Analyzing Interview Questions via Bloom's Taxonomy to Enhance the Design Thinking Process

Fatemeh Kazemi Vanhari and Christopher Anand and Charles Welch
Department of Computing and Software
McMaster University
Hamilton, Ontario, Canada
{kazemivf, anandc, cwelch}@mcmaster.ca

Abstract

Interviews are central to the Empathy phase of Design Thinking, helping designers uncover user needs and experience. Although interviews are widely used to support human-centered innovation, evaluating their quality, especially from a cognitive perspective, remains underexplored. This study introduces a structured framework for evaluating interview quality in the context of Design Thinking, using Bloom's Taxonomy as a foundation. We propose the Cognitive Interview Quality Score, a composite metric that integrates three dimensions: Effectiveness, Bloom Coverage, and Distribution Balance Score. Using human-annotations, we assessed 15 interviews across three domains to measure cognitive diversity and structure. We compared CIQS-based rankings with human experts and found that the Bloom Coverage Score aligned more closely with expert judgments. We evaluated the performance of LMA-3-8B-Instruct and GPT-4o-mini, using zero-shot, few-shot, and chain-of-thought prompting, finding GPT-4o-mini, especially in zero-shot mode, showed the highest correlation with human annotations in all domains. Error analysis revealed that models struggled more with mid-level cognitive tasks (e.g., Apply, Analyze) and performed better on Create, likely due to clearer linguistic cues. These findings highlight both the promise and limitations of using NLP models for automated cognitive classification and underscore the importance of combining cognitive metrics with qualitative insights to comprehensively assess interview quality.

1 Introduction

Design Thinking is a widely adopted framework for creative problem-solving, particularly in areas that require deep user understanding and human-centered innovation. It typically progresses through five iterative stages: Empathize, Define, Ideate, Prototype, and Test. At the heart of this process is the first stage, "Empathize", enabling designers to deeply understand users' experiences, emotions, and needs. It distinguishes Design Thinking from purely analytical approaches by emphasizing a human-centered perspective. This phase often involves interviews, observations, and immersive techniques, such as simulating real user experiences, to uncover pain points and inform meaningful design interventions (Brown, 2009; Org, 2015). Among these methods, interviews play a vital role by fostering open-ended, direct dialogue between researchers and users. The quality of the interview questions during this phase is especially critical, as it shapes the depth, clarity, and diversity of responses, and ultimately influences the effectiveness of the application's design.

Despite the central role of interviews, there is limited systematic guidance on how to structure interview questions to encourage deeper cognitive engagement. Many interviews rely on intuitive or ad hoc question writing, often leading to unbalanced questioning that skews toward lower-order thinking, such as remembering or understanding, while neglecting higher-order processes such as analyzing, evaluating, and creating (Anderson and Krathwohl, 2001).

To address this gap, we propose a novel approach that takes advantage of Bloom's Taxonomy, a widely used hierarchical framework that classifies cognitive tasks into six categories: Remember, Understand, Apply, Analyze, Evaluate, and Create (Bloom, 1956; Anderson and Krathwohl, 2001). Originally developed for educational settings, Bloom's Taxonomy has been effectively adapted in recent years for use in question generation (Hwang et al., 2023), question classification (Mohammed and Omar, 2018; Gani et al., 2023), and curriculum evaluation (and, 2002). In this study, we apply our approach to the domain of interview question design within the context of Design Thinking. Specifically, we investigate whether

Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications, pages 582–593

covering different levels of Bloom's Taxonomy in interview questions and responses has an effect on the overall quality of interviews and contributes to enhancing the Design Thinking process. Our analysis covers three interview subjects: AI Regulation, Math Visualizer, and Grandfather Game.

We use Large Language Models (LLMs) to automatically classify both interview questions and their responses according to Bloom's Taxonomy. Leveraging recent advances in prompt engineering techniques including zero-shot (Brown et al., 2020), few-shot (Liu et al., 2023), and chain-of-thought (CoT) prompting (Wei et al., 2022), we use LLaMA-3-8B-Instruct, an instruction-tuned open-source LLM, and GPT-4o-mini, a lightweight proprietary model from OpenAI optimized for fast reasoning tasks, to assign Bloom levels based on the cognitive demands of the text. We also introduce a composite evaluation metric, the Cognitive Interview Quality Score (CIQS), which integrates Bloom effectiveness, coverage, and distribution balance scores into a single measure to assess the overall quality of interview questions.

To guide our investigation into the cognitive quality of interviews and the role of automated classification, this study is driven by the following research questions:

RQ1: Does covering multiple levels of Bloom's Taxonomy in interview questions and responses contribute to higher-quality interviews within the Design Thinking process?

RQ2: Can LLMs, such as LLaMA-3-8B-Instruct and GPT-4o-mini, reliably classify interview content into Bloom's cognitive levels across different prompting strategies?

RQ3: To what extent do our CIQS-based automated rankings of interview quality align with expert human evaluations across diverse interview subjects?

2 Related Work

2.1 Automated Classification of Questions Using Bloom's Taxonomy

Bloom's Taxonomy, originally introduced by Bloom (1956) and later revised by Anderson and Krathwohl (2001), has long served as a framework for classifying learning objectives and designing educational assessments. Numerous studies have leveraged this taxonomy to guide the construction of questions that effectively target various cognitive levels, from simple recall (Remember) to complex creative tasks (Create). Chang and Chung (2009) developed a keyword-based system aimed at automatically classifying teachers' questions according to Bloom's Taxonomy. By constructing a dictionary that maps specific keywords to corresponding cognitive levels, their system achieved a 75% accuracy in identifying questions at the Remember level. However, its performance declined for higher-order levels, with accuracy ranging between 25% and 59%. Yahya and Osman (2011) explored the effectiveness of machine learning techniques by employing TF-IDF features combined with Support Vector Machine (SVM) classifiers to categorize 190 exam questions across Bloom's six cognitive categories. Haris and Omar (2012) employed a rule-based classifier to categorize 135 computer programming examination questions according to Bloom's Taxonomy.

