Beyond Generalist LLMs: The Power of Personality-Driven Expert Agents Ecosystem

Anonymous ACL submission

Abstract

The rapid advancement of natural language processing (NLP) has been propelled by large language models (LLMs), yet their monolithic nature often results in inefficiencies, particularly in specialized training for complex tasks. Addressing this, we introduce *Pool of Experts*, a novel multi-agent LLM framework that facilitates role specialization through prompt-based agentification, circumventing the computational cost of fine-tuning. The methodology involves a structured two-stage process: initialization, where agents are configured with distinct expert roles based on task context, and inference, where agents collaboratively generate responses. We evaluate the impact of expert role selection on task accuracy across multiple datasets, employing decision-making strategies like Majority Voting and a Final Decision Maker. Our system outperforms state-of-the-art systems in complex tasks, such as Strategy QA and Last Letters Concat. While no single framework consistently excels, the choice of framework significantly influences performance in tasks. This research paves the way for more sophisticated and structured NLP systems, contributing to the advancement of multi-agent LLMs.

1 Introduction

The swift evolution of natural language processing (NLP) has been primarily fueled by the emergence of large language models (LLMs) that demonstrate exceptional performance across diverse tasks. However, the monolithic nature of these models often leads to inefficiencies in terms of time and energy consumption, particularly when specialized training or fine-tuning is required for complex tasks. This paper addresses a critical gap in the field: the need for scalable, interpretable, and adaptable architectures that can efficiently handle multifaceted challenges without the overhead of traditional specialization methods.

Recent work in NLP has explored various strategies for improving model efficiency and performance, such as transfer learning and model distillation. However, these approaches often fall short in terms of scalability and adaptability, particularly when applied to tasks requiring diverse skill sets. Our research builds on these foundations by proposing a novel paradigm that leverages the strengths of multi-agent architectures. Unlike previous work, which typically focuses on single-agent models, our approach envisions a society of LLMs, each specializing in distinct roles such as perception, reasoning, and decision-making. This multi-agent framework draws inspiration from the layered approach of perceptrons in deep learning, where low-level features integrate with higher-level abstractions to achieve sophisticated outcomes.

The contribution of this paper is threefold. First, the *"Pool of Expert"* framework. We introduce the concept of "Pool of Expert" through prompt-based agentification, which allows for efficient role specialization without the need for extensive retraining.

Second, the analysis of Expert Role Selection. We provide a comprehensive analysis of the impact of expert role selection on task accuracy across various datasets, demonstrating that our approach can lead to significant improvements in performance.

Third, a comparison of Decision-Making Strategies. We present a detailed comparison between the Final Decision Maker and Majority Vote methods, highlighting the conditions under which each method excels.

Our findings are supported by rigorous statistical analyses, including weighted ANOVA and Kruskal-Wallis tests, ensuring the robustness of our conclusions. The code for our experiments is publicly available[1].

[1] https://anonymous.4open.science/r/PoE_Small_originalPaper-4500

Our approach aligns with the broader vision of creating a multi-agent society architecture in NLP, where model collaboration mirrors human interactions, akin to how deep learning structures reflect brain architecture. By transitioning from isolated, monolithic models to an interconnected ecosystem, we unlock new possibilities for scalability, interpretability, and adaptability. This paradigm shift not only addresses current limitations but also sets the stage for future research in multi-agent interactions, ultimately contributing to the development of more sophisticated and efficient NLP systems.

2 Related Work

Since this is a relatively new research field, the literature is limited. Therefore, we include some non-peer-reviewed works to illustrate current efforts. We analyze two areas of the literature: multi-agent systems and equipping agents with personality.

Multi-Agent Systems (MASs) Recent advancements in Multi-Agent Systems (MASs) highlight the potential of LLMs to enable dynamic and collaborative behaviors. Liang et al. (2024) address the Degeneration-of-Thought (DoT) problem in LLMs by introducing a debate framework that leverages structured interactions and controlled disagreement to foster divergent thinking, demonstrating the importance of strategic agent coordination. Similarly, Zhang et al. (2024a) propose ProAgent, which equips agents with proactive behavior and dynamic adaptation capabilities through intention prediction and alignment, though it does not explicitly optimize for domain-critical requirements.

Nascimento et al. (2023) leverage the MAPE-K model to integrate LLMs into MASs, enhancing self-adaptation through improved communication and decision-making in dynamic environments. Complementing these efforts, Xu et al. (2024) design a MAS where agents interact competitively and cooperatively in social deduction games. Using a probabilistic graphical model, their approach models dependencies among agents' beliefs and actions, leading to the emergence of strategic behaviors in multi-agent scenarios.

To address the challenge of scaling MASs, LONGAGENT (Zhao et al., 2024) introduces a collaborative framework capable of processing text inputs exceeding 128k tokens. This system employs a hierarchical structure where a leader agent coordinates task decomposition, conflict resolution, and communication, yielding emergent behaviors driven by roles and strategies rather than explicitly defined personalities. Similarly, MALLM (Becker, 2024) demonstrates how assigning expert personas to agents facilitates task-specific interactions, enabling adaptation across complex problem-solving paradigms.

Applications extend beyond traditional domains. He et al. (2025) demonstrate MASs in software engineering, where specialized agents collaborate through structured communication, while Lim et al. (2024) apply MASs to manufacturing, emphasizing adaptability under dynamic conditions. Finally, Zhang et al. (2024b) propose a Cooperative Embodied Language Agent framework for decentralized environments, where emergent behaviors arise from role-based communication and decision-making in resource-constrained settings.

Agent Personality Early work by Kim et al. (2019) emphasized integrating personality traits into conversational agents using psychological models, such as the Big Five traits (Roccas et al., 2002a). Recent studies have focused on enhancing interaction quality, personalization, and reliability. In healthcare, personality-based approaches have gained attention. For instance, Hwang et al. (2021) demonstrates how familiar personas can enhance trust and engagement in healthcare interactions, highlighting the role of empathy and personalized design. Similarly, Ahmad et al. (2022) explore aligning agent responses with users' personality traits to improve therapeutic efficiency. Further advancements include methods to express "roles" that enhance LLM reasoning capabilities. Kong et al. (2024) employ a prompt-based approach to improve context-aware responses, while in (Serapio-García et al., 2023), the authors embed personality traits in LLMs, demonstrating their capability to produce diverse and customizable interactions when guided by appropriate prompts. Recently, Dong et al. (2024) investigated using LLMs for personalized preference evaluation, emphasizing challenges such as simplistic personas and the need for better alignment. Similarly, Liu et al. (2022) integrate persona data to enrich conversational context and response quality. Efforts to create diverse and realistic personas have been introduced by Schuller et al. (2024), aiming to enhance human-technology interactions. In addition, Jandaghi et al. (2024) generate persona-aligned dialogues to build Synthetic-Persona-Chat (SPC) datasets. Finally, integrating expert-driven approaches (Long et al., 2024; Chai

et al., 2024) has shown how expert contributions can bolster LLM reliability and reduce biases.

3 Expert-based Approach

This section outlines the proposed methodology, called **Pool of Experts**, which leverages an ecosystem of LLM-based agents to enhance decision-making in complex tasks. This approach integrates specialized agents to tackle challenges such as multi-step decision-making, operational research requiring domain expertise (Xiao et al., 2023), and tasks affected by data scarcity. By mimicking real-world collaborative environments, agents are designed to reflect professional roles, ensuring that the decision-making process is structured, efficient, and aligned with human-like reasoning.

Our approach is implemented through a multi-agent system structured in two main stages: (i) initialization and (ii) inference. The initialization stage defines and configures agents based on the task context, ensuring alignment with the problem domain. A description framework guides this process to generate the personality description of agents within our system. This allows agents to shape their behavioral and operational characteristics. The inference stage processes incoming queries by leveraging agent interactions to provide structured responses.

Agent Initialization The initialization phase begins with the creation of the **Psychologist Agent** (PA), responsible for generating the profiles of all agents in the system (Figure 1-A). The PA defines personality profile descriptions by incorporating behavioral and operational characteristics based on psychological frameworks. For example, when generating the profile for the **Project Manager** (PM) agent, the PA emphasizes strategic thinking and organizational skills. Following the PA's creation, the system generates the **Project Manager (PM)** agent (Figure 1-B). Unlike expert agents who directly solve tasks, the PM focuses on identifying the expertise fields required for problem-solving. By evaluating project objectives and constraints, the PM ensures the selection of agents with the necessary skills for the task (e.g., selecting a physician and an obstetrician for pregnancy-related queries). This adaptive process means different tasks or contexts generate different PM profiles, ensuring flexibility and task-specific expertise allocation. Once the PM has compiled a list of required expertise fields, the **Expert Agents** (EAs) are generated (Figure 1-C). The PA assigns each EA a specialized domain, guiding profile generation based on the chosen description framework (e.g., UDP). A distinctive feature of the EAs is their ability to self-evaluate their responses by assigning confidence scores that quantify their certainty and familiarity with the subject matter. Additionally, EAs provide a justification grade, measuring the alignment of their reasoning steps with the task's requirements, ensuring transparency in the decision-making process.

The final agent in the framework is the **Final Decision Maker (FDM)** (Figure 1-D). Directly generated by the PA, the FDM aggregates and evaluates responses from the expert agents. It systematically reviews expert answers, identifies inconsistencies, and selects the most well-supported solution based on confidence levels and logical coherence. This step ensures that the final response to a query is reliable, consistent, and fully justified.

Inference Process During inference, queries are presented to the team of expert agents, who independently analyze the problem within their respective domains. Each agent generates a response based on its expertise and provides an associated confidence score. Once all expert responses are collected, the FDM reviews their reasoning, confidence levels, and supporting justifications to synthesize a final, well-supported answer. This hierarchical structured multi-agent system enhances problem-solving capabilities by distributing expertise across specialized agents, ensuring that solutions are accurate and explainable.

4 Research Questions

In this Section, we outline the motivation of our research questions, detailing the importance of each aspect in the development and evaluation of Multi-Agents System LLM-based.

RQ1 Is there a universally optimal description framework for constructing a profile that transforms an LLM into an agent?

RQ2 How does the choice of a description framework influence performance on different datasets and tasks?

RQ3 How does the expertise field of an agent influence performance, and does agent position bias (e.g., Agent 1 consistently outperforms Agent 2, etc.) affect the outcomes?

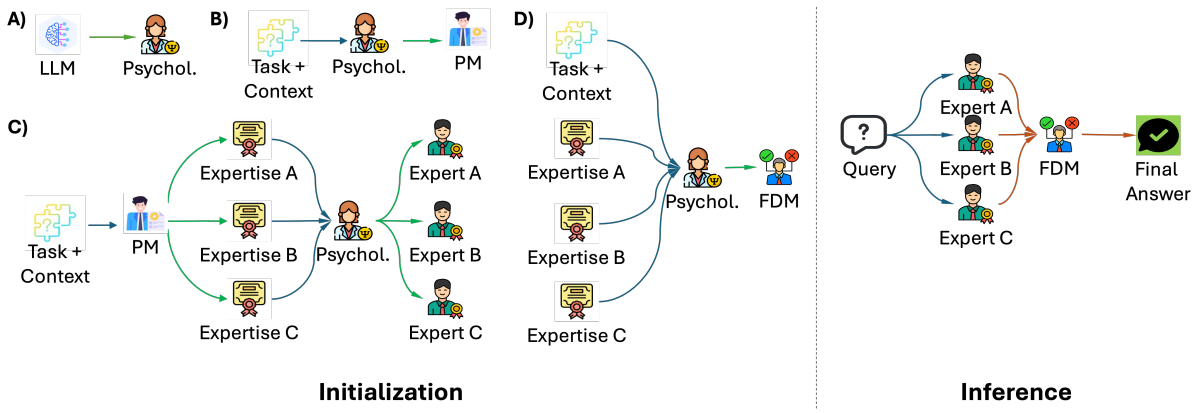

Figure 1: Blue arrows indicate data an agent considers; green arrows indicate a profile generation; orange arrows indicate answer generation.

RQ4 What impact does the Project Manager agent have on guiding the generation of expertise fields for expert agents and adapting them to improve problem-solving based on the specific task at hand?

The effectiveness of an LLM-based agent depends heavily on how its identity and behavioral traits are structured. Thus, we first investigate **RQ1**, which explores the most suitable description framework for constructing a profile that transforms an LLM into an agent. Establishing a well-defined framework is crucial, as it dictates how an agent processes information, interacts with users, and makes decisions. Additionally, different description frameworks may influence the model's performance in various ways, which motivates **RQ2**. It is important to determine whether certain frameworks consistently lead to superior results or if their impact is negligible. This insight would inform the best practices for agent design, ensuring that descriptions align with the intended application. Once an agent's description is established, its expertise field plays a critical role in determining its effectiveness. **RQ3** examines whether defining expertise fields leads to performance differences and whether agent position bias exists (e.g., whether one agent consistently outperforms another due to ordering effects). Understanding these factors allows us to assess whether expertise specialization enhances or limits an agent's capabilities and whether additional architectural adjustments are needed to mitigate potential biases. The role of the *Project Manager* agent is also a key aspect of agent coordination and task-solving. **RQ4** investigates how the Project Manager agent influences the generation of expertise fields for expert agents and adapts them to improve problem-solving based on the specific task at hand. If the Project Manager effectively enhances expertise allocation, it could be a pivotal component in structured agent-based reasoning systems.

5 Experimental Settings

This Section describes the experimental manipulations conducted to address our research questions. We implement our approach using the *Meta Llama 3.1 70B Instruct* LLM. The specific implementation details and model parameters are provided in Appendix G. To explore our research questions, we systematically manipulate the description frameworks used to generate agent personality descriptions. These frameworks differ in their emphasis on personality, cognition, behavior, and design. To evaluate the impact of different personality description methods, we test them on diverse datasets covering a range of tasks, including commonsense and social reasoning, symbolic manipulation, and implicit multi-step inference.

5.1 Description Frameworks to Generate Agent Personality

The examined description frameworks provide diverse approaches to modeling personality, cognition, behavior, and user interaction. Trait-based models such as the *Big Five* and *Myers-Briggs* define stable personality characteristics, while *Freudian Psychoanalysis* and *Erikson's Psychosocial Stages* focus on unconscious processes and lifelong development. Cognitive theories, including *Cognitive Behavioral Theory*, *Cognitive Load Theory*, and *Dual-Process Theory*, explore how individuals process information and make decisions.

Social Cognitive Theory and *Flow Theory* highlight the impact of environmental and motivational factors on human behavior.

In our approach, these frameworks serve as the profile description models we use to generate personality-based descriptions of agents. By considering user-centered approaches like *User Design Persona*, *User-Centered Design*, and *Mental Models*, we create a broad comparison of very different perspectives, showing how people are understood across multiple domains. Indeed, user-centered approaches like *User Design Persona*, *User-Centered Design*, and *Mental Models* focus on how people interact with systems and technology to enhance usability and efficiency.

