.813684

16. Ding, Z., & Chan, J. (2023a). Mapping the design space of interactions in Human-AI text co-creation tasks. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2303.06430
17. Ding, Z., & Chan, J. (2023b). Mapping the design space of interactions in Human-AI text co-creation tasks. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2303.06430
18. Duan, J., Zhang, M., Wang, J., Han, S., Chen, X., & Yang, X. (2020). VCC-DASH: A Video Content Complexity-Aware DASH Bitrate Adaptation Strategy. Electronics, 9(2), 230. https://doi.org/10.3390/electronics9020230
19. Eastman, C. M. (2001). New Directions in Design Cognition: Studies of Representation and recall. ResearchGate. https://www.researchgate.net/publication/246935473_New_Directions_in_Design_Cognition_Studies_of_Representation_and_Recall
20. Eckert, C., & Stacey, M. (2000). Sources of inspiration: a language of design. Design Studies, 21(5), 523–538. https://doi.org/10.1016/s0142-694x(00)00022-3
21. Ekenel, H. K., Semela, T., & Stiefelhagen, R. (2010). Content-based video genre classification using multiple cues. https://doi.org/10.1145/1877850.1877858
22. ElMaraghy, W., ElMaraghy, H., Tomiyama, T., & Monostori, L. (2012). Complexity in engineering design and manufacturing. CIRP Annals, 61(2), 793–814. https://doi.org/10.1016/j.cirp.2012.05.001
23. Free online appointment scheduling software | Calendly. (n.d.). Calendly.com. https://calendly.com/
24. George, T. (2023, June 22). Types of interviews in research | Guide & Examples. scribbr.com. https://www.scribbr.com/methodology/interviews-research/
25. Gonçalves, M. (2016). Decoding designers' inspiration process. https://doi.org/10.4233/uuid:a270cdf2-d46b-4085-8f4f-328b823ccdee
26. Green Video Complexity Analysis for efficient encoding in adaptive video streaming. (n.d.). Ar5iv. https://ar5iv.labs.arxiv.org/html/2304.12384
27. Guilford, J. G. (1967). The Nature of Human Intelligence. J. P. Guilford. McGraw-Hill, New York. Science.
28. Guilford, J. P. (1950). Creativity. American Psychologist/ the American Psychologist, 5(9), 444–454. https://doi.org/10.1037/h0063487
29. Halskov, K., & Nielsen, R. (2006). Virtual Video Prototyping. Human-computer Interaction, 21(2), 199–233. https://doi.org/10.1207/s15327051hci2102_2
30. Harrison, S., Minneman, S., Stults, B., & Weber, K. (1990). Video. SIGCHI Bulletin/ACM SIGCHI Bulletin, 21(3), 86–90. https://doi.org/10.1145/379088.379103
31. Hines, A., Gillen, E., Kelly, D., Skoglund, J., Kokaram, A., & Harte, N. (2014). Perceived Audio Quality for Streaming Stereo Music. https://doi.org/10.1145/2647868.2655025
32. Hong, Y., Zhen, H., Chen, P., Zheng, S., Du, Y., Chen, Z., & Gan, C. (2023). 3D-LLM: Injecting the 3D World into Large Language Models. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2307.12981
33. Höök, K., Caramiaux, B., Erkut, C., Forlizzi, J., Hajinejad, N., Haller, M., Hummels, C. C. M., Isbister, K., Jonsson, M., Khut, G., Loke, L., Lottridge, D., Marti, P., Melcer, E., Müller, F. F., Petersen, M. G., Schiphorst, T., Segura, E. M., Ståhl, A., . . . Tobiasson, H. (2018). Embracing First-Person perspectives in Soma-Based design. Informatics, 5(1), 8. https://doi.org/10.3390/informatics5010008
34. Hu, W., Xu, Y., Li, Y., Li, W., Chen, Z., & Tu, Z. (2023). BLIVA: a simple multimodal LLM for better handling of Text-Rich visual questions. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2308.09936

35. Human Research Ethics. (n.d.). TU Delft. https://www.tudelft.nl/over-tu-delft/strategie/integriteitsbeleid/human-research-ethics
36. Kalyan, K. S. (2023). A survey of GPT-3 family large language models including ChatGPT and GPT-4. Social Science Research Network. https://doi.org/10.2139/ssrn.4593895
37. Kersten, W. C., Diehl, J. C., & Van Engelen, J. M. (2018). Facing complexity through varying the clarification of the design task. Formakademisk, 11(4). https://doi.org/10.7577/formakademisk.2621
38. Korhonen, J., & Reiter, U. (2009). Analysis on the perceptual impact of bit errors in practical video streaming applications. https://doi.org/10.1109/imsaa.2009.5439481
39. Leahy, K., Daly, S. R., Murray, J. K., McKilligan, S., & Seifert, C. M. (2018). Transforming early concepts with Design Heuristics. International Journal of Technology and Design Education, 29(4), 759–779. https://doi.org/10.1007/s10798-018-9473-0
40. Li, J., Chen, D., Hong, Y., Chen, Z., Chen, P., Shen, Y., & Gan, C. (2023a). COVLM: Composing visual entities and relationships in large language models via communicative decoding. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2311.03354
41. Li, J., Chen, D., Hong, Y., Chen, Z., Chen, P., Shen, Y., & Gan, C. (2023b). COVLM: Composing visual entities and relationships in large language models via communicative decoding. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2311.03354
42. Licklider, J. C. R. (1960). Man-Computer symbiosis. IRE Transactions on Human Factors in Electronics, HFE-1(1), 4–11. https://doi.org/10.1109/thfe2.1960.4503259
43. Makatura, L., Foshey, M., Wang, B., HähnLein, F., Ma, P., Deng, B., Tjandrasuwita, M., Spielberg, A., Owens, C. E., Chen, P. Y., Zhao, A., Zhu, A., Norton, W. J., Gu, E., Jacob, J., Li, Y., Schulz, A., & Matusik, W. (2023). How can large language models help humans in design and manufacturing? arXiv (Cornell University). https://doi.org/10.48550/arxiv.2307.14377
44. Mednick, S. (1962). The associative basis of the creative process. Psychological Review, 69(3), 220–232. https://doi.org/10.1037/h0048850
45. Menon, V. V., Feldmann, C., Amirpour, H., Ghanbari, M., & Timmerer, C. (2022). VCA. https://doi.org/10.1145/3524273.3532896
46. Moore, J. (2005a). Exploring how user video supports design. Nordic Design Research Conference. https://doi.org/10.21606/nordes.2005.043
47. Moore, J. (2005b). Exploring how user video supports design. Nordic Design Research Conference. https://doi.org/10.21606/nordes.2005.043
48. Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2023, July 12). A comprehensive overview of large language models. arXiv.org. https://arxiv.org/abs/2307.06435
49. Neon - Technical Specifications - Neon eye tracking module and frames. (n.d.). https://pupil-labs.com/products/neon/specs/
50. Patil, P., Pawar, V., Pawar, Y., & Pisal, S. (2021a). Video Content Classification using Deep Learning. https://www.semanticscholar.org/paper/Video-Content-Classification-using-Deep-Learning-Patil-Pawar/fa0b09c42eeffdef67f80b784e0f4d5391f2bc07
51. Patil, P., Pawar, V., Pawar, Y., & Pisal, S. (2021b). Video Content Classification using Deep Learning. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2111.13813