Building upon these methodologies, Mohammed and Omar (2020) introduced an enhanced classification model incorporating TFPOS-IDF, a variation of TF-IDF that considers part-of-speech information, and pretrained word2vec embeddings to capture semantic relationships. They evaluated their model using kNN, Logistic Regression, and SVM classifiers on datasets containing 141 and 600 questions. The SVM classifier exhibited superior performance, achieving weighted F1-scores of 83.7% and 89.7% on the respective datasets, highlighting the efficacy of integrating syntactic and semantic features in question classification.

Li et al. (2022) conducted a study to automate the classification of learning objectives according to Bloom's Taxonomy. They compiled 21,380 learning objectives from 5,558 courses at an Australian university, manually labeled these objectives based on Bloom's six cognitive levels, and applied five conventional machine learning algorithms—Naive Bayes, Logistic Regression, Support Vector Machine, Random Forest, and XGBoost—as well as a deep learning approach using the pretrained BERT language model. Their findings demonstrated that BERT-based classifiers outperformed others across all cognitive levels, achieving Cohen's κ up to 0.93 and F1 scores up to 0.95. Additionally, SVM, Random Forest, and XGBoost models delivered performance comparable to BERT-based classifiers. The study also revealed that constructing separate binary classifiers for each cognitive level slightly outperformed a single multi-class, multi-label classifier, suggesting that individualized models for

each cognitive level may enhance classification accuracy.

Gani et al. (2023) focused on automating the classification of exam questions by evaluating various pretrained word embedding techniques, both contextual and non-contextual, across two datasets. Their study highlighted that while deep learning and contextual embeddings improved classification performance, their effectiveness was significantly influenced by dataset characteristics. Similarly, Al Faraby et al. (2024) assessed the capability of ChatGPT in classifying and generating questions. They found that in generating questions from reading sections, the differences with human-generated questions were not significant, indicating ChatGPT's potential for educational content creation.

2.2 Automatic Evaluation of Questions

Recent advancements in natural language processing have facilitated the automated evaluation of open-ended question complexity using Bloom's Taxonomy. Raz et al. (2024) employed a fine-tuned LLM to predict human ratings of question complexity, demonstrating a strong correlation (r = 0.73) between LLM-generated scores and human assessments, outperforming traditional baseline measures such as semantic distance and word count. Simone A Luchini and Beaty (2025) investigated the use of LLMs to assess the originality of narratives across multiple languages. They trained three distinct LLMs to predict human originality ratings of short stories written in 11 languages. The first model, trained exclusively on English narratives, achieved a robust correlation (r = 0.81) with human assessments. When this model was applied to multilingual stories translated into English, it maintained strong predictive performance ($r \geq 0.73$). Additionally, a multilingual model trained on narratives in their original languages reliably predicted human originality scores across all languages ($r \geq 0.72$). Hwang et al. (2023), explored an AI-driven approach to generating and evaluating multiple-choice questions in introductory chemistry and biology, focusing on alignment with Bloom's Taxonomy. They employed zero-shot prompting with GPT-3.5 to create questions, validated their cognitive levels using RoBERTa, and assessed question quality based on Item Writing Flaws Moore et al. (2023). The findings indicate that GPT-3.5 is capable of generating questions at various cognitive levels, particularly excelling at producing higher-order thinking questions at the Evaluation level. However, discrepancies between AI-generated and human-assessed Bloom levels suggest the need for further refinement in question generation methodologies. Additionally, the study highlights an inverse correlation between Bloom's level and perceived question quality, indicating that while AI can generate complex questions, it may struggle with nuances in cognitive distinction and clarity at higher taxonomic levels.

3 Methodology

3.1 Dataset

This study is based on a dataset of transcribed interviews collected to evaluate the cognitive depth of questions and responses used during the "Empathy" phase of the Design Thinking process. In this study, the interviews focused on three distinct subject areas: Grandfather Game Application, Math Visualizer Software, and AI Regulation (for a description of each area see Appendix B). These topics were selected to ensure a variety of user perspectives and cognitive demands, ranging from personal storytelling to educational technology and policy discussions.

A total of 15 semi-structured interviews were conducted. Each interview consisted of both high-level and low-level open-ended questions. Not all questions were equally well-structured, as the goal was to intentionally support a range of cognitive levels in line with Bloom's Taxonomy, enabling analysis across varying depths of reasoning and understanding. The interviews were audio-recorded with participant consent, transcribed using Microsoft Teams, and manually reviewed for accuracy. Transcripts were anonymized and structured by role (interviewer/interviewee).

While the original transcripts included more entries, we removed manually segments that were not suitable for cognitive classification. This included ice-breaker exchanges (for example, "Hi, how are you today?", "Thanks for joining us!"), affirmations (for example, "yes", "okay"), and expressions of appreciation (for example, "thank you"), all of which could not be meaningfully assigned a Bloom's level. After this filtering process, the final dataset consisted of 726 entries, comprising 363 interview questions and their corresponding 363 responses. All questions and responses were manually classified by one of the authors familiar with Bloom's Taxonomy levels. Our analysis spans the three interview subjects: AI Regulation

(274 entries), Math Visualizer (244 entries), and Grandfather Game (208 entries).

3.2 Bloom-Level Classification Process

To classify each interview question and response according to Bloom's Taxonomy, we employed a prompt-based strategy using two LLMs: LLaMA-3-8B-Instruct and GPT-4o-mini. We applied three prompting techniques including: zero-shot, few-shot, and CoT to guide the models' responses.