Adopting the **Big Five Personality Traits** (Roccas et al., 2002b), a person is described based on five fundamental traits: openness, conscientiousness, extraversion, agreeableness, and neuroticism. These traits capture broad patterns of thoughts, emotions, and behaviors that shape personality. Individuals vary along these dimensions, influencing their social interactions, decision-making, and emotional responses.

The **Myers-Briggs Type Indicator** (Myers and Myers, 1995) categorizes a person into one of sixteen personality types based on four dichotomies: introversion vs. extraversion, sensing vs. intuition, thinking vs. feeling, and judging vs. perceiving. This framework describes how individuals process information, make decisions, and engage with the world. It emphasizes cognitive preferences rather than fixed traits.

Following the **Freudian Psychoanalysis** (Freud and Brill, 1997), a person is shaped by unconscious psychological drives and the dynamic interplay of the id, ego, and superego. Behavior results from internal conflicts and repressed desires manifest in personality and emotional struggles. Freud's model highlights the influence of past experiences and unconscious motivations on behavior.

Erikson's Psychosocial Stages (Escalona, 1951) describe a person's development through eight psychosocial stages, each defined by a central conflict between individual needs and societal expectations. Successfully resolving each stage leads to psychological growth and virtues such as trust, autonomy, and identity. Failure to resolve conflicts may result in recurring challenges later in life.

Differently, the **Enneagram of Personality Traits**(Riso and Hudson, 1996) classifies a person into one of nine interconnected personality types, each associated with distinct motivations, fears, and behavioral patterns. These types illustrate personal growth paths and stress responses. The Enneagram is often used for self-awareness and interpersonal understanding, notably it lacks strong empirical validation.

The **Social Cognitive Theory** (Blackwell, 1942) affirms that a person learns and develops through observation, imitation, and interaction with their environment. Cognitive processes, social experiences, and reinforcement influence behavior. This theory emphasizes the reciprocal relationship between personal factors, behavior, and situational influences.

In another dimension, **Cognitive Behavioral Theory**(Beck, 1976) claims that emotions and behaviors are the same, often through automatic cognitive patterns to shape a person's thoughts. Distorted thinking can lead to emotional distress, but restructuring these thoughts can improve mental well-being.

User Design Persona (Cooper and Saffo, 1999) is adopted to represent a fictional person. It captures the description of a person using a research-based user archetype that captures key goals, behaviors, and frustrations. Personas are widely adopted to guide product design by reflecting real user needs and expectations. Effective personas ensure that designs align with user motivations and enhance usability.

Cognitive Load Theory (Sweller, 1988) focuses on the limited cognitive resources of a person. Thus, a person can become overwhelmed when processing excessive information. Learning is optimized when extraneous cognitive load is minimized, allowing focus on essential content. Effective instructional design structures information to align with cognitive capacity.

Following another perspective taken by the**User-Centered Design** (Norman and Draper, 1986), a person's needs, behaviors, and experiences should drive the design of products and systems. Intuitive designs align with users' mental models, minimizing frustration and enhancing engagement. Systems that adapt to users create seamless and efficient interactions.

Mental Models(Johnson-Laird, 1980) theory constructs internal representations of reality to guide a person's reasoning, decision-making, and problem-solving. These models shape how they interpret experiences and anticipate outcomes. The theory claims effective communication and learn-

ing depend on aligning mental models with real-world structures.

Flow Theory (Csikszentmihalyi, 1975) focus and intrinsic motivation of a person when his/her skills match the challenge of an activity. In this so-called flow state, a person loses track of time and engages effortlessly. This state fosters creativity, productivity, and personal fulfillment.

Finally, in the **Dual-Process Theory** (Wason and Evans, 1974), a person thinks using two cognitive processes: an automatic, intuitive system and a deliberate, analytical system. Implicit processes drive quick, instinctive decisions, while explicit reasoning enables reflective thought. These systems interact to shape judgments, learning, and behavior.

5.2 Datasets

We evaluate our approach on diverse datasets spanning commonsense reasoning, social understanding, symbolic manipulation, and implicit multi-step inference. These datasets were chosen to comprehensively test the ability of language models across varied reasoning tasks.

The **Common Sense QA** dataset (Talmor et al., 2019) is a multiple-choice question-answering benchmark designed to assess commonsense reasoning. It requires models to combine factual and contextual knowledge to predict the correct answer. Human performance on this dataset is 88.9%, while the current state-of-the-art (SoTA) model, DeBERTaV3-large + KEAR (Xu et al., 2022), achieves 91.2% accuracy, surpassing human performance. This model incorporates external knowledge and is fine-tuned on retrieved examples from related datasets, enabling improved commonsense reasoning.

The **Last Letter Concat** dataset (Wei et al., 2022) evaluates symbolic reasoning by requiring models to concatenate the last letters of words to form new strings. This abstract task challenges the models' ability to perform transformations unrelated to language understanding. The SoTA method, NUDGING (Fei et al., 2024), employs a training-free, inference-time alignment strategy, which inject alignment tokens during uncertain predictions to improve accuracy up to 86.0%.

The **Social IQa** dataset (Sap et al., 2019) assesses commonsense reasoning in social contexts, requiring models to infer motivations, intentions, and emotions from everyday scenarios. Human performance is 87%, while large models like GPT-3 achieve a maximum accuracy of 58%. The SoTA model, UNICORN (Lourie et al., 2021), employs multitask learning to enable generalization. It achieves 83.2% accuracy.

The **Strategy QA** dataset (Geva et al., 2021) is designed to evaluate implicit multi-step reasoning, where models must infer information beyond what is explicitly stated in the question. Human performance on this dataset is 87%, while fine-tuned models typically achieve around 66%, highlighting the task's inherent complexity. The SoTA model, ROBERTA* last-step ORA-P-D, proposed by the authors of the dataset, achieves an accuracy of 72.0%. This approach involves fine-tuning single-step reasoning tasks , combined with a retrieval-based method incorporating gold decompositions and relevant context.

The **Social Support QA** dataset (Wang and Jurgens, 2018) presents a ternary classification task to assess a model's ability to classify interactions as supportive, neutral, or unsupportive. Despite advancements in large-scale pre-trained models, performance remains challenging under zero-shot conditions. The SoTA method, a Random Forest classifier trained on 23,903 features, achieves a Macro-F1 score of 0.52, outperforming baselines. However, the paper does not specify which portion of the published dataset[2] was used for training and testing. Thus, we compare our results with the repository's best-performing zero-shot model baseline, GPT-3, which achieves a Macro-F1 score of 0.29.

6 Discussion

In this section, we analyze and discuss the experimental results by summarizing the statistical findings. Complete statistical analyses are reported in the Appendixes.

We begin by comparing the performance of our approach with human performance and state-of-the-art (SOTA) models on each dataset. The results are reported in Table 1. Since we tested different subsets of the datasets for each description framework, the table shows the confidence intervals approximating the true mean of the Final Decision Maker scores across the description frameworks.

Starting with the *Strategy QA* dataset, our approach significantly outperforms the current SOTA, setting a new upper bound that is closer to human performance. Furthermore, considering that the

[2] github.com/google/BIG-bench/tree/main/bigbench/benchmark_tasks/social_support

Dataset	Our CI Low	Our CI High	Human Performance	SoTA
Strategy QA	78.09%	79.21%	87.00%	72.00%
Common Sense QA	81.90%	83.70%	88.90%	91.20%
Last Letter Concat	90.16%	91.56%	100.00%	86.00%
Social IQa	78.37%	79.45%	87.00%	83.20%
Social Support (Acc.)	40.89%	42.49%	n/a	n/a
Social Support (F1)	0.24	0.30	n/a	0.29

Table 1: Performance comparison across datasets, showing confidence intervals (CI) for our approach, as well as SoTA and human-level benchmarks where applicable. For the Social Support dataset we report both accuracy and Macro F1 scores. Refer to Appendix A for the computation of the CI.

current SOTA is a fine-tuned system, our approach represents a considerable advancement. On the *Last Letter* dataset, our approach also establishes a new upper-bound performance. In the case of the *Common Sense QA* dataset, the performance is statistically lower than both human performance and the SOTA. However, given that the SOTA is a strongly fine-tuned model that extensively leverages external knowledge, we can conclude that our approach still performs competitively. A similar observation holds for the *Social IQa* dataset, where our results are closer to both the multitask learning system and human performance.

To address **RQ1**, we performed a global analysis using chi-square tests (reported in Appendix B) to compare the performance of different description frameworks based on both the Majority Vote and Final Decision Maker outcomes. These tests aimed to determine whether there were statistically significant differences between the frameworks. The statistical results considering the Majority Vote outcomes in isolation do not indicate a significant effect. In other words, when a majority vote method is adopted to aggregate decisions among expert agents, the choice of description framework does not substantially affect the overall outcome. Similarly, we did not observe any statistically significant effect when considering the Final Decision Maker outcomes; although variations exist, they are minimal and not statistically valid. Finally, we performed a global chi-square test to compare the Final Decision Maker outcomes against the Majority Vote outcomes. The high p-value led us to fail to reject the null hypothesis. Thus, we cannot conclude that the choice of description framework for defining the personality of the agents in our approach is superior overall.

To investigate our **RQ2** we adopt Weighted Mann–Whitney U test for pairwise comparison within datasets and the Weighted Wilcoxon Signed-Rank test for overall comparisons between the Final Decision Maker and Majority Vote approaches (we report the complete anslyses in Appendix C). Overall, qur analysis reveals that the choice of description framework significantly influences performance, though the extent and nature of this effect vary by dataset and task. For the **Strategy QA**, dataset, both the Final Decision Maker and Majority Vote approaches yield statistically significant differences between any pair of description frameworks. This finding indicates that when optimizing for Strategy QA, one cannot assume that all frameworks are interchangeable, as each framework produces distinct performance outcomes. In the case of **Common Sense QA**, although the overall weighted distribution of accuracy differences does not deviate sufficiently to declare one framework as systematically superior, pairwise comparisons via the Weighted Mann–Whitney U test reveal that even small raw differences become statistically significant. Thus, while the aggregate impact may appear modest, subtle but consistent differences exist between the frameworks. For the **Last Letter Concat** dataset, our analyses clearly demonstrate a significant difference between the Final Decision Maker and Majority Vote approaches. In this dataset, certain frameworks (e.g., those based on Big Five Personality Traits) yield similar accuracy distributions under the Final Decision Maker approach. This suggests that, in practice, some frameworks may be effectively interchangeable for this particular task. In the **Social IQa** dataset, the statistical results underscore that, in tasks requiring social reasoning, the specific description framework is critical to performance. Finally, in the **Social Support** dataset, demonstrating that even though the absolute performance may be low, significant differences exist between the aggregation

methods.

In summary, these results demonstrate that the influence of the description framework is highly task-dependent. Therefore, we can answer positively to our **RQ2** since the adoption of different description frameworks (and ensembling methods) leads to statistically different outcomes.

To investigate whether the expertise field of an agent influences performance—and to confirm that expert agent position is not biased—we conducted weighted ANOVA and weighted Kruskal–Wallis analyses (reported in Section D). The results demonstrate that the influence of expert role selection is highly task-dependent. We further performed a finer comparison by contrasting the roles within each dataset using Tukey's HSD post-hoc pairwise test. The results revealed that specific expert roles significantly differ in performance. Our analysis of expert role selection shows a strong correlation between domain expertise and task performance. Furthermore, the results indicate that the assigned roles align with the nature of the task at hand, suggesting that certain specializations are more suited to specific tasks. For example, in Social IQa, the most frequent expert roles (i.e., *Emotional Intelligence*) indicate that agents specializing in human interaction were more successful in this context. Next, we adopted chi-square tests (reported in Section E) to statistically demonstrate that the ordering of agents does not affect expert agent performance. These findings further support our conclusion that expert performance is dependent on the role assigned rather than on the agent's position in the ordering. Each agent contributes fairly to the aggregated result without any positional bias. Thus, we can affirm our hypothesis (**RQ3**) that expert role selection has a direct impact on the outcomes and that this effect is not due to any positional bias of the agents.

The analysis conducted in Appendix D statistically demonstrates that the influence of expert role selection is highly task-dependent. These findings are supported by the role distributions, which reveal distinct sets of roles selected based on the dataset. In Common Sense QA and Last Letter Concat, the impact is relatively modest, whereas in other datasets the effect is more pronounced. The Project Manager agent is therefore crucial in selecting the most appropriate expert roles to shape performance. This underscores not only the importance of task-dependent role allocation strategies in our multi-agent system but also the efficacy of the Project Manager agent within the system. The Project Manager agent exhibits the flexibility necessary to successfully select the appropriate roles for the expert agents in solving tasks. Thus, we answer positively to our **RQ4** that the Project Manager agent has a positive impact on guiding the generation of expertise fields for the expert agents and is able to adapt to task-specific requirements with flexibility. The results further demonstrate that the PM agent played a crucial role in optimizing expert selection by dynamically adjusting role allocation based on the task at hand, with its decisions strongly correlating with task-specific performance improvements.

7 Conclusions

This study introduces *Pool of Experts*, a multi-agent LLM-based framework that employs a consortium of specialized agents, each dedicated to distinct functions such as perception, reasoning, and decision-making. Drawing inspiration from human interaction akin to the influence of brain structure on deep learning, our approach demonstrates that distributing expertise among agents enhances accuracy and robustness without necessitating costly fine-tuning. We conducted a systematic analysis of expert role selection, decision-making strategies, and agent collaboration mechanisms, underscoring the efficacy of structured multi-agent architectures.

Key insights include: (1) While no single description framework consistently excels, the choice of framework significantly impacts performance across tasks and datasets. (2) Expertise field selection is critical and task-specific, with the Project Manager agent crucially guiding expertise allocation to enhance task-specific outcomes. (3) Ensemble methods, particularly those utilizing a FDM offer more accurate and reliable decisions. Notably, our approach surpasses SoTA in Strategy QA and Last Letter Concat, while remaining competitive in Common Sense QA and Social IQa.

Despite its advantages, our approach faces limitations such as increased computational costs and reliance on predefined role assignments. Future work should explore adaptive methods for role selection and decision aggregation to enhance flexibility and generalization. This research contributes to the advancement of multi-agent LLMs, highlighting structured agent collaboration as a promising alternative to traditional single-model architectures.

8 Limitations

Despite the effectiveness of our *Pool of Experts* framework, several limitations must be acknowledged.

The first limitation concerns the computational cost associated with multi-agent decision-making. Since multiple agents contribute responses before reaching a final decision, the approach requires significantly more inference calls compared to single-agent models. This increased resource consumption may pose scalability challenges in scenarios with computational constraints.