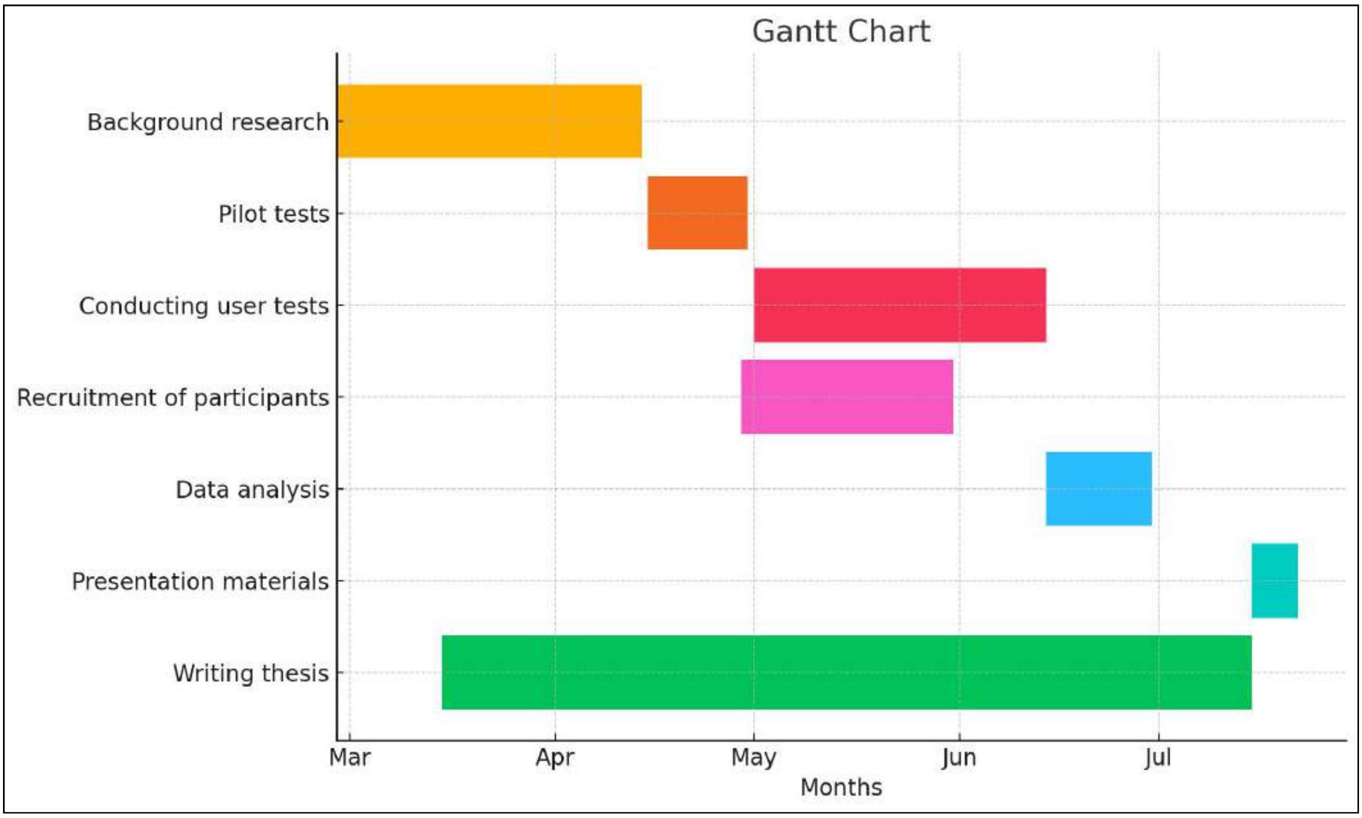
52. Peng, I., Lin, M., Chen, Y., Yang, F., & Su, A. Y. (2013, July 1). Improvement of Streaming Video in Differential Service Networks by Using Opportunity RED Mechanism. https://doi.org/10.1109/cisis.2013.116
53. Plucker, J. A., & Beghetto, R. A. (2006). Why creativity is domain general, why it looks domain specific, and why the distinction does not matter. In American Psychological Association eBooks (pp. 153–167). https://doi.org/10.1037/10692-009
54. Ramesh, I., Sivakumar, I., Ramesh, K., Venkatesh, V. P. P., & Vetriselvi, V. (2020). Categorization of YouTube Videos by Video Sampling and Keyword Processing. https://doi.org/10.1109/iccsp48568.2020.9182158
55. Roach, M., Mason, J., & Pawlewski, M. (2002). Video genre classification using dynamics. https://doi.org/10.1109/icassp.2001.941230
56. Rouvier, M., Linarès, G., & Matrouf, D. (2009). Robust audio-based classification of video genre. https://doi.org/10.21437/interspeech.2009-337
57. Setchi, R., & Bouchard, C. (2010). In search of Design Inspiration: A Semantic-Based approach. Journal of Computing and Information Science in Engineering, 10(3). https://doi.org/10.1115/1.3482061
58. Shamsi, F., Muhammad, S., & Shaikh, S. (2019a). Content-based automatic video genre identification. International Journal of Advanced Computer Science and Applications/International Journal of Advanced Computer Science & Applications, 10(6). https://doi.org/10.14569/ijacsa.2019.0100677
59. Shamsi, F., Muhammad, S., & Shaikh, S. (2019b). Content-based automatic video genre identification. International Journal of Advanced Computer Science and Applications/International Journal of Advanced Computer Science & Applications, 10(6). https://doi.org/10.14569/ijacsa.2019.0100677
60. Shyu, M., Xie, Z., Chen, M., & Chen, S. (2008). Video Semantic Event/Concept detection using a Subspace-Based multimedia data mining framework. IEEE Transactions on Multimedia, 10(2), 252–259. https://doi.org/10.1109/tmm.2007.911830
61. Sternberg, R. J., Sternberg, R. J., Ackerman, P. L., Bouchard, T. J., Ceci, S. J., Conway, A. R. A., Deary, I. J., Ericsson, K. A., Flynn, J. R., Gardner, H., Gottfredson, L. S., Grigorenko, E. L., Haier, R. J., Halpern, D. F., Kaufman, A. S., Kaufman, S. B., Lubinski, D., Lynn, R., Mayer, J. D., . . . Sternberg, R. J. (2018). The nature of human intelligence. In Cambridge University Press eBooks. https://doi.org/10.1017/9781316817049
62. Studying what people do. (2007). In Springer eBooks (pp. 36–85). https://doi.org/10.1007/978-1-84628-961-3_2
63. Su, S., Hong, J. P., Shi, J., & Park, H. S. (2017, July 1). Predicting Behaviors of Basketball Players from First Person Videos. https://doi.org/10.1109/cvpr.2017.133
64. Suh, S., Chen, M., Min, B., Li, T. J., & Xia, H. (2023a). Structured Generation and Exploration of Design Space with Large Language Models for Human-AI Co-Creation. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2310.12953
65. Suh, S., Chen, M., Min, B., Li, T. J., & Xia, H. (2023b). Structured Generation and Exploration of Design Space with Large Language Models for Human-AI Co-Creation. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2310.12953
66. Suh, S., Chen, M., Min, B., Li, T. J., & Xia, H. (2023c). Structured Generation and Exploration of Design Space with Large Language Models for Human-AI Co-Creation. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2310.12953
67. The NASA TLX Tool: Task Load Index. (n.d.). TLX @ NASA Ames - NASA TLX Paper/Pencil Version. https://humansystems.arc.nasa.gov/groups/tlx/tlxpaperpencil.php