In the zero-shot prompting approach, the model receives a direct instruction to classify the input into one of the six Bloom levels including: Remember, Understand, Apply, Analyze, Evaluate, or Create, without being given any prior examples. This method tests the model's ability to rely on its internalized knowledge of Bloom's Taxonomy and produces a fast baseline classification.

In few-shot prompting, we provide the model with one labeled example for each Bloom's Taxonomy level before introducing the target input. These examples help calibrate the model's understanding of the classification task.

Finally, we apply CoT prompting, which instructs the model to explain its reasoning before presenting a final classification. This method encourages step-by-step cognitive processing, making the model's decision-making process transparent and auditable.

The purpose of this classifications is to evaluate their alignment with human judgment and to inform future efforts toward automating cognitive-level assessment in interviews (see Section 4.1 for the results).

3.3 Evaluation Framework

To assess the cognitive quality of interviews, we used human-annotated Bloom's Taxonomy classifications for each question and response. Based on these annotations, we calculated three key evaluation metrics: Effectiveness Score (ES), Bloom Coverage Score (BCS), and Distribution Balance Score (BDS), developed by the authors to capture different dimensions of cognitive engagement. Together, these metrics represent the Cognitive Interview Quality Score (CIQS), a composite measure reflecting the cognitive richness and structural diversity of each interview.

In this study, CIQS and its components were derived from human classifications due to their higher reliability. The following sections review the components of the CIQS metric.

3.3.1 Effective Score (ES)

The Effectiveness Score measures how well each interview question succeeds in eliciting the intended level of cognitive engagement, as defined by Bloom's Taxonomy. Rather than evaluating the question in isolation, this score is grounded in a comparison between the cognitive level of the question and the cognitive depth observed in the interviewee's response. This approach aligns with the goals of the "Empathy" phase in Design Thinking, where the primary objective is not only to ask meaningful questions but to generate equally meaningful insights Brown (2009).

To calculate ES, first each question–response pair is evaluated by comparing the intended cognitive level of the question with the actual level of the response, and rated according to this criteria:

- Highly Effective (2 points): The response exceeds the intended cognitive level (for example, a question aimed at "Analysis" receives a "Creative" response).
- Effective (1 point): The response matches the intended cognitive level of the question.
- Needs Improvement (0 points): The response falls below the intended level, indicating limited cognitive engagement.

After assigning these numerical values to each pair, the ES for each interview is calculated as the average score across all pairs:

$$\text{Effectiveness Score} = \frac{\sum_{i=1}^{n} s_i}{n} \tag{1}$$

where $s_i \in \{0, 1, 2\}$ is the score assigned to the i-th question–response pair based on the mentioned criteria, and n is the total number of pairs in the interview. The resulting score ranges from 0 (all questions need improvement) to 2 (all questions are highly effective). This metric captures not only the cognitive intent behind the questions but also their real-world impact as demonstrated through participant responses.

3.3.2 Bloom Coverage Score (BCS)

The Bloom Coverage Score evaluates the extent to which an interview engages participants across the six levels of Bloom's Taxonomy. A higher BCS indicates greater cognitive diversity, reflecting an intentional design that stimulates a broad range of thinking processes.

This diversity is particularly important in the context of Design Thinking, where complex problem-solving requires movement across multiple cognitive domains. Wu et al. (2021) propose a design thinking model explicitly structured around Bloom's Taxonomy, arguing that design thinking can be taught and structured through cognitive processes, from basic understanding to advanced creative generation. They emphasize that aligning design tasks with Bloom's full spectrum enables learners and participants to progress systematically from comprehension to innovation.

We define BCS as the number of cognitive levels covered in the interview divided by the total number of levels (6). The ideal BCS is 1.0, indicating that all six Bloom's levels are present at least once. The metric focuses on whether each level appears, not how often, encouraging diverse cognitive coverage in interview design.

3.3.3 Distribution Balance Score (BDS)

While the BCS measures the number of Bloom's cognitive levels represented in an interview, it does not reflect how evenly those levels are distributed. A cognitively rich interview is not only diverse in coverage but also balanced, ensuring that no single level dominates. To address this, we introduce BDS, which quantifies the uniformity of the cognitive distribution across Bloom's levels.

Let p_i represent the proportion of questions classified into the i-th Bloom's Taxonomy level, and let n be the total number of Bloom levels ($n = 6$). The BDS is defined as:

$$\text{BDS} = 1 - \frac{\sum_{i=1}^{n} \left(p_i - \frac{1}{n}\right)^2}{\frac{n-1}{n}} \quad (2)$$

This formula computes the squared deviation of the observed distribution $\{p_i\}$ from a uniform distribution $\frac{1}{n}$, and normalizes it by the maximum possible imbalance, which occurs when all items are concentrated in a single Bloom level. The squared term ensures that both over and underrepresentation contribute equally to the imbalance score, while penalizing larger deviations more. The BDS value ranges between 0 and 1.0. A BDS of 1.0 indicates a perfectly balanced distribution across all Bloom levels, reflecting equal representation. Conversely, a BDS of 0 signifies complete imbalance, with all items concentrated in a single Bloom level.

The formulation of the BDS is adapted from Pielou's Evenness Index Pielou (1966), traditionally used in ecology to assess distributional uniformity. We apply this concept to measure cognitive balance across Bloom's levels. Unlike entropy-based alternatives, our variance-based approach offers greater simplicity and sensitivity to cognitive imbalances. This metric encourages interviews that span multiple cognitive levels in a well-distributed and cognitively meaningful way.