Furthermore, our system operates purely in a zero-shot manner, without task-specific fine-tuning. While this ensures broad applicability and reduces the computational cost associated with training, it also limits adaptation to domain-specific tasks where fine-tuned models may offer better performance. As a result, our approach may not always be competitive with domain-specialized systems that leverage additional training data, although it remains a strong candidate for tasks where no training data are available.

Regarding the chosen datasets, our evaluation relies on existing benchmark datasets which, while widely used, are limited to inference-type tasks. This restriction stems from the nature of our approach, which does not modify the model's internal knowledge. Consequently, the datasets may not fully capture the complexity and variability posed by more challenging tasks.

In our experiments, we limit the generation of expert agents to three. This constraint may negatively impact overall performance and lead to an underestimation of the importance of the Final Decision Maker role. Furthermore, we did not explicitly assess whether biases emerge during the generation of expert agents, which could influence decision-making outcomes. More tests are required to provide more robust evidence for our findings.

Lastly, we restrict our experimental investigation to the Llama 3.1 70B instruct model. Although this limits our ability to generalize findings to other large-language models, it is very probable that the observed behaviors and results will extend to similar models given the shared architectural and operational principles of state-of-the-art LLMs. Expanding the framework to other architectures remains an important direction for future work.

9 Ethical Considerations

The use of multi-agent systems powered by large language models (LLMs) raises several ethical considerations. One primary concern is the potential for biased decision-making. While our system distributes expertise among multiple agents, it ultimately relies on models that inherit biases from their training data. This could lead to unintended amplification of stereotypes or skewed reasoning, particularly in high-stakes applications such as legal or medical decision-making.

Although our framework ensures that each agent's reasoning process is explicitly visible in the responses it provides, transparency does not necessarily equate to interpretability. Further research is needed to evaluate how users perceive and trust multi-agent LLM decisions in real-world scenarios. Additionally, the configuration of the system with a FDM introduces a vulnerability to hallucinations, similar to those observed in single-agent systems. This limitation is particularly concerning in domains where incorrect or fabricated information could have significant consequences.

Finally, the environmental impact of large-scale LLMs must be acknowledged. Multi-agent architectures inherently involve generating multiple outputs per task, increasing computational costs. While our method avoids expensive fine-tuning, the energy consumption associated with repeated inference calls remains a concern. Efforts should be made to balance the benefits of multi-agent reasoning with the need for sustainable AI practices.

References

Rangina Ahmad et al. 2022. Designing personality-adaptive conversational agents for mental health care. *Information Systems Frontiers*, 24(3).

A.T. Beck. 1976. *Cognitive Therapy and the Emotional Disorders*. International Universities Press.

Jonas Becker. 2024. Multi-agent large language models for conversational task-solving. *Preprint*, arXiv:2410.22932.

Gordon W. Blackwell. 1942. Social learning and imitation. by neal e. miller and john dollard. new haven: Yale university press, 1941. 341pp. $3.50. *Social Forces*, 21(2):256–256.

Ziwei Chai et al. 2024. An expert is worth one token: Synergizing multiple expert LLMs as generalist via expert token routing. In *Proceedings of the 62nd Annual Meeting of the ACL (Volume 1: Long Papers)*.

W. J. Conover. 2006. *Practical Nonparametric Statistics 3Rd Ed*. Wiley India Pvt. Limited.

Alan Cooper and Paul Saffo. 1999. *The Inmates Are Running the Asylum*. Macmillan Publishing Co., Inc., USA.

M. Csikszentmihalyi. 1975. *Beyond Boredom and Anxiety*. Jossey-Bass behavioral science series. Jossey-Bass Publishers.

Yijiang River Dong et al. 2024. Can LLM be a personalized judge? In *Findings of the ACL: EMNLP 2024*. ACL.

Sibylle Escalona. 1951. childhood and society. erik h. erikson. new york: Norton, 1950. 397 pp. $4.00. *Science*, 113(2931):253–253.

Yu Fei, Yasaman Razeghi, and Sameer Singh. 2024. Nudging: Inference-time alignment via model collaboration. *CoRR*, abs/2410.09300.

S. Freud and A.A. Brill. 1997. *The Interpretation of Dreams*. Classics of World Literature. Wordsworth Editions.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Junda He, Christoph Treude, and David Lo. 2025. Llm-based multi-agent systems for software engineering: Literature review, vision and the road ahead. *ACM Trans. Softw. Eng. Methodol.* Just Accepted.

Youjin Hwang, Donghoon Shin, Sion Baek, Bongwon Suh, and Joonhwan Lee. 2021. Applying the persona of user's family member and the doctor to the conversational agents for healthcare. *Preprint*, arXiv:2109.01729.

Pegah Jandaghi et al. 2024. Faithful persona-based conversational dataset generation with llms. In *Proceedings of the 6th Workshop on NLP4ConvAI 2024*, pages 114–139. ACL.

Philip N. Johnson-Laird. 1980. Mental models in cognitive science. *Cogn. Sci.*, 4:71–115.

Hankyung Kim et al. 2019. Designing personalities of conversational agents. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*.

Aobo Kong et al. 2024. Better zero-shot reasoning with role-play prompting. In *Proceedings of the 2024 Conference of the North American Chapter of ACL: HLT (Volume 1: Long Papers)*.

William H. Kruskal and W. Allen Wallis. 1952. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260):583–621.

Tian Liang et al. 2024. Encouraging divergent thinking in llms through multi-agent debate. In *Proceedings of the 2024 Conference on EMNLP*, pages 17889–17904. ACL.

Jonghan Lim, Birgit Vogel-Heuser, and Ilya Kovalenko. 2024. Large language model-enabled multi-agent manufacturing systems. *Preprint*, arXiv:2406.01893.

Junfeng Liu, Christopher Symons, and Ranga Raju Vatsavai. 2022. Persona-based conversational ai: State of the art and challenges. *Preprint*, arXiv:2212.03699.

Do Xuan Long et al. 2024. Multi-expert prompting improves reliability, safety and usefulness of llms. In *Proceedings of the 2024 Conference on Empirical Methods in NLP*. ACL.

Nicholas Lourie, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. UNICORN on RAINBOW: A universal commonsense reasoning model on a new multitask benchmark. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13480–13488. AAAI Press.

H. B. Mann and D. R. Whitney. 1947. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1):50 – 60.

I.B. Myers and P.B. Myers. 1995. *Gifts Differing: Understanding Personality Type*. Mobius.

Nathalia Nascimento et al. 2023. Self-adaptive llms-based multiagent systems. In *2023 IEEE ACSOS-C*.

Donald A. Norman and Stephen W. Draper. 1986. *User Centered System Design; New Perspectives on Human-Computer Interaction*. L. Erlbaum Associates Inc., USA.

D.R. Riso and R. Hudson. 1996. *Personality Types: Using the Enneagram for Self-Discovery*. Houghton Mifflin.

Sonia Roccas, Sagiv, et al. 2002a. The big five personality factors and personal values. *Personality and social psychology bulletin*, 28(6):789–801.

Sonia Roccas, Lilach Sagiv, Shalom H. Schwartz, and Ariel Knafo. 2002b. The big five personality factors and personal values. *Personality and Social Psychology Bulletin*, 28(6):789–801.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Andreas Schuller et al. 2024. Generating personas using llms and assessing their viability. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '24. ACM.

Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models. *Preprint*, arXiv:2307.00184.

John Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2):257–285.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4149–4158. Association for Computational Linguistics.

John W. Tukey. 1949. Comparing individual means in the analysis of variance. *Biometrics*, 5(2):99–114.

Zijian Wang and David Jurgens. 2018. It's going to be okay: Measuring access to support in online communities. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 33–45, Brussels, Belgium. Association for Computational Linguistics.

P.C. Wason and J.ST.B.T. Evans. 1974. Dual processes in reasoning? *Cognition*, 3(2):141–154.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

B. L. WELCH. 1947. The generalization of 'student's' problem when several different population varlances are involved. *Biometrika*, 34(1-2):28–35.

Ziyang Xiao, Dongxiang Zhang, Wu, et al. 2023. Chain-of-experts: When llms meet complex or problems. In *The Twelfth ICLR*.

Lin Xu, Zhiyuan Hu, Daquan Zhou, Hongyu Ren, Zhen Dong, Kurt Keutzer, See Kiong Ng, and Jiashi Feng. 2024. Magic: Investigation of large language model powered multi-agent in cognition, adaptability, rationality and collaboration. *Preprint*, arXiv:2311.08562.

Yichong Xu, Chenguang Zhu, Shuohang Wang, Siqi Sun, Hao Cheng, Xiaodong Liu, Jianfeng Gao, Pengcheng He, Michael Zeng, and Xuedong Huang. 2022. Human parity on commonsenseqa: Augmenting self-attention with external attention. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 2762–2768. ijcai.org.

Ceyao Zhang, Yang, et al. 2024a. Proagent: building proactive cooperative agents with llms. In *Proceedings of the 38th AAAI Conference on AI, 36th Conference on Innovative Applications of AI, and 14th Symposium on Educational Advances in AI*, AAAI'24/IAAI'24/EAAI'24. AAAI Press.

Hongxin Zhang, Weihua Du, Jiaming Shan, Qinhong Zhou, Yilun Du, Joshua B. Tenenbaum, Tianmin Shu, and Chuang Gan. 2024b. Building cooperative embodied agents modularly with large language models. *Preprint*, arXiv:2307.02485.

Jun Zhao, Can Zu, Hao Xu, Yi Lu, Wei He, Yiwen Ding, Tao Gui, Qi Zhang, and Xuanjing Huang. 2024. Longagent: Scaling language models to 128k context through multi-agent collaboration. *Preprint*, arXiv:2402.11550.

Appendices

A Final Decision Maker Comparison Against State-of-the-Art and Human Performance

In this section, we statistically compare our approach against the state-of-the-art (SoTA) method for each dataset. A straightforward way to perform a single, overall significance test comparing the Final Decision Maker's accuracy to the reported SoTA is to compute the weighted overall accuracy across all frameworks and then perform a binomial-proportion (z) test against the SoTA scores.

Because different frameworks were tested on different numbers of samples, we first aggregate our results by summing the total number of correct answers and the total number of items tested across all frameworks. Next, we convert each accuracy to a proportion by multiplying it by the corresponding number of samples, and then we sum these values to obtain the overall number of correct answers. Dividing this sum by the total number of samples across all frameworks gives us the overall weighted accuracy.

The null hypothesis is: H_0: *There is no difference between the overall weighted accuracy of our Final Decision Maker and the state-of-the-art accuracy.*

To compare against human performance, we computed the confidence interval of our weighted results.

A.1 Strategy QA Dataset

Framework	Accuracy (%)	Samples
Big Five Personality Traits	78.14	1560
Cognitive Behavioral Theory	78.41	1598
Cognitive Load Theory	79.40	1558
Dual-Process Theory	78.33	1546
Enneagram of Personality Traits	77.97	1593
Erikson's Psychosocial Stages	79.23	1666
Flow Theory	78.77	1639
Freudian Psychoanalysis	79.10	1603
Mental Models	79.28	1578
Myers-Briggs Type Indicator	78.58	1601
Social Cognitive Theory	77.85	1612
User Design Persona	78.54	1631
User-Centered Design	78.51	1629

Table 2: Per-framework performance on the *Strategy QA* dataset.

We report in Table 2 the outcomes of the Final Decision Maker. We want to test whether our model's true accuracy, denoted by p, differs from the state-of-the-art baseline, $p_0 = 0.72$. We conducted $N = 20{,}814$ trials, out of which we observed an overall accuracy of $\hat{p} = 0.7865$. Thus, we define the hypotheses as: $H_0 : p = 0.72$ and $H_1 : p \neq 0.72$. Under the null hypothesis H_0, the standard error of the proportion is approximately 0.0031. The test statistic is $Z \approx 21.3$ with a p-value $< 10^{-90}$, and thus we *reject* the null hypothesis. Our model's true accuracy is significantly higher than the 72% state-of-the-art; specifically, the point estimate is about 78.65%, which is a $+6.65\%$ improvement over the state-of-the-art. Statistically, the difference is extremely unlikely to be due to chance alone.

Our aggregated results yield an observed accuracy of $\hat{p} = 0.7865$ and a standard error (SE) of approximately 0.00284. A 95% confidence interval for the true accuracy is between 0.7809 and 0.7921. This indicates that if we were to repeat this evaluation on different samples of the same size from the dataset, we would expect the true accuracy of the Final Decision Maker to lie within this range 95% of the time. The narrow width of the interval reflects the large sample size and suggests that our estimate of approximately 78.65% accuracy is both precise and robust when generalizing to the entire dataset.

A.2 Common Sense QA Dataset

Framework	Accuracy (%)	Samples
Big Five Personality Traits	83.66	557
Cognitive Behavioral Theory	81.61	571
Cognitive Load Theory	83.75	560
Dual-Process Theory	81.68	546
Enneagram of Personality Traits	84.01	563
Erikson's Psychosocial Stages	82.93	574
Flow Theory	82.75	545
Freudian Psychoanalysis	82.72	567
Mental Models	81.00	579
Myers-Briggs Type Indicator	82.61	575
Social Cognitive Theory	82.48	548
User Design Persona	84.47	573
User-Centered Design	82.28	553

Table 3: Per-framework performance on the *Common Sense QA* dataset.

We report in Table 3 the performance of the Final Decision Maker across 13 different description frameworks. We want to test whether our model's true accuracy p, differs from the state-of-the-art baseline, $p_0 = 0.912$. We conducted $N = 7311$ trials, from which we computed an aggregated (weighted) accuracy of $\hat{p} \approx 0.828$. Thus,

we define the hypotheses as: $H_0 : p = 0.912$ and $H_1 : p \neq 0.912$. Under the null hypothesis H_0, the standard error of the proportion is approximately $\text{SE}_{H_0} \approx 0.00331$. The test statistic is $Z \approx -25.38$, with a p-value $< 10^{-100}$, and thus we *reject* the null hypothesis. Our model's true accuracy is significantly lower than the 91.20% state-of-the-art. To quantify the precision of our estimate, we calculate a 95% confidence interval for the model's true accuracy using the observed proportion $\hat{p} = 0.828$. The standard error is computed as SE ≈ 0.00442, yielding a 95% confidence interval between 0.819 and 0.837.

A.3 Last Letter Concat Dataset

Framework	Accuracy (%)	Samples
Big Five Personality Traits	92.20	500
Cognitive Behavioral Theory	92.00	500
Cognitive Load Theory	88.80	500
Dual-Process Theory	93.60	500
Enneagram of Personality Traits	91.60	500
Erikson's Psychosocial Stages	83.80	500
Flow Theory	89.00	500
Freudian Psychoanalysis	92.80	500
Mental Models	92.20	500
Myers-Briggs Type Indicator	92.60	500
Social Cognitive Theory	92.40	500
User Design Persona	92.60	500
User-Centered Design	87.60	500

Table 4: Per-framework performance on the *Last Letter Concat* dataset.