68. Thomas, R., & Izatt, J. (2003). A taxomony of Engineering Design tasks and its applicability to university Engineering Education. European Journal of Engineering Education, 28(4), 535–547. https://doi.org/10.1080/0304379032000101872
69. User Acceptance of Information Technology: Toward a unified View on JSTOR. (n.d.-a). www.jstor.org. https://www.jstor.org/stable/30036540
70. User Acceptance of Information Technology: Toward a unified View on JSTOR. (n.d.-b). www.jstor.org. https://www.jstor.org/stable/30036540
71. User Experience Questionnaire (UEQ). (n.d.). https://www.ueq-online.org/
72. Welcome to EGO4D! | Ego4D. (n.d.). https://ego4d-data.org/docs/
73. What is Brainwriting. (2024, March 22). The Interaction Design Foundation. https://www.interaction-design.org/literature/topics/brainwriting
74. What is Divergent Thinking? (2024, March 28). The Interaction Design Foundation. https://www.interaction-design.org/literature/topics/divergent-thinking
75. Whitney, P., Rinehart, C. A., & Hinson, J. M. (2008). Framing effects under cognitive load: The role of working memory in risky decisions. Psychonomic Bulletin & Review, 15(6), 1179–1184. https://doi.org/10.3758/pbr.15.6.1179
76. Woelfel, C., Krzywinski, J., & Drechsel, F. (2013). Knowing, reasoning and visualizing in industrial design. Knowledge Engineering Review, 28(3), 287–302. https://doi.org/10.1017/s0269888913000258
77. Xu, Y., Zhou, Y., & Chiu, D. (2014). Analytical QOE models for Bit-Rate switching in dynamic adaptive streaming systems. IEEE Transactions on Mobile Computing, 13(12), 2734–2748. https://doi.org/10.1109/tmc.2014.2307323
78. Yan, N. Y., Ricci, E., Liu, N. G., & Sebe, N. (2015). Egocentric daily activity recognition via multitask clustering. IEEE Transactions on Image Processing, 24(10), 2984–2995. https://doi.org/10.1109/tip.2015.2438540
79. Yew, J., & Shamma, D. A. (2011). Know your data: understanding implicit usage versus explicit action in video content classification. Proceedings of SPIE, the International Society for Optical Engineering/Proceedings of SPIE. https://doi.org/10.1117/12.878807
80. Zhang, D., Yu, Y., Dong, J., Li, C., Su, D., Chu, C., & Yu, D. (2024, January 24). MM-LLMS: Recent Advances in MultiModal Large Language Models. arXiv.org. https://arxiv.org/abs/2401.13601
81. Zhao, H., Cai, Z., Si, S., Ma, X., An, K., Chen, L., Liu, Z., Wang, S., Han, W., & Chang, B. (2023). MMICL: Empowering Vision-language Model with Multi-Modal In-Context Learning. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2309.07915
82. Zhao, S., Yao, H., & Sun, X. (2013). Video classification and recommendation based on affective analysis of viewers. Neurocomputing, 119, 101–110. https://doi.org/10.1016/j.neucom.2012.04.042
83. Zhou, C., Chai, C., & Liao, J. (2021). Analysis of problem decomposition strategies of novice industrial designers using network-based cognitive maps. International Journal of Technology and Design Education, 32(2), 1293–1315. https://doi.org/10.1007/s10798-020-09647-1
84. Zhou, C., Zhang, X., & Yu, C. (2023). How does AI promote design iteration? The optimal time to integrate AI into the design process. Journal of Engineering Design, 1–28. https://doi.org/10.1080/09544828.2023.2290915
85. Zhou, G., Hong, Y., & Wu, Q. (2023). NavGPT: Explicit Reasoning in Vision-and-Language Navigation with Large Language Models. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2305.16986
86. Zimmerman, J. (2005). Video sketches: exploring pervasive computing interaction designs. IEEE Pervasive Computing, 4(4), 91–94. https://doi.org/10.1109/mprv.2005.91

68. Thomas, R., & Izatt, J. (2003). A taxomony of Engineering Design tasks and its applicability to university Engineering Education. European Journal of Engineering Education, 28(4), 535–547. https://doi.org/10.1080/0304379032000101872
69. User Acceptance of Information Technology: Toward a unified View on JSTOR. (n.d.-a). www.jstor.org. https://www.jstor.org/stable/30036540
70. User Acceptance of Information Technology: Toward a unified View on JSTOR. (n.d.-b). www.jstor.org. https://www.jstor.org/stable/30036540
71. User Experience Questionnaire (UEQ). (n.d.). https://www.ueq-online.org/
72. Welcome to EGO4D! | Ego4D. (n.d.). https://ego4d-data.org/docs/
73. What is Brainwriting. (2024, March 22). The Interaction Design Foundation. https://www.interaction-design.org/literature/topics/brainwriting
74. What is Divergent Thinking? (2024, March 28). The Interaction Design Foundation. https://www.interaction-design.org/literature/topics/divergent-thinking
75. Whitney, P., Rinehart, C. A., & Hinson, J. M. (2008). Framing effects under cognitive load: The role of working memory in risky decisions. Psychonomic Bulletin & Review, 15(6), 1179–1184. https://doi.org/10.3758/pbr.15.6.1179
76. Woelfel, C., Krzywinski, J., & Drechsel, F. (2013). Knowing, reasoning and visualizing in industrial design. Knowledge Engineering Review, 28(3), 287–302. https://doi.org/10.1017/s0269888913000258
77. Xu, Y., Zhou, Y., & Chiu, D. (2014). Analytical QOE models for Bit-Rate switching in dynamic adaptive streaming systems. IEEE Transactions on Mobile Computing, 13(12), 2734–2748. https://doi.org/10.1109/tmc.2014.2307323
78. Yan, N. Y., Ricci, E., Liu, N. G., & Sebe, N. (2015). Egocentric daily activity recognition via multitask clustering. IEEE Transactions on Image Processing, 24(10), 2984–2995. https://doi.org/10.1109/tip.2015.2438540
79. Yew, J., & Shamma, D. A. (2011). Know your data: understanding implicit usage versus explicit action in video content classification. Proceedings of SPIE, the International Society for Optical Engineering/Proceedings of SPIE. https://doi.org/10.1117/12.878807
80. Zhang, D., Yu, Y., Dong, J., Li, C., Su, D., Chu, C., & Yu, D. (2024, January 24). MM-LLMS: Recent Advances in MultiModal Large Language Models. arXiv.org. https://arxiv.org/abs/2401.13601
81. Zhao, H., Cai, Z., Si, S., Ma, X., An, K., Chen, L., Liu, Z., Wang, S., Han, W., & Chang, B. (2023). MMICL: Empowering Vision-language Model with Multi-Modal In-Context Learning. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2309.07915
82. Zhao, S., Yao, H., & Sun, X. (2013). Video classification and recommendation based on affective analysis of viewers. Neurocomputing, 119, 101–110. https://doi.org/10.1016/j.neucom.2012.04.042
83. Zhou, C., Chai, C., & Liao, J. (2021). Analysis of problem decomposition strategies of novice industrial designers using network-based cognitive maps. International Journal of Technology and Design Education, 32(2), 1293–1315. https://doi.org/10.1007/s10798-020-09647-1
84. Zhou, C., Zhang, X., & Yu, C. (2023). How does AI promote design iteration? The optimal time to integrate AI into the design process. Journal of Engineering Design, 1–28. https://doi.org/10.1080/09544828.2023.2290915
85. Zhou, G., Hong, Y., & Wu, Q. (2023). NavGPT: Explicit Reasoning in Vision-and-Language Navigation with Large Language Models. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2305.16986
86. Zimmerman, J. (2005). Video sketches: exploring pervasive computing interaction designs. IEEE Pervasive Computing, 4(4), 91–94. https://doi.org/10.1109/mprv.2005.91