3.3.4 Cognitive Interview Quality Score (CIQS)

To provide a comprehensive assessment of interview quality from a cognitive perspective, we propose the Cognitive Interview Quality Score. This metric combines three core dimensions: practical effectiveness, cognitive coverage, and structural balance. CIQS is calculated using the following weighted formula:

$$\text{CIQS} = 0.5 \times \text{ES} + 0.3 \times \text{BCS} + 0.2 \times \text{BDS} \quad (3)$$

In this formula, Effectiveness is emphasized most heavily to reflect the importance of empirical success: questions must not only be well-designed but must also stimulate the intended cognitive engagement, as evidenced by actual responses Anderson and Krathwohl (2001). Bloom Coverage receives moderate emphasis for its role in encouraging diverse thinking pathways, while Distribution Balance contributes structural integrity without dominating the evaluation. The weighting scheme (0.5 for ES, 0.3 for BCS, and 0.2 for BDS) was determined to prioritize cognitive alignment in actual responses while still valuing breadth and balance. This design is informed by principles from educational assessment and cognitive taxonomy theory Anderson and Krathwohl (2001), though the metric itself is introduced as part of this work. The CIQS serves as a unified cognitive quality rating for each interview, enabling systematic comparison across topics or participant groups while supporting iterative improvement in interview design.

3.4 Human Evaluation of Interview Quality

To validate the CIQS framework, we conducted a human evaluation in which an expert (tenured Professor) in design thinking independently ranked the interviews across all three subjects. The expert ranked each interview based on its effectiveness in uncovering useful information about the user and their practices and needs. This qualitative judgment served as a benchmark to assess how well CIQS scores aligned with human-perceived

interview quality. Comparing the CIQS rankings with the expert's rankings helps determine whether cognitively focused metrics reflect what a human evaluator considers a high-quality, informative interview.

4 Experiments & Results

4.1 Evaluating Human–LLM Cognitive Classification Agreement

One of the authors annotated all question-response pairs in our dataset for their Bloom's Taxonomy level. To measure the agreement between LLM-assigned and human-assigned Bloom's Taxonomy levels, we used Kendall's Tau (τ), which is well-suited for ordinal data and provides a robust estimate of correlation, particularly with small sample sizes and tied ranks (Kendall, 1938). The results are presented in Table 1, indicate that the zero-shot GPT-4o-mini achieved the strongest alignment with human judgments in all domains: AI Regulation ($\tau = 0.58$), Math Visualizer ($\tau = 0.47$), and Grandfather Game ($\tau = 0.56$). Among LLaMA-3-8B-Instruct models, the few-shot prompting yielded the highest correlations overall, particularly in AI Regulation ($\tau = 0.33$). In contrast, zero-shot prompting under LLaMA showed very weak agreement across subjects.

These findings suggest that GPT-4o-mini, especially in zero-shot, is more reliable for capturing cognitive-level distinctions in interview data, while open-source LLaMA models show more limited alignment with expert assessments. Performance differences can be attributed to the models' architectures and training methodologies. GPT-4o-mini (OpenAI's distilled model) balances efficiency and advanced reasoning, excelling in nuanced tasks.[1] LLaMA-3-8B-Instruct, while optimized for dialogue and instruction-following, may require further fine-tuning to match the classification accuracy demonstrated by GPT-4o-mini in this study.[2]

To identify which Bloom's levels posed the greatest challenges for LLMs, we generated separate confusion matrices comparing the aggregated predictions of LLaMA-3-8B-Instruct models and GPT-4o-mini models against human classifications across Bloom's Taxonomy levels, as presented in Figures 1 and 2. The LLaMA ensemble, based on

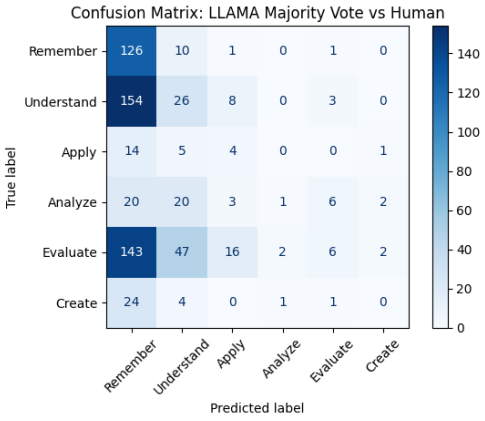

Figure 1: Confusion Matrix: LLaMA-3-8B-Instruct Majority Vote Vs Human Classification.

majority voting, exhibited a strong overprediction of the "Remember" category, leading to widespread misclassification of responses originally labeled as "Understand", "Evaluate", and "Create". This pattern suggests a tendency to default to lower-order cognitive categories. In contrast, the GPT-4o-mini ensemble produced a more balanced distribution across predicted classes, with higher accuracy in identifying"Remember", "Understand" and "Evaluate", and notably less confusion between the levels.

These findings are further supported by the quantitative results reported in Tables 2 and 3. The LLaMA-3-8B-Instruct models showed limited alignment with human labels, with accuracy ranging from 23.9% to 29.1% and macro F1-scores below 0.19. Their highest macro precision and recall were 0.312 and 0.226, respectively, under the Chain-of-Thought setting. In contrast, all GPT-4o-mini variants outperformed LLaMA across metrics. The Zero-shot GPT model achieved 53.7% accuracy and a macro F1-score of 0.511, while Few-shot prompting reached a macro precision of 0.642. GPT models also showed stronger weighted F1-scores, indicating better overall balance across Bloom levels.

4.2 Evaluating Cognitive Dimensions of Interviews with CIQS

To evaluate and compare the cognitive quality of interviews across different topics, we applied our proposed scoring framework, the Cognitive Interview Quality Score, which combines three key dimensions: Effectiveness Score, Bloom Coverage Score, and Distribution Balance Score. As illustrated in Table 4, AI Regulation achieved the highest CIQS (0.88), supported by strong effectiveness (ES =

[1] https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence

[2] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

Model and Prompting Technique	AI Regulation	Math Visualizer	Grandfather Game
LLaMA-3-8B-Instruct Zero-shot	0.08	-0.01	0.01
LLaMA-3-8B-Instruct Few-shot	0.33	0.26	0.26
LLaMA-3-8B-Instruct Chain-of-Thought	0.18	0.09	0.33
GPT-4o-mini Zero-shot	**0.58**	**0.47**	**0.56**
GPT-4o-mini Few-shot	0.45	0.41	0.51
GPT-4o-mini Chain-of-Thought	0.52	0.41	0.47

Table 1: Kendall's Tau (τ) correlation coefficients between model predictions and human annotations with highest scoring models in bold.