We report in Table 4 the performance of the Final Decision Maker. We want to test whether our model's true accuracy, denoted by p, differs from SoTA. From the aggregated data (weighted by the number of samples), we compute an accuracy of $\hat{p} \approx 0.9086$.

We first compare our aggregated accuracy to the state-of-the-art baseline, $p_0 = 0.86$. Thus, we define the hypotheses as: $H_0 : p = 0.86$ and $H_1 : p \neq 0.86$. With $N = 6500$, we have $\hat{p} = 0.9086$ and $p_0 = 0.86$. Under H_0, the standard error is $\text{SE}_{H_0} \approx 0.0043$. The test statistic is then $Z \approx 11.3$, which is extremely large. The corresponding p-value is effectively $\ll 10^{-20}$, and thus we *reject* H_0. Our model's accuracy is significantly higher than the 86% state-of-the-art.

We further compare our results to human performance, treating human performance as $p_0 = 1.0$, then under $H_0 : p = 1.0$ there is zero variance (since the probability of an error is 0). In this case, a standard Z-test is not defined; however, from a binomial perspective, if $p = 1.0$ then even a single error would have probability 0. Given that we observed 594 errors, the p-value under this scenario is trivially $< 10^{-100}$. Hence, our model's accuracy is *significantly below* the 100% human-performance level.

To quantify the precision of our estimate, we compute a 95% confidence interval for the model's true accuracy p. The standard error calculated is SE ≈ 0.00358. This indicates that we can be 95% confident that the model's true accuracy on this dataset lies between 90.16% and 91.56%.

In summary, when we aggregate results across all trials, the Final Decision Maker achieves an overall accuracy of about 90.86%. Statistical tests confirm that this accuracy is significantly higher than the 86% state-of-the-art approach ($Z \approx 11.3$, $p \ll 10^{-20}$) and significantly lower than a hypothetical 100% human-performance. The 95% confidence interval further indicates that the model's true performance is stable around approximately 91% when generalized to the entire dataset.

A.4 Social IQa Dataset

Framework	Accuracy (%)	Samples
Big Five Personality Traits	79.29	1700
Cognitive Behavioral Theory	79.47	1646
Cognitive Load Theory	78.89	1644
Dual-Process Theory	78.28	1625
Enneagram of Personality Traits	78.98	1670
Erikson's Psychosocial Stages	78.70	1737
Flow Theory	79.02	1654
Freudian Psychoanalysis	79.09	1722
Mental Models	78.48	1710
Myers-Briggs Type Indicator	78.05	1695
Social Cognitive Theory	78.99	1623
User Design Persona	80.58	1715
User-Centered Design	78.45	1666

Table 5: Per-framework performance on the *Social IQa* dataset.

We report in Table 5 the performance of the Final Decision Maker. We want to test whether our model's true accuracy, denoted by p, differs from established SoTA. From the aggregated data weighted by the number of samples, we compute an accuracy of $\hat{p} \approx 0.7891$.

We first compare our aggregated accuracy to the state-of-the-art baseline, $p_0 = 0.832$. Thus, we define the hypotheses as: $H_0 : p = 0.832$ and $H_1 : p \neq 0.832$. With $N = 21807$ trials, we have $\hat{p} = 0.7891$ and $p_0 = 0.832$. Under H_0, the standard error $\text{SE}_{H_0} \approx 0.00253$. The test statistic is then

$Z \approx -16.93$, with a p-value effectively close to zero ($p \ll 10^{-30}$). Thus, we *reject* H_0, concluding that our model's accuracy is significantly lower than the 83.20% state-of-the-art.

We further compare our results to a human performance of $p_0 = 0.87$. Here, the hypotheses are defined as: $H_0 : p = 0.87$ and $H_1 : p \neq 0.87$. Under H_0, the standard error $\text{SE}_{H_0} \approx 0.00228$. The resulting test statistic is $Z \approx -35.5$, yielding a small p-value of $\ll 10^{-200}$). We, therefore, *reject* H_0, conclude that our model's accuracy is significantly below the human performance of 87.00%.

To quantify the precision of our estimate, we calculate a 95% confidence interval for the model's true accuracy p. The standard error SE ≈ 0.00276, and the 95% confidence interval lies between 78.37% and 79.45%.

In summary, when aggregating results across all trials, the Final Decision Maker achieves an overall accuracy of about 78.91%. Statistical tests confirm that this performance is significantly lower than both the state-of-the-art and human performance.

A.5 Social Support Dataset

A.6 Accuracy

Since no state-of-the-art approaches have reported accuracy for this dataset, we only calculate the confidence interval for our approach. We treat these 13 outcomes (reported in Table 6) of the description frameworks as a sample of size $n = 13$. The sample mean is $\bar{X} = 41.69\%$, and the sample standard deviation is $s \approx 1.33\%$. To approximate the true mean performance μ of the method, we construct a 95% confidence interval using a t-based approach. With $n - 1 = 12$ degrees of freedom, the critical value is $t_{0.025,\,12} \approx 2.18$. The standard error is SE $\approx 0.37\%$, yielding a margin of error of $2.18 \times 0.37\% \approx 0.80\%$. Thus, the 95% confidence interval for the true mean lies between 40.89% and 42.49%.

This relatively narrow interval reflects the modest spread of the sample, as indicated by a standard deviation of about 1.33%.

A.6.1 Macro F1

Data, Accuracy, and Macro-F1 Scores We report in Table 6 the Final Decision Maker's macro-F1 scores across the 13 description frameworks. We want to test whether the average macro-F1 across these 13 runs (μ), differs from the state-of-the-art. From the data, we compute a sample mean of $\bar{X} = 27.60\%$ and a sample standard deviation of $s = 5.06\%$. Treating these 13 scores as independent observations, we perform a one-sample t-test with $n = 13$.

Framework	Accuracy (%)	Macro-F1 (%)	Samples
Big Five Personality Traits	40.02	22.86	897
Cognitive Behavioral Theory	43.48	30.34	897
Cognitive Load Theory	40.02	22.84	897
Dual-Process Theory	40.58	23.29	897
Enneagram of Personality Traits	42.59	29.85	897
Erikson's Psychosocial Stages	40.25	22.98	897
Flow Theory	42.25	29.84	897
Freudian Psychoanalysis	41.36	29.34	897
Mental Models	41.58	29.35	897
Myers-Briggs Type Indicator	42.81	40.25	897
Social Cognitive Theory	43.48	24.33	897
User Design Persona	40.47	23.15	897
User-Centered Design	43.03	30.40	897

Table 6: Per-framework performance on the *Social Support* dataset.

The hypotheses are defined as $H_0 : \mu = 29\%$ and $H_1 : \mu \neq 29\%$. The standard error is SE $\approx 1.40\%$. The test statistic $t \approx -1.00$. With $n - 1 = 12$ degrees of freedom, this t-value corresponds to a two-sided p-value of approximately 0.34, which is not statistically significant.

We also construct a 95% confidence interval for the true mean macro-F1, that lies between 24.54% and 30.66%. Since 29% lies within this interval, we fail to reject the null hypothesis.

In summary, across the 13 expert-model configurations, the Final Decision Maker achieves an average macro-F1 of 27.60% (with individual scores ranging from about 22.8% to 40.3%). Although this average is slightly below the state-of-the-art approach of 29%, the one-sample t-test indicates that there is insufficient evidence to conclude that the mean performance is statistically different from 29%.

B Description Framework Analyses

To understand which is the best-performing description framework, we performed a global analysis to compare the performance of different description frameworks based on both the Majority Vote outcome and the Final Decision Maker outcome.

For each framework, we aggregated the total number of successes and failures across all datasets and constructed a contingency table. A chi-square test was then applied to determine whether there were statistically significant differences in the outcomes performance among the frameworks. The chi-square test null hypothesis H_0 is: *there is no association between the description framework and the success/failure outcome of the ensembling*

method adopted, i.e., all frameworks perform equivalently.

B.1 Majority Vote Analysis

Framework	Successes	Failures
Big Five Personality Traits	3864.9936	1349.0064
Cognitive Behavioral Theory	3904.0689	1307.9311
Cognitive Load Theory	3827.0044	1331.9956
Dual-Process Theory	3784.0509	1329.9491
Enneagram of Personality Traits	3887.0043	1335.9957
Erikson's Psychosocial Stages	3995.9253	1378.0747
Flow Theory	3915.9107	1319.0893
Freudian Psychoanalysis	3950.0357	1338.9643
Mental Models	3899.0131	1364.9869
Myers-Briggs Type Indicator	3904.1322	1363.8678
Social Cognitive Theory	3875.9580	1304.0420
User Design Persona	4002.9301	1313.0699
User-Centered Design	3888.0289	1356.9711

Table 7: Global Contingency Table of Majority Vote Successes and Failures by Framework.

The chi-square test yielded a statistic of 5.54 with 12 degrees of freedom, corresponding to a p-value of 0.9376. Since the p-value is very high, we fail to reject the null hypothesis. Thus, there is no statistically significant difference in the Majority Vote performance among the various description frameworks. Table 7 shows the global contingency table with the aggregated majority vote successes and failures for each description framework. As the table indicates, the number of successes and failures are very similar across the frameworks. Such close results are consistent across the frameworks and reinforce the conclusion from the chi-square test.

Thus, these results suggest that, when a majority vote ensembling method is used to aggregate decisions, the choice of description framework does not substantially affect the outcome. Consequently, factors other than majority vote performance, such as theoretical underpinnings, interpretability, or computational efficiency, may be more critical when selecting a description framework for a given task or application.

B.2 Final Decision Maker Analysis

The chi-squared test on the aggregated contingency table yielded a chi-squared statistic of 4.54 with 12 degrees of freedom and a p-value of 0.9715. This very high p-value suggests that any observed differences in the number of successes and failures across the description frameworks are entirely attributable to random variation.

Framework	Successes	Failures
Big Five Personality Traits	3852.88	1361.12
Cognitive Behavioral Theory	3877.08	1334.92
Cognitive Load Theory	3805.98	1353.02
Dual-Process Theory	3761.01	1352.99
Enneagram of Personality Traits	3874.04	1348.96
Erikson's Psychosocial Stages	3943.05	1430.95
Flow Theory	3873.00	1361.99
Freudian Psychoanalysis	3933.92	1355.08
Mental Models	3896.01	1367.99
Myers-Briggs Type Indicator	3903.03	1364.97
Social Cognitive Theory	3840.96	1339.04
User Design Persona	3972.96	1343.04
User-Centered Design	3864.89	1380.11

Table 8: Global Contingency Table of Final Decision Maker Successes and Failures by Framework.

Table 8 reports the aggregated contingency table for the Final Decision Maker, showing the total number of successes and failures for each description framework. As observed, the success and failure counts across the frameworks vary only slightly. Such minimal variation supports the conclusion that no single description framework offers a performance advantage in terms of Final Decision Maker accuracy. Thus, as in the case of the majority vote outcome, the choice of description framework, when evaluated based on the final decision maker, should be guided by other factors.

B.3 Comparison Analysis: Final Decision Maker vs. Majority Vote

	Successes	Failures
Final Decision Maker	50398.81	17694.19
Majority Vote	50699.06	17393.94

Table 9: Global 2×2 Contingency Table.

We performed a global, weighted comparison of the Final Decision Maker method and the Majority Vote method to determine the most effective ensembling strategy. To this end, we aggregated the total number of successes and failures across all datasets. The overall accuracy for Final Decision Maker was 74.01%, while Majority Vote attained 74.46%. Although the Majority Vote shows a marginal advantage, the difference between the two methods is not statistically significant. We formally test this by stating the null hypothesis: H_0: *There is no association between the ensembling method (Final Decision Maker vs. Majority Vote) and the success*

or failure outcome. A chi-square test performed on the aggregated 2×2 contingency table (Table 9) yielded a chi-square statistic of 3.4609 with 1 degree of freedom and a p-value of 0.0628, leading us to fail to reject the null hypothesis.

Thus, we cannot conclude that one method is superior to the other overall.

C Final Decision Maker and Majority Vote Finer Comparison

Dataset	Wilcoxon Statistic	p-value	Significant?
Strategy QA	0.0000	0.000000	Yes
Common Sense QA	9422141.5000	0.277116	No
Last Letters	125250.0000	0.000000	Yes
Social IQa	108766137.0000	0.000000	Yes
Social Support	17301336.0000	0.000000	Yes

Table 10: Weighted Wilcoxon Test Results: Final Decision Maker vs. Majority Vote.

Weighted Mann–Whitney U test for pairwise comparison within datasets and the Weighted Wilcoxon Signed-Rank test for overall comparisons between the Final Decision Maker and Majority Vote approaches.

In this Section, we determine whether different description frameworks yield systematically distinct results to test the statistical differences between various description frameworks across datasets using non-parametric testing. Specifically, we compare the accuracy distributions of the Final Decision Maker and the Majority Vote ensembling methods. To achieve this, we employ the Weighted Mann–Whitney U test for pairwise comparisons (Mann and Whitney, 1947) within datasets and the Weighted Wilcoxon Signed-Rank test (Conover, 2006) for overall comparisons between the Final Decision Maker and Majority Vote approaches. The Mann–Whitney U test is a powerful non-parametric statistical method employed to evaluate the null hypothesis. This hypothesis asserts that, for randomly selected values X and Y from two distinct populations, the probability of X being greater than Y is precisely equal to the probability of Y being greater than X. A comprehensive formulation establishes that observations from both groups are assumed to be independent, responses are at least ordinal to enable the comparison of any two observations to determine which is greater, and under the null hypothesis (H_0), the distributions of both populations are considered identical while the alternative hypothesis (H_1) asserts that the distributions are not identical. The Wilcoxon signed-rank test is a non-parametric statistical method used for hypothesis testing. For two matched samples, the Wilcoxon test functions as a paired difference test, similar to the paired Student's t-test. This test is particularly useful when the assumption of normal distribution for differences is not met, as it only requires that the distribution of differences is symmetric around a central value. Its primary aim is to determine whether this central value significantly differs from zero.

The results of the Weighted Wilcoxon Signed-Rank test, summarized in Table 10, reveal whether the two decision paradigms (Final Decision Maker vs. Majority Vote) produce significantly different results for each dataset. In Strategy QA, Last Letter Concat, Social IQa, and Social Support the Final Decision Maker and Majority Vote approaches yield significantly different performance distributions. This implies that the choice of decision paradigm substantially affects model evaluation outcomes. No Significant Difference in Common Sense QA suggests that, for this dataset, either approach can be used interchangeably.