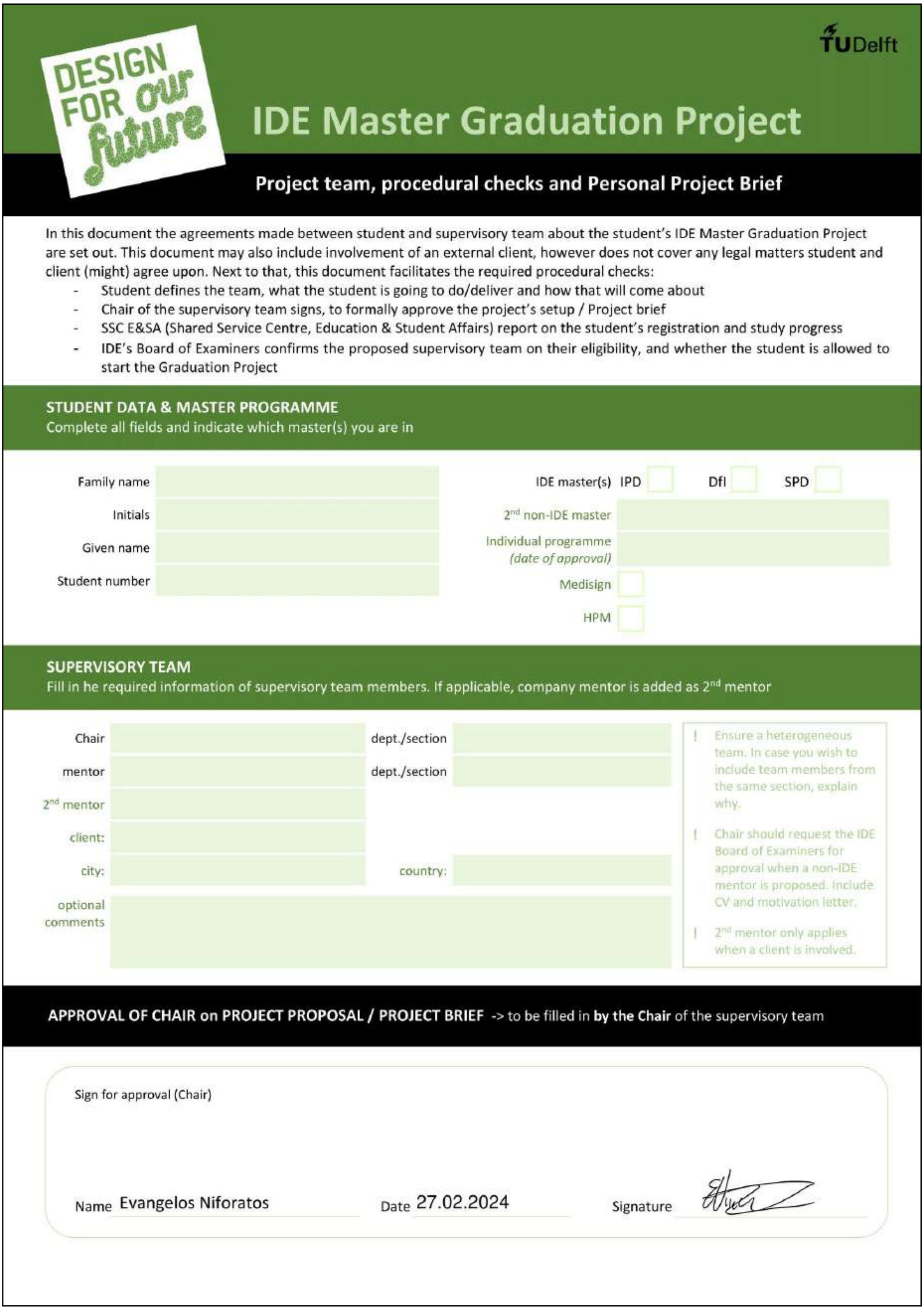
7. Appendix

A: Gantt chart of activities



7. Appendix

B: Project brief


DESIGN FOR our future
TUDelft

IDE Master Graduation Project

Project team, procedural checks and Personal Project Brief

In this document the agreements made between student and supervisory team about the student's IDE Master Graduation Project are set out. This document may also include involvement of an external client, however does not cover any legal matters student and client (might) agree upon. Next to that, this document facilitates the required procedural checks:

- Student defines the team, what the student is going to do/deliver and how that will come about
- Chair of the supervisory team signs, to formally approve the project's setup / Project brief
- SSC E&SA (Shared Service Centre, Education & Student Affairs) report on the student's registration and study progress
- IDE's Board of Examiners confirms the proposed supervisory team on their eligibility, and whether the student is allowed to start the Graduation Project

STUDENT DATA & MASTER PROGRAMME
Complete all fields and indicate which master(s) you are in

Family name

Initials

Given name

Student number

IDE master(s) IPD □ DfI □ SPD □

2nd non-IDE master

Individual programme *(date of approval)*

Medisign □

HPM □

SUPERVISORY TEAM
Fill in he required information of supervisory team members. If applicable, company mentor is added as 2nd mentor

Chair dept./section

mentor dept./section

2nd mentor

client:

city: country:

optional comments

! Ensure a heterogeneous team. In case you wish to include team members from the same section, explain why.

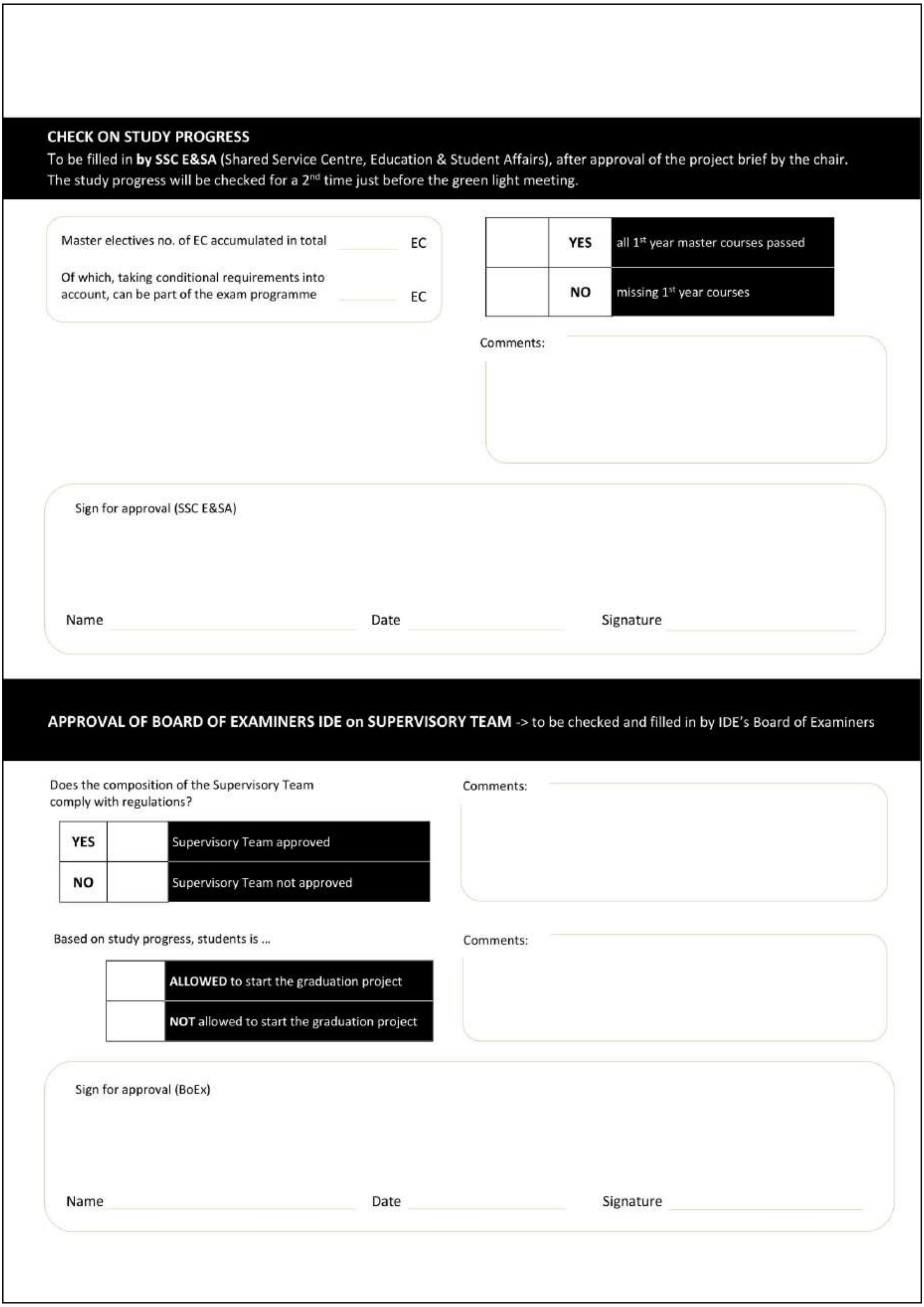
! Chair should request the IDE Board of Examiners for approval when a non-IDE mentor is proposed. Include CV and motivation letter.