Model	Accuracy	Macro Precision	Macro Recall	Macro F1	Weighted F1
LLaMA-3-8B-Instruct Zero-shot	0.239	0.218	0.171	0.129	0.180
LLaMA-3-8B-Instruct Few-shot	0.285	0.222	0.225	0.155	0.205
LLaMA-3-8B-Instruct Chain-of-Thought	**0.291**	**0.312**	**0.226**	**0.185**	**0.230**

Table 2: Performance of LLaMA-3-8B-Instruct models across prompting techniques.

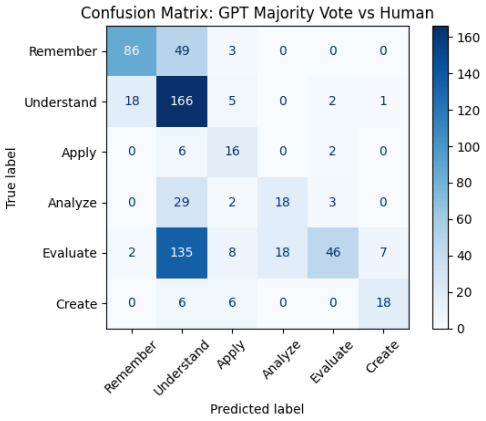

Figure 2: Confusion Matrix: GPT-4o-mini Majority Vote vs Human Classification.

1.01) and distribution balance (BDS = 0.80), despite slightly lower Bloom coverage (BCS = 0.71). This suggests that responses in AI-related interviews were well-aligned with the intended cognitive levels and well-distributed, though not all Bloom levels were equally represented. In contrast, Math Visualizer interviews exhibited the lowest CIQS (0.82), mainly due to a lower effectiveness score, suggesting that responses did not consistently reach the cognitive depth expected from the questions. Grandfather Game fell in the middle CIQS (0.84), showing relatively strong alignment but narrower cognitive coverage.

This automated scoring approach enables an objective comparison of interviews based on cognitive dimensions. However, cognitive depth is only one aspect of interview quality. As part of future work, we aim to explore additional metrics, such as emotional engagement, relevance to interview goals, procedural coverage, and question neutrality. These dimensions emerged from the feedback we received during our interview sessions on different topics, where participants highlighted aspects that contributed to more meaningful and engaging conversations. These dimensions may offer a more complete view of interview quality beyond what Bloom's taxonomy captures.

4.3 Evaluating the Alignment Between CIQS and Human Rankings

Figures 3-5 compare CIQS-based rankings with human expert judgments. Each CIQS score reflects a weighted combination of ES, BCS, and BDS.

In AI Regulation, the expert ranked Interview 3 as the most effective and Interview 1 as the least, while our CIQS-based scoring produced the opposite order and ranked Interview 3 as the most effective, highlighting a misalignment between cognitive structure (as captured by CIQS) and the expert's judgment, which was based on how well each interview uncovered useful information about the user and their practices and needs. For Math Visualizer, Interview 5 ranked highest by CIQS due to perfect BCS and strong ES, while the expert preferred Interview 2 for its insightfulness. In Grandfather Game, both approaches aligned on Interview 1 as the best, though discrepancies appeared in the middle ranks. Notably, further analysis revealed that Bloom Coverage Score more closely aligned with human expert rankings than CIQS or other individual metrics. BCS showed moderate to strong correlations with expert judgments across all domains: ($\rho = 0.50$) in Math Visualizer, ($\rho = 0.90$) in Grandfather Game, and ($\rho = 0.71$) in AI Regulation. These results suggest that interviews with broader cognitive coverage were more likely to be perceived as informative and high-quality by experts, contradicting our initial hypothesis that

Model	Accuracy	Macro Precision	Macro Recall	Macro F1	Weighted F1
GPT-4o-mini Zero-shot	**0.537**	0.584	**0.537**	**0.511**	0.521
GPT-4o-mini Few-shot	0.491	**0.642**	0.481	0.477	0.453
GPT-4o-mini Chain-of-Thought	0.518	0.529	0.527	0.494	**0.518**

Table 3: Performance of GPT-4o-mini models across prompting techniques.

Metric	AI Regulation	Math Visualizer	Grandfather Game
Effectiveness Score	**1.01**	0.84	1.00
Bloom Coverage Score	0.71	**0.80**	0.64
Distribution Balance Score	0.80	**0.82**	0.75
Cognitive Interview Quality Score	**0.88**	0.82	0.84

Table 4: Cognitive evaluation scores across interview subjects with highest scores in bold.

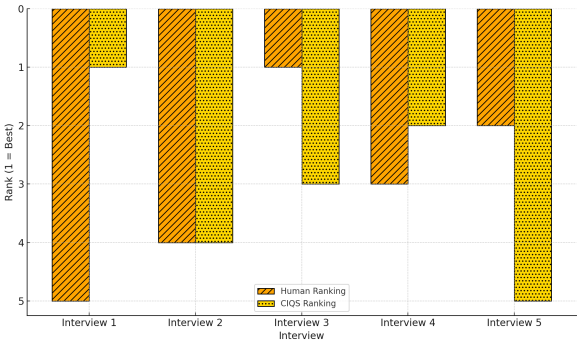

Figure 3: AI Regulation: CIQS vs Human Rankings.