Given these findings, we proceed with pairwise comparisons of description frameworks within each dataset, focusing separately on the Final Decision Maker and Majority Vote approaches to determine whether certain frameworks consistently outperform others in specific dataset tasks.

C.1 Strategy QA Dataset

The p-value analysis in 10 indicates a very strong statistical difference between the Final Decision Maker and Majority Vote approaches when weighing by the number of samples tested. Thus, we start the pair-wise analysis of the different description frameworks using the Final Decision Maker. According to the results, every comparison is statistically significant, meaning that here the choice of the description framework matters.

We move on pair-wise analysis using the majority vote (Table 12). The table shows that nearly every pair comparison has a statistical significance, with p-values close to 0 for the vast majority of comparisons. As with the Final Decision Maker, each framework yields a unique accuracy distribution for the Majority Vote as well. The result is that description framework selection heavily impacts performance under the Majority Vote in this dataset, much as it does under the Final Decision Maker. In this dataset, both Final Decision Maker and Majority Vote results show that any two de-

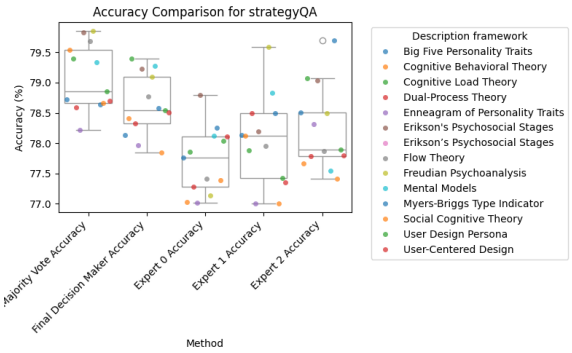

Figure 2: Box plot comparison Strategy QA dataset.

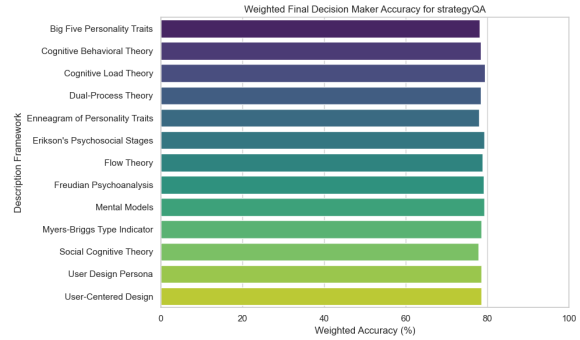

Figure 3: Weighted mean for the Strategy QA dataset.

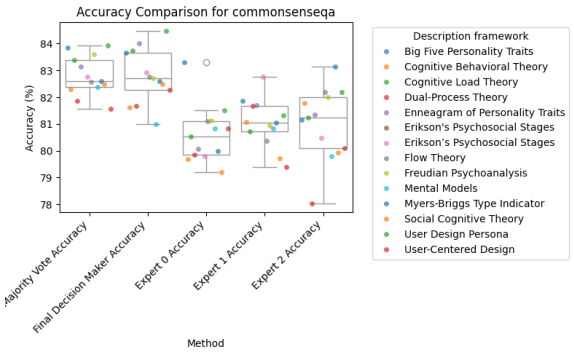

Figure 4: Box plot comparison Common Sense QA dataset.

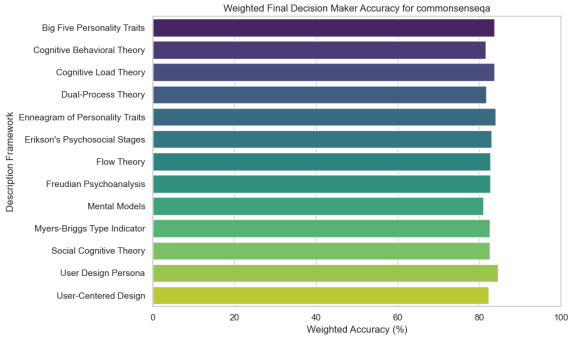

Figure 5: Weighted mean for the Common Sense QA dataset.

scription frameworks significantly differ. Practically, if you are optimizing for performance on this dataset, you cannot assume frameworks are interchangeable. This finding is also supported by looking at the weighted mean differences shown in Figure 3. Also, as shown in Figure 2, the individual performance of the agents is typically lower than the adoption of an ensemble method.

C.2 Common Sense QA Dataset

The results reported in Table 10 revealed no statistical difference between choosing the Final Decision Maker and the Majority Vote ensembling method. Even though certain individual frameworks might show small differences, the overall weighted distribution of accuracy differences does not deviate enough to conclude that one approach is systematically superior, both methods perform comparably. But, comparing the description frameworks against each other, revealed consistent differences in every pair. Weighted Mann–Whitney test, reported in Table 14 shows that even small raw differences in final accuracy become significant. In practice, the results imply that each description framework yields a distinct Final Decision Maker accuracy distribution.

Much like the Final Decision Maker results, the statistical comparison reported in Table 13 reveals that all frameworks differ significantly under the Weighted Mann–Whitney for the Majority Vote approach as well.

These findings are also supported by analyzing the weighted mean accuracies shown in Figure 5. Indeed, a graphical representation of the results obtained in this dataset is proposed in Figure 4. Adopting an ensemble method generally provides better performance w.r.t. the single expert agents, such as Big Five or Cognitive Load Theory.

C.3 Last Letter Concat Dataset

The analysis results reported in Table 10 revealed a p-value near zero for this dataset, indicating a clear difference between the Final Decision Maker and Majority Vote. Indeed, the Weighted Wilcoxon test strongly rejects the null hypothesis that the two methods produce the same distribution of accuracies. Both ensembling approaches achieve overall high accuracies (often above 90%), making them particularly interesting since the difference is consistent and robust even within a high-performing environment.

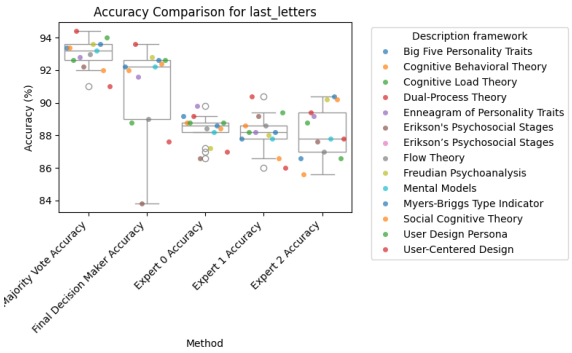

Figure 6: Boxplot Last Letter Concat dataset.

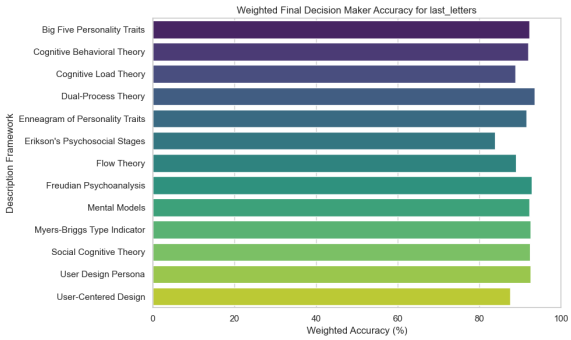

Figure 7: Weighted mean for the Last Letter Concat dataset.

Starting a deeper analysis of the Final Decision Maker, proposed in Table 15, we see that almost all pairs comparisons are statistically significant, with only a few exceptions (for example, Big Five Personality Traits and Mental Models, Myers-Briggs Type Indicator and User Design Persona). This indicates that certain frameworks effectively yield the same final accuracy distribution for the Final Decision Maker. The large majority of frameworks stand out from each other, but those scattered no statistically significant results show that for certain frameworks, the Weighted Mann–Whitney test sees no meaningful difference. Analyzing the statistically significant comparisons for the Majority Vote (Table 16) we note the same phenomenon. Indeed, almost every comparison is significant, but again, there are a small number of exceptions. For instance, Big Five Personality Traits and Cognitive Behavioral Theory, Freudian Psychoanalysis and Myers-Briggs Type Indicator are not significant. The overall pattern remains that most pairs differ significantly, implying that in Last Letters, the chosen framework heavily influences final accuracy with the Majority Vote approach. Interestingly, in this dataset, our approach typically has very high accuracies (90%+). Still, small differences are statistically significant.

The statistical findings are also supported by the analysis of the weighted mean accuracy plot shown in Figure 7. Furthermore, a visual analysis of the performance distribution presented in Figure 6 indicates a significant variability in scores for the Final Decision Maker, whereas the scores for the Majority Vote demonstrate a more compact distribution. Furthermore, the performance levels of the individual agents are considerably lower when compared to the two ensemble methods.

C.4 Social IQa Dataset

The results of the Weighted Wilcoxon test, shown in Table 10, indicate a highly significant difference between the Final Decision Maker and Majority Vote methods in this dataset, with a p-value effectively close to 0. This suggests that, from a weighted perspective, one of these methods consistently outperforms the other for the SocialIQa dataset. The findings strongly imply that the two methods are not interchangeable regarding their final accuracy outcomes.

The statistical analysis of the Final Decision Maker presented in Table 17 shows that all or nearly all pairs of frameworks are statistically relevant. For instance, both Dual-Process Theory and Flow Theory exhibit significant differences. This means that every pair of description frameworks demonstrates a statistically significant difference when evaluated using the Weighted Mann–Whitney test. As a result, even small raw differences can yield highly significant outcomes. Practically speaking, this suggests that we cannot treat any two descriptive frameworks as equivalent—each framework produces a distinct accuracy distribution for the Final Decision Maker from the perspective of the Weighted Mann–Whitney analysis.

The Majority Vote analysis, reported in Table 18, shows an extensive series of statistically valid comparisons, with very few exceptions (e.g. Dual-Process Theory and Flow Theory). As with Final Decision Maker, Majority Vote analysis demonstrates that almost all frameworks differ in Weighted Mann–Whitney.

Also, the weighted mean accuracy presented in Figure 9 supports the findings. Furthermore, the distribution of performance scores illustrated in Figure 8 highlights two significant observations. Firstly, the Majority Vote ensembling method

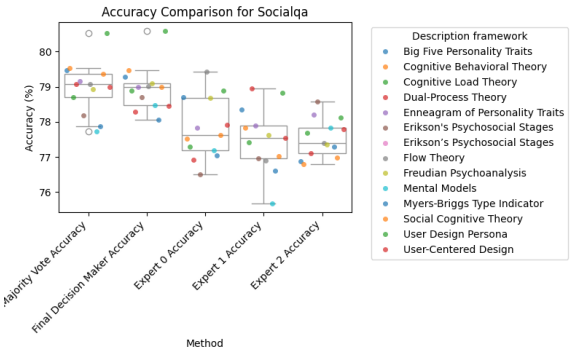

Figure 8: Boxplot Social IQa dataset.

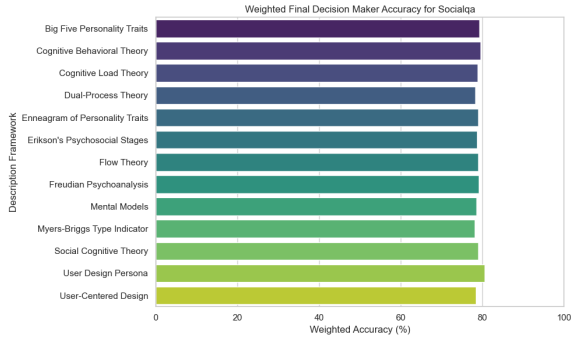

Figure 9: Weighted mean for the Social IQa dataset.

demonstrates a greater degree of variability in performance compared to the Final Decision Maker. Secondly, the Cognitive Load Theory description framework emerges as the highest scorer, whereas the Myers-Briggs Type Indicator description framework consistently ranks as the lowest performer across all assessments.

C.5 Social Support Dataset

In this dataset, all the description frameworks have fairly low overall accuracies (in the 40% range). The Weighted Wilcoxon test results presented in Table 10 yield a p-value near zero, implying a significant difference between the two methods once weighting (based on the number of samples tested) is accounted for. Even though the absolute performance might be low, the difference is systematically observed across the various description frameworks. In practice, this indicates that, in this dataset, one approach has a measurable advantage.

The analysis of the Final Decision Maker proposed in Table 19, demonstrates many statistically valid comparisons, with few exceptions. For instance, Big Five Personality Traits and Cognitive Load Theory are not statistically different. Although many comparisons are significant, certain frameworks produce statistically indistinguishable Weighted Mann–Whitney outcomes. Considering this dataset tends to have low accuracies (in the 40% range), the significance indicates that even small improvements are consistent. The large number of statistically valid comparisons confirms that the choice of framework typically matters under the Final Decision Maker approach. The comparative analysis of the Majority Vote (reported in Table 20) shows a similar trend of statistically significant comparisons for the Final Decision Maker (with few exceptions such as Cognitive Behavioral Theory and Enneagram of Personality Traits, or Cognitive Behavioral Theory and Flow Theory).

The permutation test (21) is particularly suitable for comparing macro F1-scores due to its non-parametric nature and flexibility in handling complex, aggregated metrics. Unlike parametric tests, permutation tests do not assume a specific distribution of the data, making them ideal for evaluating metrics like the macro F1-score, which is the unweighted average of F1-scores across all classes in a multi-class classification problem. In practice, the permutation test involves repeatedly shuffling the labels of the dataset and recalculating the performance metric to generate a distribution of scores under the null hypothesis—that there is no difference between the models. By comparing the observed difference in macro F1-scores to this null distribution, we can determine the empirical p-value, indicating the likelihood of observing such a difference by chance. This method is particularly advantageous when dealing with metrics like the macro F1-score, which can be sensitive to class imbalances and may not meet the assumptions required for parametric tests. By not relying on such assumptions, the permutation test provides a more reliable assessment of statistical significance in these contexts. Therefore, the permutation test is a robust and appropriate choice for comparing macro F1-scores, offering a non-parametric approach that accommodates the complexities inherent in multi-class classification evaluations.

Finally, the analysis of the performance score distributions presented in Figure 10 indicates that cognitive theory-based description frameworks consistently yield superior results. This finding underscores the importance of cognitive skills for agents in effectively addressing the task at hand. The same finding is supported by the visual analysis of the weighted mean of accuracy presented in Figure 11.

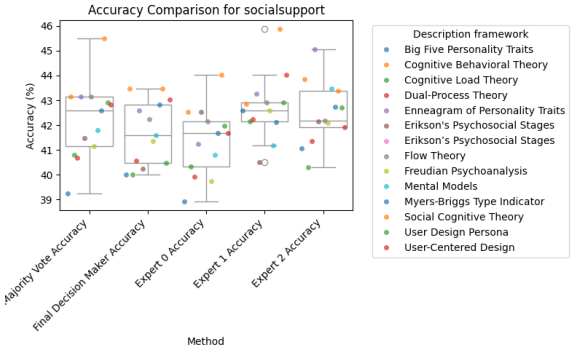

Figure 10: Boxplot Social Support dataset.