! 2nd mentor only applies when a client is involved.

APPROVAL OF CHAIR on PROJECT PROPOSAL / PROJECT BRIEF -> to be filled in **by the Chair** of the supervisory team

Sign for approval (Chair)

Name Evangelos Niforatos Date 27.02.2024 Signature



CHECK ON STUDY PROGRESS

To be filled in **by SSC E&SA** (Shared Service Centre, Education & Student Affairs), after approval of the project brief by the chair. The study progress will be checked for a 2nd time just before the green light meeting.

Master electives no. of EC accumulated in total ________ EC

Of which, taking conditional requirements into account, can be part of the exam programme ________ EC

	YES	all 1st year master courses passed
	NO	missing 1st year courses

Comments:

Sign for approval (SSC E&SA)

Name ________ Date ________ Signature ________

APPROVAL OF BOARD OF EXAMINERS IDE on SUPERVISORY TEAM -> to be checked and filled in by IDE's Board of Examiners

Does the composition of the Supervisory Team comply with regulations?

YES		Supervisory Team approved
NO		Supervisory Team not approved

Comments:

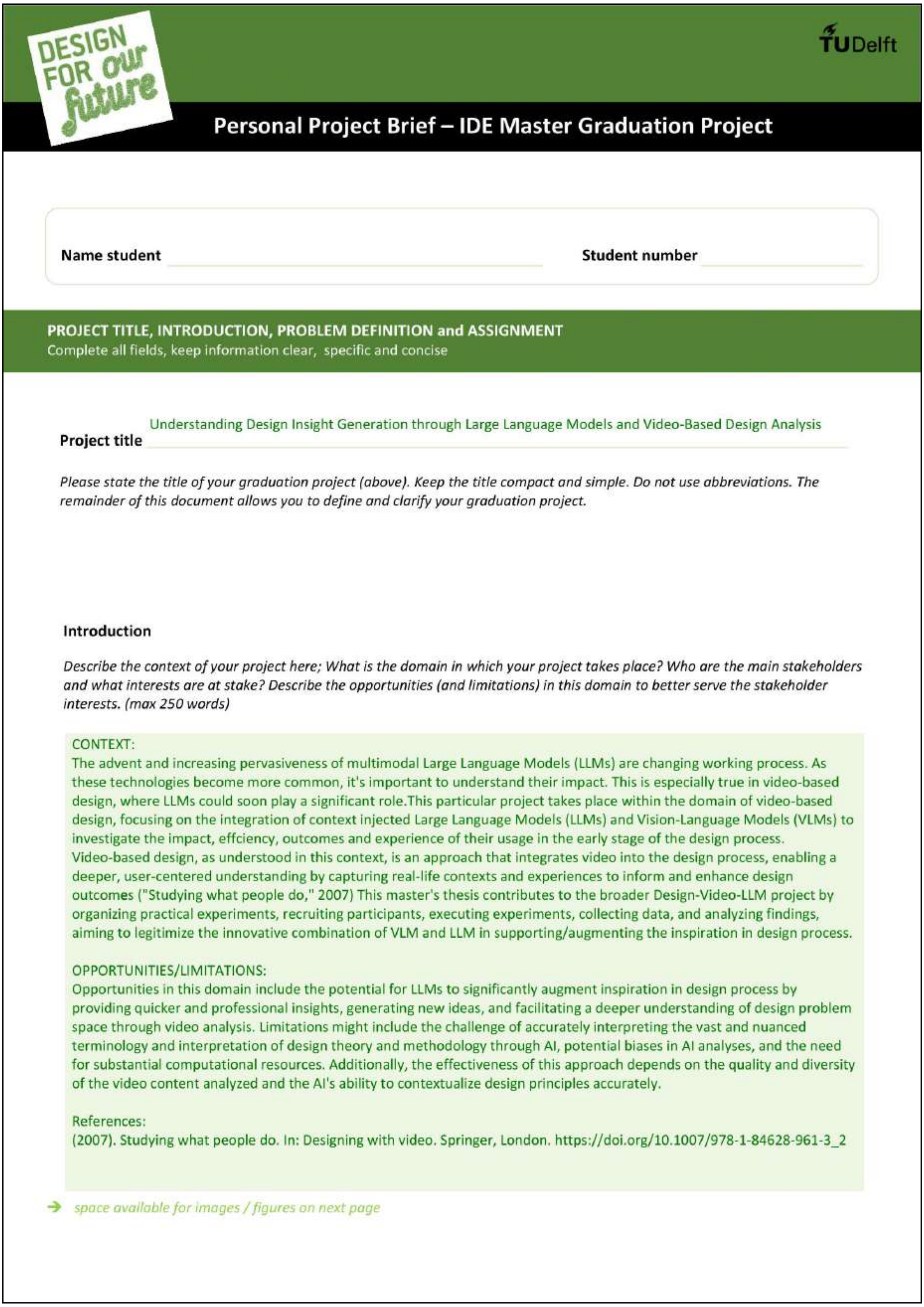
Based on study progress, students is ...

	ALLOWED to start the graduation project
	NOT allowed to start the graduation project

Comments:

Sign for approval (BoEx)

Name ________ Date ________ Signature ________



DESIGN FOR our future

TUDelft

Personal Project Brief – IDE Master Graduation Project

Name student ______________________ **Student number** ____________

PROJECT TITLE, INTRODUCTION, PROBLEM DEFINITION and ASSIGNMENT
Complete all fields, keep information clear, specific and concise

Project title Understanding Design Insight Generation through Large Language Models and Video-Based Design Analysis

Please state the title of your graduation project (above). Keep the title compact and simple. Do not use abbreviations. The remainder of this document allows you to define and clarify your graduation project.

Introduction

Describe the context of your project here; What is the domain in which your project takes place? Who are the main stakeholders and what interests are at stake? Describe the opportunities (and limitations) in this domain to better serve the stakeholder interests. (max 250 words)

CONTEXT:
The advent and increasing pervasiveness of multimodal Large Language Models (LLMs) are changing working process. As these technologies become more common, it's important to understand their impact. This is especially true in video-based design, where LLMs could soon play a significant role.This particular project takes place within the domain of video-based design, focusing on the integration of context injected Large Language Models (LLMs) and Vision-Language Models (VLMs) to investigate the impact, effciency, outcomes and experience of their usage in the early stage of the design process. Video-based design, as understood in this context, is an approach that integrates video into the design process, enabling a deeper, user-centered understanding by capturing real-life contexts and experiences to inform and enhance design outcomes ("Studying what people do," 2007) This master's thesis contributes to the broader Design-Video-LLM project by organizing practical experiments, recruiting participants, executing experiments, collecting data, and analyzing findings, aiming to legitimize the innovative combination of VLM and LLM in supporting/augmenting the inspiration in design process.