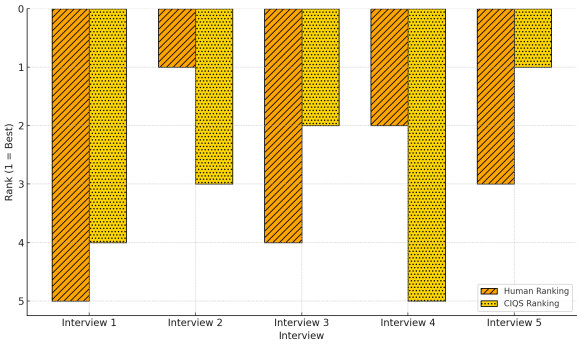

Figure 4: Math Visualizer: CIQS vs Human Rankings.

Effectiveness Score would play the most influential role in overall evaluation. To further investigate this we performed a linear regression to learn the coefficients for Equation 3 that best align with the human expert rankings. We found that BCS had the highest coefficient but that values varied across domains with ES and BDS less consistently in their impact. While more work is needed to determine which factors most correlate with human judgments, these preliminary results suggest that BCS is more impactful and that other attributes of the topic may be relevant in expert decisions (for full regression details, see Appendix C).

The results suggest that While CIQS captures

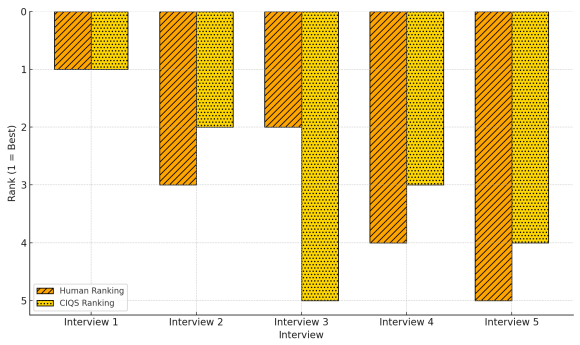

Figure 5: Grandfather Game: CIQS vs Human Rankings.

the cognitive structure of interviews, human evaluations often consider additional factors such as relevance, clarity, emotional engagement, and procedural detail. This highlights the value of combining cognitive metrics with qualitative insights for a more complete assessment of interview quality.

5 Discussion

RQ1: Our results suggest that interviews covering a broader range of Bloom's cognitive levels (higher BCS) tend to be ranked more favorably by the human expert, indicating greater cognitive diversity. This supports the hypothesis that cognitive richness, particularly through varied questioning strategies, enhances the quality of interviews in the Design Thinking context. However, alignment with human expert rankings was not always consistent, implying that additional qualitative dimensions (for example, emotional engagement and the inclusion of procedural information) also influence perceived interview quality.

RQ2: The outputs from LLaMA and GPT-4o-mini demonstrated partial alignment with human annotations, showing that LLMs have the potential to support cognitive level classification. GPT-

4o-mini, in particular, showed stronger agreement with human labels across prompting strategies, especially in zero-shot settings. However, inconsistencies across domains and between models reveal that current LLMs are not yet fully reliable as standalone evaluators. While their performance is promising for future automation efforts, fine-tuning and prompt engineering may be necessary to achieve consistent, human-comparable accuracy.

RQ3: The CIQS rankings showed partial alignment with expert human evaluations, with higher consistency in the Grandfather Game domain (τ = 0.40) and greater divergence in Math Visualizer (τ = –0.8) domain. These differences suggest that while CIQS effectively captures the cognitive structure and balance of interviews, human experts often consider additional qualitative dimensions, such as emotional engagement, relevance to user needs, and the inclusion of procedural information, that are not directly encoded in cognitive metrics. As such, CIQS serves as a valuable and scalable starting point for evaluating interview quality, but it should complement qualitative assessments.

6 Conclusion & Future Work

This study introduced a cognitive evaluation framework for interview quality based on Bloom's Taxonomy, applied within the context of Design Thinking. We proposed the CIQS, a composite metric incorporating effectiveness, coverage, and distribution of cognitive levels. Using human-annotations, we collected and evaluated 15 interviews across three domains to measure the cognitive diversity and structure of interview content. We compared CIQS rankings with expert judgments, finding that while they are partially aligned, BCS correlates more strongly with human rankings than CIQS or other individual metrics, suggesting that breadth is especially valued by experts. GPT-4o-mini, particularly in zero-shot, showed the highest agreement with human Bloom level annotations (up to τ = 0.58), outperforming LLaMA-3-8B-Instruct.

These findings suggest that while CIQS effectively captures the cognitive structure of interviews, human evaluations often prioritize additional factors such as relevance to user needs, clarity, emotional engagement, and procedural depth. This highlights the importance of complementing cognitive metrics with broader qualitative dimensions for a more comprehensive assessment of interview quality. In future work, we plan to refine CIQS by exploring alternative weighting, incorporating additional qualitative indicators, and fine-tuning LLMs for more accurate, autonomous classification of interview content based on Bloom's Taxonomy. To support continued research, we will release our corpus of 726 question–response pairs spanning three domains to support future work.

Limitations

The main challenge of this and any study of Design Thinking effectiveness is the maxim "savour surprises", by which design thinkers mean that the most important information is usually the information which was not anticipated and not planned for. This is because this information is the most likely to invalidate a design made without in-depth user interviews, or to lead to a new product category which was not previously contemplated Furr and Dyer (2014). At this stage, we are not trying to identify such surprises, but ultimately, a research program aiming to improve design education will have to address it.

A more immediate limitation of this study is the use of a single human expert. Experts in teaching and evaluating design thinking are uncommon and in demand in academia and industry. To increase the number of evaluators, it will be necessary to streamline the process so that it is less time-consuming.

Another limitation of this study is that even the human evaluator is not evaluating what we ultimately care about: the acceleration of the innovation process through better design interviews. We do not know whether interviews ranked highly by human experts actually lead to higher rates of innovation. Once automated metrics are found with higher levels of agreement with human experts, validation studies including the full development cycle from initial interviews to product validation will be necessary.