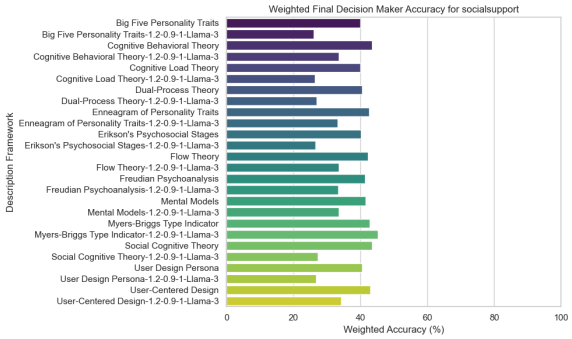

Figure 11: Weighted mean for the Social Support dataset.

D Agent Roles Analyses

We first tested whether expert roles differ statistically in their accuracy scores within each dataset by applying weighted analyses that account for the number of samples tested per framework. Specifically, we used a weighted ANOVA (also known as Welch's ANOVA)(WELCH, 1947) to compare the weighted means and a weighted Kruskal–Wallis test. We chose Welch's t-test, or unequal variances t-test since it is a two-sample test to test the (null) hypothesis that two populations have equal means. This statistic is more reliable when the two samples have unequal variances and possibly unequal sample sizes. This is the right choice in our cases since we cannot assume that the distributions analyzed have equal variances and they do not have equal sample sizes. Furthermore, we make the ANOVA findings more robust by also applying the Kruskal–Wallis test. The Kruskal–Wallis test (Kruskal and Wallis, 1952), or one-way ANOVA on ranks, is a non-parametric statistical test to test whether samples originate from the same distribution. However, a significant test indicates that at least one sample stochastically dominates one other sample without identifying where this

Figure 12: Expert roles distribution for the Strategy QA dataset.

stochastic dominance occurs or for how many pairs of groups of stochastic dominance obtains. Thus, when the weighted ANOVA yielded a significant overall effect ($p < 0.05$), we conducted Tukey's HSD post-hoc pairwise comparisons (Tukey, 1949) to determine which specific roles differ. Turkey's HSD post-hoc pairwise test is a single-step multiple comparison procedure and statistical test used to correctly interpret the statistical significance of the difference between means. The test, based on a studentized range distribution, compares all possible pairs of means in a distribution similar to the distribution of t from the t-test. The test compares the means of every treatment to the means of every other treatment; that is, it applies simultaneously to the set of all pairwise comparisons.

The overall weighted ANOVA and Kruskal–Wallis results are summarized in Table 22 and Table 23, respectively, while the detailed Tukey's HSD results are reported in Tables 25, 26, and 27. The distribution of expert roles within each dataset is presented in Table 24. Here, we discuss each dataset in isolation.

D.1 Strategy QA Dataset

The weighted ANOVA analysis reported in Table 22 indicates a statistically significant difference in the weighted mean accuracy among expert roles, while the weighted Kruskal–Wallis test (Table 23) did not detect differences. But, since the results of the multiple comparisons may inflate the Type I error rate [3] we proceed with deeper analyses.

Analyzing the distribution of roles, shown in Figure 12, we noted that the most adopted role is Artificial Intelligence, followed by Cognitive Psychology, and Cognitive Science. This indicates that

[3] Type I error rate is the probability of incorrectly rejecting a true null hypothesis (H_0). It is the likelihood of concluding that there is an effect or difference when none actually exists. This is often referred to as a *false positive* result.

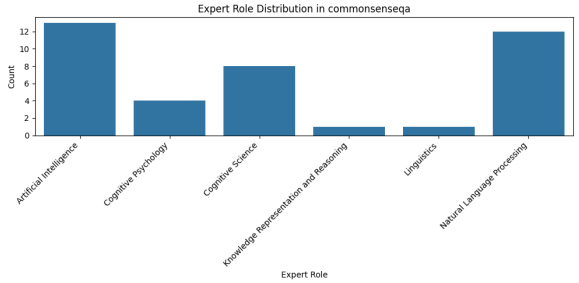

Figure 13: Expert roles distribution for the Common Sense QA dataset.

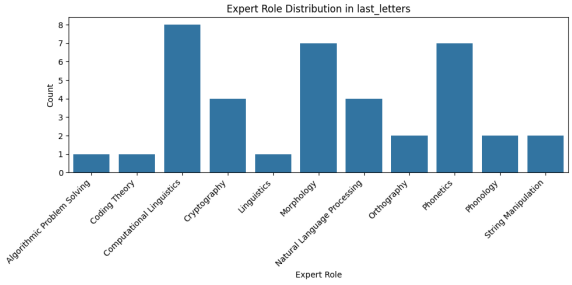

Figure 14: Expert roles distribution for the Last Letter Concat dataset.

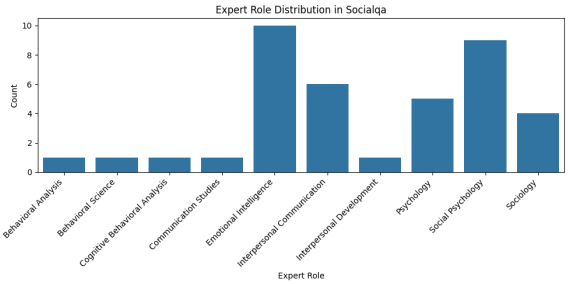

Figure 15: Expert roles distribution for the Social IQa dataset.

the Project Manager Agent thinks that Cognitive skills are important and so required to efficiently perform the task.

We performed the Tukey's HSD pairwise comparisons (Table 25) and we discover that :

- *Artificial Intelligence* significantly differs from *Cognitive Psychology*, *Logic and Formal Reasoning*, *Machine Learning*, and *Natural Language Processing*.
- *Cognitive Psychology* is significantly different from *Mathematical Logic*.
- *Cognitive Science* is significantly different from both *Logic and Formal Reasoning* and *Machine Learning*.
- *Formal Logic* differs significantly from *Logic and Formal Reasoning* and *Machine Learning*.
- *Machine Learning* is significantly different from *Mathematical Logic*, and *Mathematical Logic* differs significantly from *Natural Language Processing*.

In this dataset, the weighted ANOVA indicates that expert role selection significantly affects accuracy. The distribution of roles shown in Figure 12 together with Tukey's HSD test reveal that roles selection has an impact on performance. Indeed *Artificial Intelligence*, *Machine Learning*, *Logic and Formal Reasoning*, and *Mathematical Logic* exhibit statistically significant differences.

D.2 Common Sense QA Dataset

The analysis of the weighted ANOVA test and the Kruskal–Wallis test, reported respectively in Table 22 and Table 23, revealed very high p-values, indicating that role selection does not have a decisive impact on accuracy. Although the role distribution shown in Figure 13 exhibits variability, with a strong preference for *Artificial Intelligence* and *Natural Language Processing* roles, the differences in the selection are not statistically significant.

D.3 Last Letter Concat Dataset

The statistical results of the weighted ANOVA test and the Kruskal–Wallis test, reported respectively in Table 22 and Table 23, did not indicate a stastical valid difference in role selection. Similar to the Common Sense QA dataset.

The role distribution, shown in Figure 14, we discover a preference for *Computational Linguistics*, *Morphology*, *Phonetics* and *Cryptography* roles. However, this selection is not statistically valid. The uniform sample size and the non-significant test results imply that expert role selection does not affect accuracy in this dataset.

D.4 Social IQa Dataset

The statistical test results reported in Table 22 and Table 23 revealed that while the weighted ANOVA is highly significant, the weighted Kruskal–Wallis test is marginal. We further analyzed the possibility of a role selection effect by conducting Tukey's HSD analysis. The test results, reported in Table 26, indicate that only the *Behavioral Analysis* role is significantly different from *Behavioral Science*, *Cognitive Behavioral Analysis*, *Communica-*

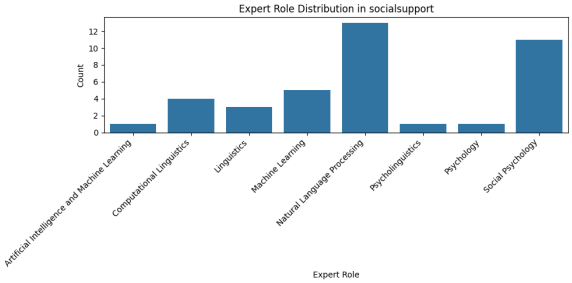

Figure 16: Expert roles distribution for the Social Support dataset.

tion Studies, *Emotional Intelligence*, *Interpersonal Communication*, *Interpersonal Development*, *Psychology*, *Social Psychology*, and *Sociology*.

We further analyzed the role distribution, as shown in Figure 15. The distribution demonstrates that the set of roles comprises strong task-dependent roles, including *Emotional Intelligence*, *Interpersonal Communication*, *Psychology*, *Social Psychology*, and *Sociology*.

Overall, the significant ANOVA and detailed Tukey's HSD pairwise comparisons indicate that, in the Social IQa dataset, expert role selection plays an important role. In particular, the fact that *Behavioral Analysis* is statistically distinct from several other roles suggests that its inclusion (or exclusion) could meaningfully affect system accuracy.

D.5 Social Support Dataset

The statistical tests (reported in Table 22 and Table 23) both indicate statistically significant values. We therefore proceed with a finer comparison using Tukey's HSD pairwise test. The results, reported in Table 27, identify only one significant difference—between *Natural Language Processing* and *Social Psychology*.

We also analyzed the role distribution shown in Figure 16, which reveals a strong preference for both *Natural Language Processing* and *Social Psychology* roles. To sum up, while the overall tests indicate that expert role selection affects accuracy, the limited number of significant pairwise differences suggests that the variations among roles are relatively modest.

D.6 Overall Analysis

Across the examined datasets, our analyses reveal a strong impact of expert role selection on task accuracy. In the Common Sense QA and Last Letter Concat datasets (see Sections D.2 and D.3), the weighted ANOVA and Kruskal–Wallis tests did not show statistically significant differences in role selection. However, an inspection of the role distributions indicates that each dataset exhibits distinct preferred roles. This pattern suggests that the Project Manager agent is able to select the appropriate expertise based on the specific task requirements, resulting in strong task-dependent role allocation despite overall non-significant differences.

In contrast, the analysis of the Social IQa dataset (subsection D.4), Strategy QA dataset (subsection D.1), and Social Support dataset (subsection D.5) indicate that expert role selection plays an important role. The weighted ANOVA yielded highly significant results, and Tukey's HSD pairwise comparisons reveal statistically valid role selection. This implies that the inclusion (or exclusion) of some roles can meaningfully affect system accuracy in these tasks.

Overall, these findings demonstrate that the influence of expert role selection is highly task-dependent. In some datasets, such as Common Sense QA and Last Letter Concat, the impact is relatively modest, whereas in others, like Social IQa, role selection plays a critical role in shaping performance. This underscores not only the importance of task-dependent role allocation strategies in multi-agent systems, but also the efficacy of the Project Manager agent in selecting the most important role for solving a text. The Project Manager agent exhibits the flexibility necessary to successfully select the appropriate roles for the expert agents to solve tasks.

E Agent Position Bias

We investigated whether the position of an agent (i.e., Expert 0, Expert 1, or Expert 2) influences performance. To conduct this evaluation, we aggregated the accuracy data from all datasets by weighting each observation by the number of samples tested. The aggregated counts yield the contingency table presented in Table 28. A chi-square test was then performed on this 3×2 contingency table. The null hypothesis H_0 states that: *the expert's position does not affect performance*, i.e., all three positions yield the same success-to-failure ratio. The test produced a chi-square statistic of $\chi^2 = 2.28$ with 2 degrees of freedom and a p-value of approximately 0.32. Since the p-value is greater than 0.05 we fail to reject the null hypothesis.

These results indicate that there is no statistically

significant positional bias in the performance of the agents. Although minor numerical differences exist among the aggregated successes of Experts 0, 1, and 2, these differences are well within the bounds of random variation given the sample sizes. Consequently, we conclude that the agents perform equivalently regardless of their position.

F Task Instructions and Contexts

In this section, we report the *task instructions* and the *context* for each dataset-task.

Note: We use the same task instructions and context for all the description frameworks of the same dataset.

As *context*, we generally use the one provided in the dataset, downloaded from https://github.com/google/BIG-bench/tree/main/bigbench/benchmark_tasks, if available.

F.1 Common Sense QA Dataset

task-instruction: Answer closed-ended questions by indicating the capital letter of the correct answer from the five alternatives provided
context: commonsense question answering
example of query: Question: "'Where would you find magazines along side many other printed works?"' A) doctor B) bookstore C) market D) train station E) mortuary

F.2 Last Letter Concat Dataset

task-instruction: Take the last letters of the words and concatenate them
context:linguistics puzzles
example of query: Take the last letters of each words in "Silvia Carolina Stan Chuck" and concatenate them.

F.3 Social IQa Dataset

task-instruction: Answer to multi choice questions
context: Questions are about the motivations, emotional reactions, and preceding and following events surrounding interpersonal situations
example of query: Answer the following question by reporting the capital letter of the correct answer among the alternatives: """Tracy didn't go home that evening and resisted Riley's attacks. What does Tracy need to do before this? """ A) Make a new plan B) Go home and see Riley C) Find somewhere to go

F.4 Strategy QA Dataset

task-instruction: Answer multiple-choice questions where the required reasoning steps are implicit in the question
context: The task aims to measure the ability of pretrained models on context-free question answering, multi-step, implicit reasoning, and logical reasoning
example of query: Answer the following question by reporting the capital letter of the correct answer among the alternatives: """Is it common to see frost during some college commencements?""" A) Yes B) No

F.5 Social Support Dataset

task-instruction: Classify whether a reply is unsupportive, neutral, or supportive
context: The task aims to measure the ability of pretrained models on understanding supportive (and unsupportive) language uses for social support classification
example of query: Answer the following question by reporting the capital letter of the correct answer among the alternatives: """While I don't have any source offhand, there are authentic Jewish sources that mentions the idea that 7 represents nature (7 days of creation) and 8 is supernatural. It has nothing to do with the shape of the number.""" Q: Is the following reply unsupportive, neutral, or supportive? A) unsupportive B) neutral C) supportive

G Implementation Details

We implemented our approach by adopting the Meta Llama 3.1-70B Instruct large language model (LLM). The inference process is guided by a temperature setting of 1.2, which introduces a moderate degree of randomness in the model's responses, allowing for more diverse and creative outputs. Additionally, nucleus sampling is set at 0.9, ensuring that the model considers only the most relevant and probable tokens while generating text, striking a balance between diversity and coherence. Finally, in this first set of experiments, we do not focus on obtaining the highest probability answer from the LLM. Instead, we configure the system to generate a single alternative response for each input, ensuring a deterministic output. We ensure reproducibility by setting a fixed seed value for random number generation directly within our implementation files.