OPPORTUNITIES/LIMITATIONS:
Opportunities in this domain include the potential for LLMs to significantly augment inspiration in design process by providing quicker and professional insights, generating new ideas, and facilitating a deeper understanding of design problem space through video analysis. Limitations might include the challenge of accurately interpreting the vast and nuanced terminology and interpretation of design theory and methodology through AI, potential biases in AI analyses, and the need for substantial computational resources. Additionally, the effectiveness of this approach depends on the quality and diversity of the video content analyzed and the AI's ability to contextualize design principles accurately.

References:
(2007). Studying what people do. In: Designing with video. Springer, London. https://doi.org/10.1007/978-1-84628-961-3_2

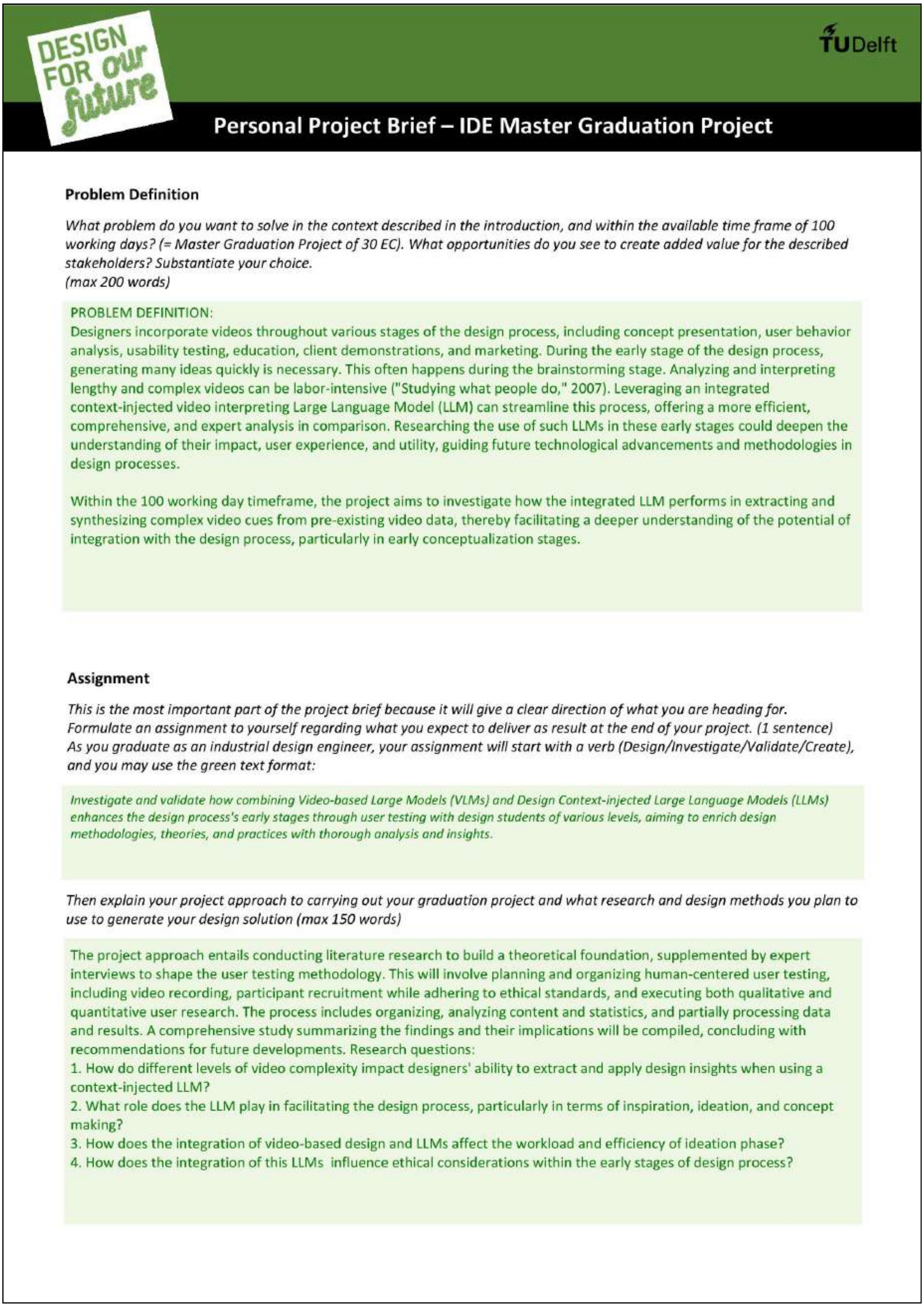
→ *space available for images / figures on next page*

introduction (continued): space for images

click to add picture

image / figure 1

image / figure 2



Personal Project Brief – IDE Master Graduation Project

Problem Definition

What problem do you want to solve in the context described in the introduction, and within the available time frame of 100 working days? (= Master Graduation Project of 30 EC). What opportunities do you see to create added value for the described stakeholders? Substantiate your choice.
(max 200 words)

PROBLEM DEFINITION:
Designers incorporate videos throughout various stages of the design process, including concept presentation, user behavior analysis, usability testing, education, client demonstrations, and marketing. During the early stage of the design process, generating many ideas quickly is necessary. This often happens during the brainstorming stage. Analyzing and interpreting lengthy and complex videos can be labor-intensive ("Studying what people do," 2007). Leveraging an integrated context-injected video interpreting Large Language Model (LLM) can streamline this process, offering a more efficient, comprehensive, and expert analysis in comparison. Researching the use of such LLMs in these early stages could deepen the understanding of their impact, user experience, and utility, guiding future technological advancements and methodologies in design processes.

Within the 100 working day timeframe, the project aims to investigate how the integrated LLM performs in extracting and synthesizing complex video cues from pre-existing video data, thereby facilitating a deeper understanding of the potential of integration with the design process, particularly in early conceptualization stages.