Role-playing can be challenging depending on the task. For our AI interviews, we noticed a lack of procedural information and emotion, where we expected more of both. We think it is not trivial for most people to role-play older individuals or versions of themselves. We suggest future work in this direction borrows from more established fields to set up experiments involving perspective-taking, e.g. work on empathy Batson et al. (2002).

Finally, since our intermediate goal is to produce tools useful for teaching design skills, it is disap-

pointing that the proprietary LLM greatly outperformed the open-source LLM. Many school boards and higher education institutions will be reluctant to submit their students' data to proprietary LLMs which they cannot control. In our study, we used role-playing by sophisticated professionals, graduate students and upper-year undergraduate students to produce a data set for training and evaluation. In teaching scenarios, it would be much harder to insure that personal information would not lead into the interviews. Moreover, when you initially describe the data set, you need to use similar language to say that this data is designed to not include personal information. Finally, a data set generated using role-playing may be fundamentally different from real design interviews in a way which effects the validity of the metrics.

While Bloom's Taxonomy provides a useful scaffold for assessing cognitive engagement, it has known limitations. The taxonomy does not explicitly model underlying mental processes such as perception, memory, and intuition, and some categories may overlap in practice—for example, extrapolation under "Understand" often resembles "Apply." Furthermore, the hierarchy between "Apply", "Analyze", and "Create" has been critiqued as insufficiently nuanced. Future extensions could explore integrating more adaptive taxonomies that better capture the fluid and context-dependent nature of reasoning in design interviews (Madaus et al., 1973).

References

Said Al Faraby, Ade Romadhony, and Adiwijaya. 2024. Analysis of LLMs for educational question classification and generation. *Computers and Education: Artificial Intelligence*, 7:100298.

David R. Krathwohl and. 2002. A Revision of Bloom's Taxonomy: An Overview. *Theory Into Practice*, 41(4):212–218.

Lorin W Anderson and David R Krathwohl. 2001. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives: complete edition*. Addison Wesley Longman, Inc.

C Daniel Batson, Nadia Ahmad, David A Lishner, J Tsang, CR Snyder, and SJ Lopez. 2002. Empathy and altruism. *The Oxford handbook of hypo-egoic phenomena*, pages 161–174.

Benjamin S. Bloom. 1956. Taxonomy of educational objectives: The classification of educational goals. *Handbook; Cognitive domain*, 1.

Tim Brown. 2009. *Change by Design: How Design Thinking Creates New Alternatives for Business and Society*. Harper Business, New York, NY.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Wen-Chih Chang and Ming-Shun Chung. 2009. Automatic applying Bloom's Taxonomy to classify and analysis the cognition level of English question items. In *2009 Joint Conferences on Pervasive Computing (JCPC)*, pages 727–734. IEEE.

Nathan R Furr and Jeff Dyer. 2014. *The innovator's method: bringing the lean start-up into your organization*. Harvard Business Press.

Mohammed Osman Gani, Ramesh Kumar Ayyasamy, Anbuselvan Sangodiah, and Yong Tien Fui. 2023. Bloom's Taxonomy-based exam question classification: The outcome of CNN and optimal pre-trained word embedding technique. *Education and Information Technologies*, 28(12):15893–15914.

Syahidah Sufi Haris and Nazlia Omar. 2012. A rule-based approach in Bloom's Taxonomy question classification through natural language processing. *2012 7th International Conference on Computing and Convergence Technology (ICCCT)*, pages 410–414.

Kevin Hwang, Sai Challagundla, Maryam Alomair, Lujie Karen Chen, and Fow-Sen Choa. 2023. Towards AI-assisted multiple choice question generation and quality evaluation at scale: Aligning with Bloom's Taxonomy. In *Workshop on Generative AI for Education*.

M. G. Kendall. 1938. A New Measure of Rank Correlation. *Biometrika*, 30(1-2):81–93.

Yuheng Li, Mladen Rakovic, Boon Xin Poh, Dragan Gasevic, and Guanliang Chen. 2022. Automatic Classification of Learning Objectives Based on Bloom's Taxonomy. In *Proceedings of the 15th International Conference on Educational Data Mining*, pages 530–537, Durham, United Kingdom. International Educational Data Mining Society.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM computing surveys*, 55(9):1–35.

George F. Madaus, Elinor M. Woods, and Ronald L. Nuttall. 1973. A causal model analysis of bloom's taxonomy. *American Educational Research Journal*, 10(4):253–262.

Manal Mohammed and Nazlia Omar. 2018. Question classification based on Bloom's Taxonomy using enhanced TF-IDF. *International Journal on Advanced Science, Engineering and Information Technology*, 8(4-2):1679.

Manal Mohammed and Nazlia Omar. 2020. Question classification based on Bloom's Taxonomy cognitive domain using modified TF-IDF and word2vec. *PloS one*, 15(3):e0230442.

Steven Moore, Huy A Nguyen, Tianying Chen, and John Stamper. 2023. Assessing the quality of multiple-choice questions using GPT-4 and rule-based methods. In *European conference on technology enhanced learning*, pages 229–245. Springer.

IDEO Org. 2015. *The field guide to human centered design*. Ideo Org.

E.C. Pielou. 1966. The measurement of diversity in different types of biological collections. *Journal of Theoretical Biology*, 13:131–144.

Tuval Raz, Simone Luchini, Roger Beaty, and Yoed Kenett. 2024. Automated Scoring of Open-Ended Question Complexity: A Large Language Model Approach. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 46.