Additionally, we leverage the default *chat-*

template provided by Llama 3.1 to structure interactions between the system and the user. This template explicitly defines the roles of both the system and the user, ensuring that the model correctly interprets input prompts and maintains consistency in its responses. By adhering to this standardized template, we enhance the model's ability to generate contextually appropriate and structured outputs, improving overall response coherence and alignment with the intended person assigned to the agent.

G.1 Computational Cost Estimation

We conduct our experiments using the CINECA HPC infrastructure. This section provides an estimation of the required GPU hours.

For each of the five datasets, we run 13 description frameworks. Each experiment is executed on 3 H100 GPUs. Typically, processing 500 experimental samples requires approximately 24 hours. Below, we report the number of samples tested for each dataset and the corresponding estimated computational cost.

- **Strategy QA**: More than 1600 samples, but we consider 1500 for computational cost estimation. Since we process 500 samples per day, the experiments require 3 days. The computational cost is:

$$\textit{GPU hours cost} = 3 \times 13 \times 24 \times 3 = 2808$$

- **Common Sense QA**: Approximately 500 samples. As we process 500 samples per day, the experiments require 1 day. The computational cost is:

$$\textit{GPU hours cost} = 3 \times 13 \times 24 \times 1 = 936$$

- **Last Letter Concat**: 500 samples. Since we process 500 samples per day, the experiments require 1 day. The computational cost is:

$$\textit{GPU hours cost} = 3 \times 13 \times 24 \times 1 = 936$$

- **Social IQa**: More than 1700 samples, but we consider 1500 for computational cost estimation. Since we process 500 samples per day, the experiments require 3 days. The computational cost is:

$$\textit{GPU hours cost} = 3 \times 13 \times 24 \times 3 = 2808$$

- **Social Support**: 897 samples, rounded to 1000 for computational cost estimation. Since we process 500 samples per day, the experiments require 2 days. The computational cost is:

$$\textit{GPU hours cost} = 3 \times 13 \times 24 \times 2 = 1872$$

Thus, the total estimated computational cost is:

$$\textit{Total GPU hours cost} = 2808 + 936 + 936 + 2808 + 1872 = 9360$$

G.2 System Usage

Each parameter in the JSON file plays a specific role in guiding the system's behavior. Below we report a detailed explanation of each parameter that must be included in the JSON configuration file.

- **name**: The name of the experiment.
- **task**: the task instruction to provide to the experts.
- **context**: The broader domain of the task, indicating that it belongs to, e.g., commonsense question answering.
- **description_framework**: The theoretical description framework guiding the generation of the agent description profiles.
- **model_name**: The specific language model used, in this case, Meta Llama 3.1-70B Instruct.
- **output_dir**: The directory where the generated results are stored.
- **input**: The input file containing queries, with each query on a separate line.
- **temperature**: Controls response randomness. We adopt a value of 1.2 to introduce moderate variability, allowing for diverse outputs.
- **nucleus**: Defines the nucleus sampling threshold. We adopt a value of 0.9 ensures that the model considers only the most relevant tokens during generation.
- **alternatives**: The number of alternative outputs generated per query. Here, it is set to 1 to ensure a deterministic response.
- **resume**: A flag indicating whether to resume execution from a previous run.

- **cache_dir**: Specifies the directory for caching model weights and computations. This parameter avoids downloading the same model multiple times.
- **max_experts_number**: The maximum number of expert models used during inference. In our experiments, we fix this parameter to 3.
- **token**: The authentication token required to access the model on the HuggingFace repository.

This configuration setup allows users to customize the system's behavior while ensuring reproducibility across different experimental settings.

To execute our system, we utilize a SLURM job script to manage resource allocation and scheduling efficiently. Below, we provide an example SLURM configuration file:

```
#!/bin/bash
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=16
#SBATCH --gpus=3
#SBATCH --mem=100GB
#SBATCH --time="23:59:59"
#SBATCH --partition="cluster-partition"
#SBATCH --qos=normal
#SBATCH --job-name="job-name"
#SBATCH --output="std out"
#SBATCH --error="std err"

echo "Job starting"

source /myvenv/bin/activate

python ../PoE_Small/config_file.py
--config-file configuration.json
```

After specifying the SLURM directives, the script activates a virtual environment (‘source /myvenv/bin/activate‘) and executes the system using a Python script with a predefined JSON configuration file. This setup ensures efficient resource utilization and facilitates job execution on a high-performance computing cluster (e.g., CINECA pre-exascale HPC).

Weighted Mann–Whitney for Final Decision Maker, Dataset: strategyQA				
Framework 1	Framework 2	MW-stat	p-value	Sig?
Big Five Personality Traits	Cognitive Behavioral Theory	0.0000	0	Yes
Big Five Personality Traits	Cognitive Load Theory	0.0000	0	Yes
Big Five Personality Traits	Dual-Process Theory	0.0000	0	Yes
Big Five Personality Traits	Enneagram of Personality Traits	2485080.0000	0	Yes
Big Five Personality Traits	Erikson's Psychosocial Stages	0.0000	0	Yes
Big Five Personality Traits	Flow Theory	0.0000	0	Yes
Big Five Personality Traits	Freudian Psychoanalysis	0.0000	0	Yes
Big Five Personality Traits	Mental Models	0.0000	0	Yes
Big Five Personality Traits	Myers-Briggs Type Indicator	0.0000	0	Yes
Big Five Personality Traits	Social Cognitive Theory	2514720.0000	0	Yes
Big Five Personality Traits	User Design Persona	0.0000	0	Yes
Big Five Personality Traits	User-Centered Design	0.0000	0	Yes
Cognitive Behavioral Theory	Cognitive Load Theory	0.0000	0	Yes
Cognitive Behavioral Theory	Dual-Process Theory	2470508.0000	0	Yes
Cognitive Behavioral Theory	Enneagram of Personality Traits	2545614.0000	0	Yes
Cognitive Behavioral Theory	Erikson's Psychosocial Stages	0.0000	0	Yes
Cognitive Behavioral Theory	Flow Theory	0.0000	0	Yes
Cognitive Behavioral Theory	Freudian Psychoanalysis	0.0000	0	Yes
Cognitive Behavioral Theory	Mental Models	0.0000	0	Yes
Cognitive Behavioral Theory	Myers-Briggs Type Indicator	0.0000	0	Yes
Cognitive Behavioral Theory	Social Cognitive Theory	2575976.0000	0	Yes
Cognitive Behavioral Theory	User Design Persona	0.0000	0	Yes
Cognitive Behavioral Theory	User-Centered Design	0.0000	0	Yes
Cognitive Load Theory	Dual-Process Theory	2408668.0000	0	Yes
Cognitive Load Theory	Enneagram of Personality Traits	2481894.0000	0	Yes
Cognitive Load Theory	Erikson's Psychosocial Stages	2595628.0000	0	Yes
Cognitive Load Theory	Flow Theory	2553562.0000	0	Yes
Cognitive Load Theory	Freudian Psychoanalysis	2497474.0000	0	Yes
Cognitive Load Theory	Mental Models	2458524.0000	0	Yes
Cognitive Load Theory	Myers-Briggs Type Indicator	2494358.0000	0	Yes
Cognitive Load Theory	Social Cognitive Theory	2511496.0000	0	Yes
Cognitive Load Theory	User Design Persona	2541098.0000	0	Yes
Cognitive Load Theory	User-Centered Design	2537982.0000	0	Yes
Dual-Process Theory	Enneagram of Personality Traits	2462778.0000	0	Yes
Dual-Process Theory	Erikson's Psychosocial Stages	0.0000	0	Yes
Dual-Process Theory	Flow Theory	0.0000	0	Yes
Dual-Process Theory	Freudian Psychoanalysis	0.0000	0	Yes
Dual-Process Theory	Mental Models	0.0000	0	Yes
Dual-Process Theory	Myers-Briggs Type Indicator	0.0000	0	Yes
Dual-Process Theory	Social Cognitive Theory	2492152.0000	0	Yes
Dual-Process Theory	User Design Persona	0.0000	0	Yes
Dual-Process Theory	User-Centered Design	0.0000	0	Yes
Enneagram of Personality Traits	Erikson's Psychosocial Stages	0.0000	0	Yes
Enneagram of Personality Traits	Flow Theory	0.0000	0	Yes
Enneagram of Personality Traits	Freudian Psychoanalysis	0.0000	0	Yes
Enneagram of Personality Traits	Mental Models	0.0000	0	Yes
Enneagram of Personality Traits	Myers-Briggs Type Indicator	0.0000	0	Yes
Enneagram of Personality Traits	Social Cognitive Theory	2567916.0000	0	Yes
Enneagram of Personality Traits	User Design Persona	0.0000	0	Yes
Enneagram of Personality Traits	User-Centered Design	0.0000	0	Yes
Erikson's Psychosocial Stages	Flow Theory	2730574.0000	0	Yes
Erikson's Psychosocial Stages	Freudian Psychoanalysis	2670598.0000	0	Yes
Erikson's Psychosocial Stages	Mental Models	0.0000	0	Yes
Erikson's Psychosocial Stages	Myers-Briggs Type Indicator	2667266.0000	0	Yes
Erikson's Psychosocial Stages	Social Cognitive Theory	2685592.0000	0	Yes
Erikson's Psychosocial Stages	User Design Persona	2717246.0000	0	Yes
Erikson's Psychosocial Stages	User-Centered Design	2713914.0000	0	Yes
Flow Theory	Freudian Psychoanalysis	0.0000	0	Yes
Flow Theory	Mental Models	0.0000	0	Yes
Flow Theory	Myers-Briggs Type Indicator	2624039.0000	0	Yes
Flow Theory	Social Cognitive Theory	2642068.0000	0	Yes
Flow Theory	User Design Persona	2673209.0000	0	Yes
Flow Theory	User-Centered Design	2669931.0000	0	Yes
Freudian Psychoanalysis	Mental Models	0.0000	0	Yes
Freudian Psychoanalysis	Myers-Briggs Type Indicator	2566403.0000	0	Yes
Freudian Psychoanalysis	Social Cognitive Theory	2584036.0000	0	Yes
Freudian Psychoanalysis	User Design Persona	2614493.0000	0	Yes
Freudian Psychoanalysis	User-Centered Design	2611287.0000	0	Yes
Mental Models	Myers-Briggs Type Indicator	2526378.0000	0	Yes
Mental Models	Social Cognitive Theory	2543736.0000	0	Yes
Mental Models	User Design Persona	2573718.0000	0	Yes
Mental Models	User-Centered Design	2570562.0000	0	Yes
Myers-Briggs Type Indicator	Social Cognitive Theory	2580812.0000	0	Yes
Myers-Briggs Type Indicator	User Design Persona	2611231.0000	0	Yes
Myers-Briggs Type Indicator	User-Centered Design	2608029.0000	0	Yes
Social Cognitive Theory	User Design Persona	0.0000	0	Yes
Social Cognitive Theory	User-Centered Design	0.0000	0	Yes
User Design Persona	User-Centered Design	2656899.0000	0	Yes

Table 11: Pairwise Weighted Mann–Whitney comparisons of description frameworks for Strategy QA (Final Decision Maker).