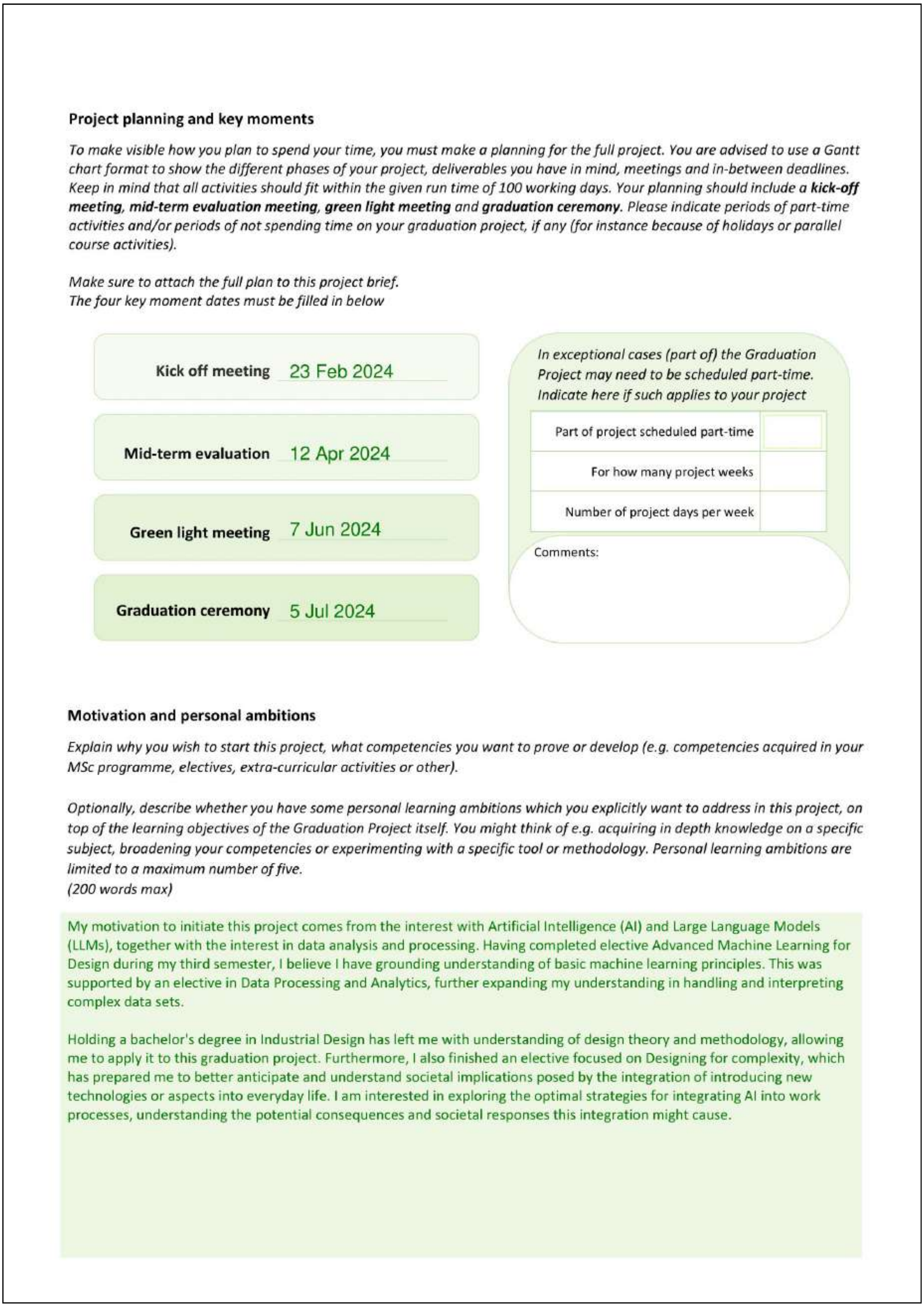
Assignment

This is the most important part of the project brief because it will give a clear direction of what you are heading for. Formulate an assignment to yourself regarding what you expect to deliver as result at the end of your project. (1 sentence) As you graduate as an industrial design engineer, your assignment will start with a verb (Design/Investigate/Validate/Create), and you may use the green text format:

Investigate and validate how combining Video-based Large Models (VLMs) and Design Context-injected Large Language Models (LLMs) enhances the design process's early stages through user testing with design students of various levels, aiming to enrich design methodologies, theories, and practices with thorough analysis and insights.

Then explain your project approach to carrying out your graduation project and what research and design methods you plan to use to generate your design solution (max 150 words)

The project approach entails conducting literature research to build a theoretical foundation, supplemented by expert interviews to shape the user testing methodology. This will involve planning and organizing human-centered user testing, including video recording, participant recruitment while adhering to ethical standards, and executing both qualitative and quantitative user research. The process includes organizing, analyzing content and statistics, and partially processing data and results. A comprehensive study summarizing the findings and their implications will be compiled, concluding with recommendations for future developments. Research questions:
1. How do different levels of video complexity impact designers' ability to extract and apply design insights when using a context-injected LLM?
2. What role does the LLM play in facilitating the design process, particularly in terms of inspiration, ideation, and concept making?
3. How does the integration of video-based design and LLMs affect the workload and efficiency of ideation phase?
4. How does the integration of this LLMs influence ethical considerations within the early stages of design process?



Project planning and key moments

*To make visible how you plan to spend your time, you must make a planning for the full project. You are advised to use a Gantt chart format to show the different phases of your project, deliverables you have in mind, meetings and in-between deadlines. Keep in mind that all activities should fit within the given run time of 100 working days. Your planning should include a **kick-off meeting**, **mid-term evaluation meeting**, **green light meeting** and **graduation ceremony**. Please indicate periods of part-time activities and/or periods of not spending time on your graduation project, if any (for instance because of holidays or parallel course activities).*

Make sure to attach the full plan to this project brief.
The four key moment dates must be filled in below

Kick off meeting	23 Feb 2024
Mid-term evaluation	12 Apr 2024
Green light meeting	7 Jun 2024
Graduation ceremony	5 Jul 2024

In exceptional cases (part of) the Graduation Project may need to be scheduled part-time. Indicate here if such applies to your project

Part of project scheduled part-time	
For how many project weeks	
Number of project days per week	

Comments:

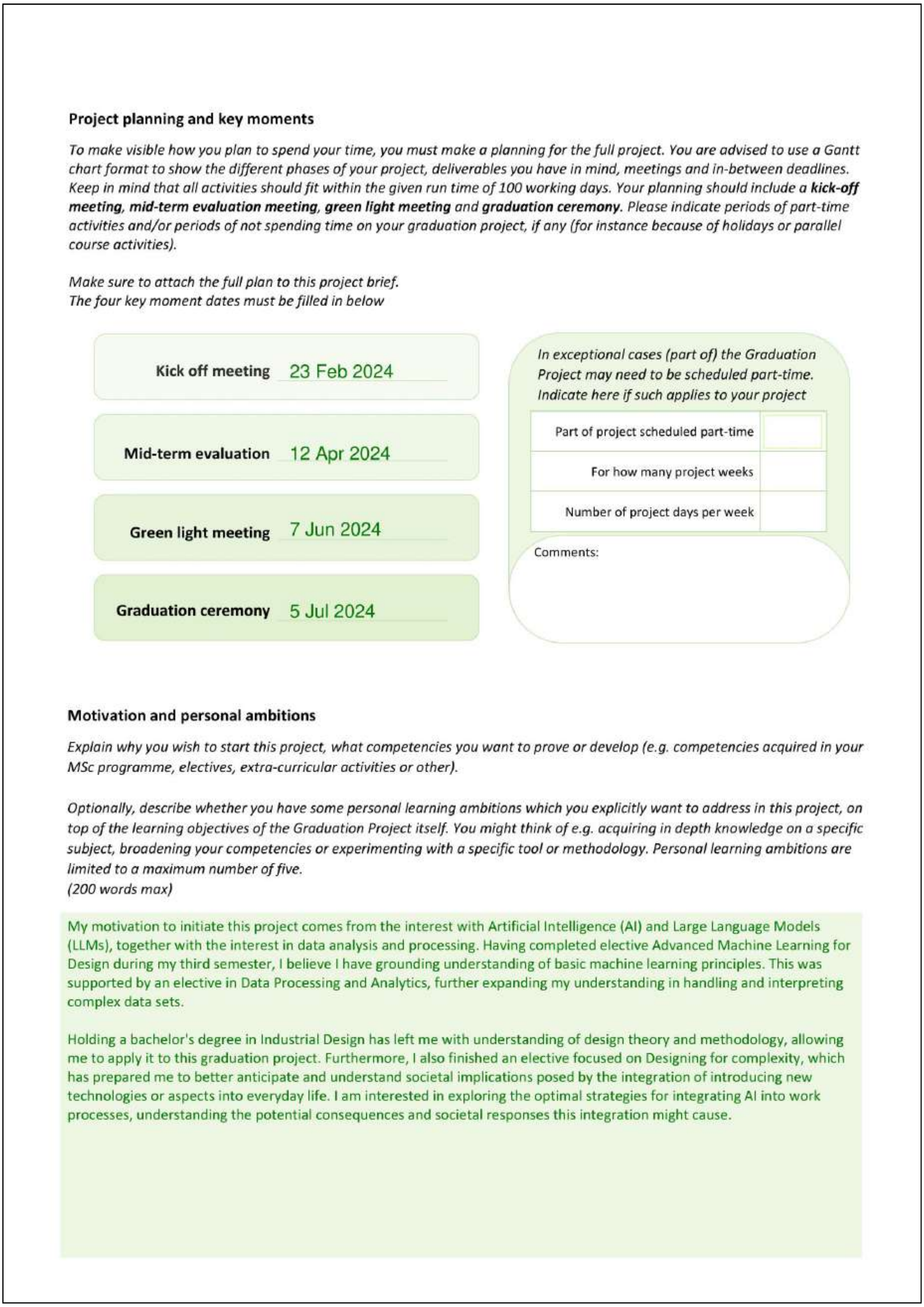
Motivation and personal ambitions

Explain why you wish to start this project, what competencies you want to prove or develop (e.g. competencies acquired in your MSc programme, electives, extra-curricular activities or other).