John D. Patterson Dan Richard Johnson Matthijs Baas Baptiste Barbot Iana P. Bashmakova Mathias Benedek Qunlin Chen Giovanni Emanuele Corazza Boris Forthmann Benjamin Goecke Sameh Said Ibrahim Maciej Karwowski Yoed Kenett Izabela Lebuda Todd Lubart Kirill G. Miroshnik Felix Kingsley Obialo Marcela Ovando-Tellez Ricardo Primi Rogelio Puente Diaz Claire Stevenson Emmanuelle Volle Aleksandra Zielińska Janet van Hell Yin Wenpeng Simone A Luchini, Moosa Ibraheem Muhammad and Roger Beaty. 2025. Automated assessment of creativity in multilingual narratives. *Psychology of Aesthetics, Creativity, and the Arts*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Fan Wu, Yang Cheng Lin, and Peng Lu. 2021. A new design thinking model based on Bloom's Taxonomy. *Learn X Design 2021: Engaging with challenges in design education*.

Anwar Ali Yahya and Addin Osman. 2011. Automatic classification of questions into Bloom's cognitive levels using support vector machines. *Computer Science, Education*.

7 Appendix

A Prompts Used for LLM Classification

This appendix provides the exact prompts used to classify interview questions and responses into cognitive levels according to Revised Bloom's Taxonomy, using different prompting techniques.

Zero-Shot Prompt

You are a cognitive science expert that categorizes text into one of the Revised Bloom's Taxonomy levels.
You must respond with only one word: one of the following levels: Remember, Understand, Apply, Analyze, Evaluate, or Create.
Do not provide any explanation, reasoning, or additional text. Only return the level name in the following format.
Classification: <One of the six Bloom's levels>

Few-Shot Prompt

You are a cognitive science expert trained in Revised Bloom's Taxonomy.
Classify the following text according to Revised Bloom's Taxonomy levels: Remember, Understand, Apply, Analyze, Evaluate, or Create.
Examples:
Text: "List the main components of design thinking."
Classification: Remember
Text: "Explain the theory of cognitive load."
Classification: Understand
Text: "How would you apply Pythagoras' theorem to calculate the height of a building?"
Classification: Apply
Text: "Identify patterns in customer behavior based on the provided dataset."
Classification: Analyze
Text: "Evaluate the effectiveness of renewable energy sources compared to fossil fuels."
Classification: Evaluate
Text: "Design a new marketing strategy for launching a product."
Classification: Create
Do not provide any explanation, reasoning, or additional text. Only return the level name in the following format.
Classification:<One of the six Bloom's levels>

Chain-of-Thought Prompt

You are a cognitive science expert in Revised Bloom's Taxonomy.
Your task is to classify a given text into one of the Revised Bloom's Taxonomy cognitive levels:
Remember, Understand, Apply, Analyze, Evaluate, or Create.
Text: {input-text}

First, explain your reasoning step by step based on what the text requires cognitively.
Then, based on your explanation, select the most appropriate Bloom's level from (Remember, Understand, Apply, Analyze, Evaluate, Create) using the following format:
Classification: <One of the six Bloom's levels>
Note: For all classifications, the following model parameters were used:
Temperature = 0.0, Max tokens = 500.

B Description of Interview Topics

This study includes interviews conducted on three different design topics, each selected to represent different cognitive and contextual demands. The topics were used to simulate early-stage Design Thinking sessions and assess the cognitive quality of interview interactions.The participants in this study included a mix of students or recent graduates and university professors, some of whom had prior experience with Design Thinking. To maintain anonymity, they were instructed to avoid disclosing any real personal information. Depending on the interview topic, participants were asked to adopt specific roles. In the Math Visualizer interviews, they were asked to act as university students; in the Grandfather Game interviews, they assumed the perspective of older adults; and in the AI Regulation interviews, they portrayed individuals using AI platforms in organizations such as schools or businesses. Interviewers were instructed to engage naturally while focusing on uncovering user needs and generating meaningful insights.

B.1 AI Regulation

This topic explores public perceptions, concerns, and expectations surrounding the regulation of artificial intelligence. The interviewees were asked about their understanding of AI technologies, trust in regulatory frameworks, and suggestions for ethical oversight. The domain encourages abstract reasoning and evaluative thinking about policy and technology.

B.2 Math Visualizer

This topic focuses on the use of visualization tools in learning mathematics. Participants discussed their personal experiences with visual learning, the challenges they face in understanding mathematical concepts, and ideas for improving visual interfaces.

B.3 Grandfather Game

This topic centers on designing a game that would appeal to older adults. Participants were asked to reflect on their childhood memories, personal interests, and previous gaming experiences to inform the creation of engaging and age-appropriate game concepts.

C Linear Regression for CIQS

We can predict the human rankings of design thinking interviews with the CIQS score by learning coefficients for Equation 3. We predict the coefficients with intercept for each conversation topic. The equation for the Grandfather Game topic is shown in Equation 4 and yields $R^2 = 0.51$ which matches human ranking. Similarly, for AI-Regulation, we get $R^2 = 0.99$ for Equation 5. Lastly, for the Math Visualiser we get Equation 6 with $R^2 = 0.58$. Here we are predicting the rank (lower is better) with the same terms, which is different than ranking by the maximum score as we did in the main part of the paper, however, it serves the same function and supports our claims that BCS is the most important term and the impact of factors appears to vary with the domain.

$$\begin{aligned}\text{CIQS} = &-11.0 \times \text{ES} - 25.0 \times \text{BCS} \\ &+ 9.2 \times \text{BDS} + 22.9 \end{aligned} \tag{4}$$

$$\begin{aligned}\text{CIQS} = &-13.6 \times \text{ES} - 3.6 \times \text{BCS} \\ &+ 53.5 \times \text{BDS} - 23.1 \end{aligned} \tag{5}$$

$$\begin{aligned}\text{CIQS} = &\ 21.0 \times \text{ES} - 18.8 \times \text{BCS} \\ &- 13.0 \times \text{BDS} + 11.1 \end{aligned} \tag{6}$$