Optionally, describe whether you have some personal learning ambitions which you explicitly want to address in this project, on top of the learning objectives of the Graduation Project itself. You might think of e.g. acquiring in depth knowledge on a specific subject, broadening your competencies or experimenting with a specific tool or methodology. Personal learning ambitions are limited to a maximum number of five.
(200 words max)

My motivation to initiate this project comes from the interest with Artificial Intelligence (AI) and Large Language Models (LLMs), together with the interest in data analysis and processing. Having completed elective Advanced Machine Learning for Design during my third semester, I believe I have grounding understanding of basic machine learning principles. This was supported by an elective in Data Processing and Analytics, further expanding my understanding in handling and interpreting complex data sets.

Holding a bachelor's degree in Industrial Design has left me with understanding of design theory and methodology, allowing me to apply it to this graduation project. Furthermore, I also finished an elective focused on Designing for complexity, which has prepared me to better anticipate and understand societal implications posed by the integration of introducing new technologies or aspects into everyday life. I am interested in exploring the optimal strategies for integrating AI into work processes, understanding the potential consequences and societal responses this integration might cause.



Project planning and key moments

*To make visible how you plan to spend your time, you must make a planning for the full project. You are advised to use a Gantt chart format to show the different phases of your project, deliverables you have in mind, meetings and in-between deadlines. Keep in mind that all activities should fit within the given run time of 100 working days. Your planning should include a **kick-off meeting**, **mid-term evaluation meeting**, **green light meeting** and **graduation ceremony**. Please indicate periods of part-time activities and/or periods of not spending time on your graduation project, if any (for instance because of holidays or parallel course activities).*

Make sure to attach the full plan to this project brief.
The four key moment dates must be filled in below

Kick off meeting	23 Feb 2024
Mid-term evaluation	12 Apr 2024
Green light meeting	7 Jun 2024
Graduation ceremony	5 Jul 2024

In exceptional cases (part of) the Graduation Project may need to be scheduled part-time. Indicate here if such applies to your project

Part of project scheduled part-time	□
For how many project weeks	
Number of project days per week	

Comments:

Motivation and personal ambitions

Explain why you wish to start this project, what competencies you want to prove or develop (e.g. competencies acquired in your MSc programme, electives, extra-curricular activities or other).

Optionally, describe whether you have some personal learning ambitions which you explicitly want to address in this project, on top of the learning objectives of the Graduation Project itself. You might think of e.g. acquiring in depth knowledge on a specific subject, broadening your competencies or experimenting with a specific tool or methodology. Personal learning ambitions are limited to a maximum number of five.
(200 words max)

My motivation to initiate this project comes from the interest with Artificial Intelligence (AI) and Large Language Models (LLMs), together with the interest in data analysis and processing. Having completed elective Advanced Machine Learning for Design during my third semester, I believe I have grounding understanding of basic machine learning principles. This was supported by an elective in Data Processing and Analytics, further expanding my understanding in handling and interpreting complex data sets.

Holding a bachelor's degree in Industrial Design has left me with understanding of design theory and methodology, allowing me to apply it to this graduation project. Furthermore, I also finished an elective focused on Designing for complexity, which has prepared me to better anticipate and understand societal implications posed by the integration of introducing new technologies or aspects into everyday life. I am interested in exploring the optimal strategies for integrating AI into work processes, understanding the potential consequences and societal responses this integration might cause.

7. Appendix

C: Motivation

The motivation to initiate this project comes from the interest in Artificial Intelligence (AI), combined with the interest in data analysis and processing. Having completed the elective Advanced Machine Learning for Design during the third semester, I have a basic understanding of machine learning principles. This was supported by an elective in Data Processing and Analytics, further expanding my understanding in handling and interpreting complex data sets.

Holding a bachelor's degree in industrial design has left me with understanding of design theory and methodology, allowing me to apply it to this graduation project. Furthermore, I also finished an elective focused on Designing for complexity, which has prepared me to better anticipate and understand societal implications posed by the integration of introducing new technologies or aspects into everyday life. I am interested in exploring the optimal strategies for integrating AI into work processes, understanding the potential consequences and societal responses this integration might cause.

D: NASA TLX Python Code

```
# Define the column names
column_names = [
    "Effort:Performance", "Temporal Demand:Effort", "Performance:Frustration",
    "Physical Demand:Performance", "Temporal Demand:Frustration",
    "Physical Demand:Frustration", "Physical Demand:Temporal Demand",
    "Temporal Demand:Mental Demand", "Frustration:Effort",
    "Performance:Temporal Demand", "Mental Demand:Physical Demand",
    "Frustration:Mental Demand", "Performance:Mental Demand",
    "Mental Demand:Effort", "Effort:Physical Demand"
]

# Function to count terms based on input numbers
def count_terms(column_names, input_numbers):
    term_counts = {}
    for col_name, num in zip(column_names, input_numbers):
        terms = col_name.split(":")
        chosen_term = terms[num - 1]  # num - 1 to convert 1/2 to 0/1 index
        if chosen_term in term_counts:
            term_counts[chosen_term] += 1
        else:
            term_counts[chosen_term] = 1
    return term_counts

# Function to process multiple rows of input numbers
def process_multiple_rows(column_names, multiple_rows):
    all_counts = []
    for row_number, input_numbers in enumerate(multiple_rows, start=1):
        term_counts = count_terms(column_names, input_numbers)
        print(f"Counts for row {row_number}:")
        for term, count in term_counts.items():
            print(f"  {term}: {count}")
        all_counts.append(term_counts)
    return all_counts

# Example multiple rows of input numbers
multiple_rows = [
    [2, 2, 1, 2, 2, 2, 2, 2, 2, 1, 1, 2, 2, 1, 1],
    [1, 2, 2, 2, 2, 2, 1, 2, 1, 1, 1, 2, 1, 1, 1],
]

# Process the multiple rows and get counts for each row
all_counts = process_multiple_rows(column_names, multiple_rows)
```

Acknowledgements

I would like to express my gratitude to Evangelos Niforatos for his generous support and understanding throughout this thesis. His guidance has been invaluable.

I am also sincerely thankful to Tianhao He for his availability and readiness to help whenever needed. Your assistance has been greatly appreciated.

Special thanks to Matteo, Ujjayan, Lori, my friends, familiy and all the others for the times spent discussing everything and anything. Time spent socializing with all was incredibly helpful.

Lastly, I extend my thanks to all the participants in the study. Your patience in completing the seemingly endless questionnaires were esential to this research.

Note on use of AI

Artificial intelligence (GPT-4) was used to:

- structure some of the sentences and make the content more coherent
- search for literature
- explain concepts
- provide suggestions