Weighted Mann–Whitney for Majority Vote, Dataset: strategyQA				
Framework 1	Framework 2	MW-stat	p-value	Sig?
Big Five Personality Traits	Cognitive Behavioral Theory	0.0000	0	Yes
Big Five Personality Traits	Cognitive Load Theory	0.0000	0	Yes
Big Five Personality Traits	Dual-Process Theory	2411760.0000	0	Yes
Big Five Personality Traits	Enneagram of Personality Traits	2485080.0000	0	Yes
Big Five Personality Traits	Erikson's Psychosocial Stages	0.0000	0	Yes
Big Five Personality Traits	Flow Theory	0.0000	0	Yes
Big Five Personality Traits	Freudian Psychoanalysis	0.0000	0	Yes
Big Five Personality Traits	Mental Models	0.0000	0	Yes
Big Five Personality Traits	Myers-Briggs Type Indicator	2497560.0000	0	Yes
Big Five Personality Traits	Social Cognitive Theory	2514720.0000	0	Yes
Big Five Personality Traits	User Design Persona	0.0000	0	Yes
Big Five Personality Traits	User-Centered Design	2541240.0000	0	Yes
Cognitive Behavioral Theory	Cognitive Load Theory	2489684.0000	0	Yes
Cognitive Behavioral Theory	Dual-Process Theory	2470508.0000	0	Yes
Cognitive Behavioral Theory	Enneagram of Personality Traits	2545614.0000	0	Yes
Cognitive Behavioral Theory	Erikson's Psychosocial Stages	0.0000	0	Yes
Cognitive Behavioral Theory	Flow Theory	0.0000	0	Yes
Cognitive Behavioral Theory	Freudian Psychoanalysis	0.0000	0	Yes
Cognitive Behavioral Theory	Mental Models	2521644.0000	0	Yes
Cognitive Behavioral Theory	Myers-Briggs Type Indicator	2558398.0000	0	Yes
Cognitive Behavioral Theory	Social Cognitive Theory	2575976.0000	0	Yes
Cognitive Behavioral Theory	User Design Persona	2606338.0000	0	Yes
Cognitive Behavioral Theory	User-Centered Design	2603142.0000	0	Yes
Cognitive Load Theory	Dual-Process Theory	2408668.0000	0	Yes
Cognitive Load Theory	Enneagram of Personality Traits	2481894.0000	0	Yes
Cognitive Load Theory	Erikson's Psychosocial Stages	0.0000	0	Yes
Cognitive Load Theory	Flow Theory	0.0000	0	Yes
Cognitive Load Theory	Freudian Psychoanalysis	0.0000	0	Yes
Cognitive Load Theory	Mental Models	2458524.0000	0	Yes
Cognitive Load Theory	Myers-Briggs Type Indicator	2494358.0000	0	Yes
Cognitive Load Theory	Social Cognitive Theory	2511496.0000	0	Yes
Cognitive Load Theory	User Design Persona	2541098.0000	0	Yes
Cognitive Load Theory	User-Centered Design	2537982.0000	0	Yes
Dual-Process Theory	Enneagram of Personality Traits	2462778.0000	0	Yes
Dual-Process Theory	Erikson's Psychosocial Stages	0.0000	0	Yes
Dual-Process Theory	Flow Theory	0.0000	0	Yes
Dual-Process Theory	Freudian Psychoanalysis	0.0000	0	Yes
Dual-Process Theory	Mental Models	0.0000	0	Yes
Dual-Process Theory	Myers-Briggs Type Indicator	0.0000	0	Yes
Dual-Process Theory	Social Cognitive Theory	0.0000	0	Yes
Dual-Process Theory	User Design Persona	0.0000	0	Yes
Dual-Process Theory	User-Centered Design	0.0000	0	Yes
Enneagram of Personality Traits	Erikson's Psychosocial Stages	0.0000	0	Yes
Enneagram of Personality Traits	Flow Theory	0.0000	0	Yes
Enneagram of Personality Traits	Freudian Psychoanalysis	0.0000	0	Yes
Enneagram of Personality Traits	Mental Models	0.0000	0	Yes
Enneagram of Personality Traits	Myers-Briggs Type Indicator	0.0000	0	Yes
Enneagram of Personality Traits	Social Cognitive Theory	0.0000	0	Yes
Enneagram of Personality Traits	User Design Persona	0.0000	0	Yes
Enneagram of Personality Traits	User-Centered Design	0.0000	0	Yes
Erikson's Psychosocial Stages	Flow Theory	2730574.0000	0	Yes
Erikson's Psychosocial Stages	Freudian Psychoanalysis	0.0000	0	Yes
Erikson's Psychosocial Stages	Mental Models	2628948.0000	0	Yes
Erikson's Psychosocial Stages	Myers-Briggs Type Indicator	2667266.0000	0	Yes
Erikson's Psychosocial Stages	Social Cognitive Theory	2685592.0000	0	Yes
Erikson's Psychosocial Stages	User Design Persona	2717246.0000	0	Yes
Erikson's Psychosocial Stages	User-Centered Design	2713914.0000	0	Yes
Flow Theory	Freudian Psychoanalysis	0.0000	0	Yes
Flow Theory	Mental Models	2586342.0000	0	Yes
Flow Theory	Myers-Briggs Type Indicator	2624039.0000	0	Yes
Flow Theory	Social Cognitive Theory	2642068.0000	0	Yes
Flow Theory	User Design Persona	2673209.0000	0	Yes
Flow Theory	User-Centered Design	2669931.0000	0	Yes
Freudian Psychoanalysis	Mental Models	2529534.0000	0	Yes
Freudian Psychoanalysis	Myers-Briggs Type Indicator	2566403.0000	0	Yes
Freudian Psychoanalysis	Social Cognitive Theory	2584036.0000	0	Yes
Freudian Psychoanalysis	User Design Persona	2614493.0000	0	Yes
Freudian Psychoanalysis	User-Centered Design	2611287.0000	0	Yes
Mental Models	Myers-Briggs Type Indicator	2526378.0000	0	Yes
Mental Models	Social Cognitive Theory	2543736.0000	0	Yes
Mental Models	User Design Persona	2573718.0000	0	Yes
Mental Models	User-Centered Design	2570562.0000	0	Yes
Myers-Briggs Type Indicator	Social Cognitive Theory	0.0000	0	Yes
Myers-Briggs Type Indicator	User Design Persona	0.0000	0	Yes
Myers-Briggs Type Indicator	User-Centered Design	0.0000	0	Yes
Social Cognitive Theory	User Design Persona	0.0000	0	Yes
Social Cognitive Theory	User-Centered Design	0.0000	0	Yes
User Design Persona	User-Centered Design	2656899.0000	0	Yes

Table 12: Pairwise Weighted Mann–Whitney comparisons of description frameworks for Strategy QA (Majority Vote).

Weighted Mann–Whitney for Majority Vote, Dataset: Common Sense QA				
Framework 1	Framework 2	MW-stat	p-value	Sig?
Big Five Personality Traits	Cognitive Behavioral Theory	318047.0000	4.48945e-247	Yes
Big Five Personality Traits	Cognitive Load Theory	311920.0000	1.10396e-244	Yes
Big Five Personality Traits	Dual-Process Theory	304122.0000	1.21835e-241	Yes
Big Five Personality Traits	Enneagram of Personality Traits	313591.0000	2.45995e-245	Yes
Big Five Personality Traits	Erikson's Psychosocial Stages	319718.0000	1.00039e-247	Yes
Big Five Personality Traits	Flow Theory	303565.0000	2.00963e-241	Yes
Big Five Personality Traits	Freudian Psychoanalysis	315819.0000	3.32321e-246	Yes
Big Five Personality Traits	Mental Models	322503.0000	8.19354e-249	Yes
Big Five Personality Traits	Myers-Briggs Type Indicator	320275.0000	6.065e-248	Yes
Big Five Personality Traits	Social Cognitive Theory	305236.0000	4.47797e-242	Yes
Big Five Personality Traits	User Design Persona	0.0000	1.65011e-247	Yes
Big Five Personality Traits	User-Centered Design	308021.0000	3.6674e-243	Yes
Cognitive Behavioral Theory	Cognitive Load Theory	0.0000	1.00039e-247	Yes
Cognitive Behavioral Theory	Dual-Process Theory	311766.0000	1.10396e-244	Yes
Cognitive Behavioral Theory	Enneagram of Personality Traits	0.0000	2.22921e-248	Yes
Cognitive Behavioral Theory	Erikson's Psychosocial Stages	0.0000	9.06613e-251	Yes
Cognitive Behavioral Theory	Flow Theory	0.0000	1.82095e-244	Yes
Cognitive Behavioral Theory	Freudian Psychoanalysis	0.0000	3.01157e-249	Yes
Cognitive Behavioral Theory	Mental Models	0.0000	7.42564e-252	Yes
Cognitive Behavioral Theory	Myers-Briggs Type Indicator	0.0000	5.49647e-251	Yes
Cognitive Behavioral Theory	Social Cognitive Theory	0.0000	4.0576e-245	Yes
Cognitive Behavioral Theory	User Design Persona	0.0000	1.49541e-250	Yes
Cognitive Behavioral Theory	User-Centered Design	315763.0000	3.32322e-246	Yes
Cognitive Load Theory	Dual-Process Theory	305760.0000	2.71479e-242	Yes
Cognitive Load Theory	Enneagram of Personality Traits	315280.0000	5.48151e-246	Yes
Cognitive Load Theory	Erikson's Psychosocial Stages	321440.0000	2.22921e-248	Yes
Cognitive Load Theory	Flow Theory	305200.0000	4.47797e-242	Yes
Cognitive Load Theory	Freudian Psychoanalysis	0.0000	7.40516e-247	Yes
Cognitive Load Theory	Mental Models	324240.0000	1.8258e-249	Yes
Cognitive Load Theory	Myers-Briggs Type Indicator	322000.0000	1.35148e-248	Yes
Cognitive Load Theory	Social Cognitive Theory	306880.0000	9.97808e-243	Yes
Cognitive Load Theory	User Design Persona	0.0000	3.67698e-248	Yes
Cognitive Load Theory	User-Centered Design	309680.0000	8.17197e-244	Yes
Dual-Process Theory	Enneagram of Personality Traits	0.0000	6.04927e-243	Yes
Dual-Process Theory	Erikson's Psychosocial Stages	0.0000	2.45996e-245	Yes
Dual-Process Theory	Flow Theory	0.0000	4.9423e-239	Yes
Dual-Process Theory	Freudian Psychoanalysis	0.0000	8.17198e-244	Yes
Dual-Process Theory	Mental Models	0.0000	2.01474e-246	Yes
Dual-Process Theory	Myers-Briggs Type Indicator	0.0000	1.49137e-245	Yes
Dual-Process Theory	Social Cognitive Theory	0.0000	1.10125e-239	Yes
Dual-Process Theory	User Design Persona	0.0000	4.0576e-245	Yes
Dual-Process Theory	User-Centered Design	301938.0000	9.01892e-241	Yes
Enneagram of Personality Traits	Erikson's Psychosocial Stages	323162.0000	4.96743e-249	Yes
Enneagram of Personality Traits	Flow Theory	306835.0000	9.97809e-243	Yes
Enneagram of Personality Traits	Freudian Psychoanalysis	0.0000	1.6501e-247	Yes
Enneagram of Personality Traits	Mental Models	325977.0000	4.06852e-250	Yes
Enneagram of Personality Traits	Myers-Briggs Type Indicator	323725.0000	3.01157e-249	Yes
Enneagram of Personality Traits	Social Cognitive Theory	308524.0000	2.22338e-243	Yes
Enneagram of Personality Traits	User Design Persona	0.0000	8.19353e-249	Yes
Enneagram of Personality Traits	User-Centered Design	311339.0000	1.82095e-244	Yes
Erikson's Psychosocial Stages	Flow Theory	312830.0000	4.0576e-245	Yes
Erikson's Psychosocial Stages	Freudian Psychoanalysis	0.0000	6.71082e-250	Yes
Erikson's Psychosocial Stages	Mental Models	332346.0000	1.65471e-252	Yes
Erikson's Psychosocial Stages	Myers-Briggs Type Indicator	330050.0000	1.22482e-251	Yes
Erikson's Psychosocial Stages	Social Cognitive Theory	314552.0000	9.04155e-246	Yes
Erikson's Psychosocial Stages	User Design Persona	0.0000	3.33231e-251	Yes
Erikson's Psychosocial Stages	User-Centered Design	317422.0000	7.40516e-247	Yes
Flow Theory	Freudian Psychoanalysis	0.0000	1.34794e-243	Yes
Flow Theory	Mental Models	315555.0000	3.32322e-246	Yes
Flow Theory	Myers-Briggs Type Indicator	0.0000	2.45996e-245	Yes
Flow Theory	Social Cognitive Theory	298660.0000	1.81649e-239	Yes
Flow Theory	User Design Persona	0.0000	6.69287e-245	Yes
Flow Theory	User-Centered Design	301385.0000	1.48765e-240	Yes
Freudian Psychoanalysis	Mental Models	328293.0000	5.49647e-251	Yes
Freudian Psychoanalysis	Myers-Briggs Type Indicator	326025.0000	4.06852e-250	Yes
Freudian Psychoanalysis	Social Cognitive Theory	310716.0000	3.00359e-244	Yes
Freudian Psychoanalysis	User Design Persona	0.0000	1.10691e-249	Yes
Freudian Psychoanalysis	User-Centered Design	313551.0000	2.45995e-245	Yes
Mental Models	Myers-Briggs Type Indicator	0.0000	1.0032e-252	Yes
Mental Models	Social Cognitive Theory	0.0000	7.40517e-247	Yes
Mental Models	User Design Persona	0.0000	2.72935e-252	Yes
Mental Models	User-Centered Design	320187.0000	6.06501e-248	Yes
Myers-Briggs Type Indicator	Social Cognitive Theory	315100.0000	5.48152e-246	Yes
Myers-Briggs Type Indicator	User Design Persona	0.0000	2.02027e-251	Yes
Myers-Briggs Type Indicator	User-Centered Design	317975.0000	4.48946e-247	Yes
Social Cognitive Theory	User Design Persona	0.0000	1.49137e-245	Yes
Social Cognitive Theory	User-Centered Design	303044.0000	3.31484e-241	Yes
User Design Persona	User-Centered Design	316869.0000	1.22145e-246	Yes

Table 13: Pairwise Weighted Mann–Whitney comparisons of description frameworks for Common Sense QA (Majority Vote).

Weighted Mann–Whitney for Final Decision Maker, Dataset: Common Sense QA				
Framework 1	Framework 2	MW-stat	p-value	Sig?
Big Five Personality Traits	Cognitive Behavioral Theory	318047.0000	4.48945e-247	Yes
Big Five Personality Traits	Cognitive Load Theory	0.0000	1.10396e-244	Yes
Big Five Personality Traits	Dual-Process Theory	304122.0000	1.21835e-241	Yes
Big Five Personality Traits	Enneagram of Personality Traits	0.0000	2.45995e-245	Yes
Big Five Personality Traits	Erikson's Psychosocial Stages	319718.0000	1.00039e-247	Yes
Big Five Personality Traits	Flow Theory	303565.0000	2.00963e-241	Yes
Big Five Personality Traits	Freudian Psychoanalysis	315819.0000	3.32321e-246	Yes
Big Five Personality Traits	Mental Models	322503.0000	8.19354e-249	Yes
Big Five Personality Traits	Myers-Briggs Type Indicator	320275.0000	6.065e-248	Yes
Big Five Personality Traits	Social Cognitive Theory	305236.0000	4.47797e-242	Yes
Big Five Personality Traits	User Design Persona	0.0000	1.65011e-247	Yes
Big Five Personality Traits	User-Centered Design	308021.0000	3.6674e-243	Yes
Cognitive Behavioral Theory	Cognitive Load Theory	0.0000	1.00039e-247	Yes
Cognitive Behavioral Theory	Dual-Process Theory	0.0000	1.10396e-244	Yes
Cognitive Behavioral Theory	Enneagram of Personality Traits	0.0000	2.22921e-248	Yes
Cognitive Behavioral Theory	Erikson's Psychosocial Stages	0.0000	9.06613e-251	Yes
Cognitive Behavioral Theory	Flow Theory	0.0000	1.82095e-244	Yes
Cognitive Behavioral Theory	Freudian Psychoanalysis	0.0000	3.01157e-249	Yes
Cognitive Behavioral Theory	Mental Models	330609.0000	7.42564e-252	Yes
Cognitive Behavioral Theory	Myers-Briggs Type Indicator	0.0000	5.49647e-251	Yes
Cognitive Behavioral Theory	Social Cognitive Theory	0.0000	4.0576e-245	Yes
Cognitive Behavioral Theory	User Design Persona	0.0000	1.49541e-250	Yes
Cognitive Behavioral Theory	User-Centered Design	0.0000	3.32322e-246	Yes
Cognitive Load Theory	Dual-Process Theory	305760.0000	2.71479e-242	Yes
Cognitive Load Theory	Enneagram of Personality Traits	0.0000	5.48151e-246	Yes
Cognitive Load Theory	Erikson's Psychosocial Stages	321440.0000	2.22921e-248	Yes
Cognitive Load Theory	Flow Theory	305200.0000	4.47797e-242	Yes
Cognitive Load Theory	Freudian Psychoanalysis	317520.0000	7.40516e-247	Yes
Cognitive Load Theory	Mental Models	324240.0000	1.8258e-249	Yes
Cognitive Load Theory	Myers-Briggs Type Indicator	322000.0000	1.35148e-248	Yes
Cognitive Load Theory	Social Cognitive Theory	306880.0000	9.97808e-243	Yes
Cognitive Load Theory	User Design Persona	0.0000	3.67698e-248	Yes
Cognitive Load Theory	User-Centered Design	309680.0000	8.17197e-244	Yes
Dual-Process Theory	Enneagram of Personality Traits	0.0000	6.04927e-243	Yes
Dual-Process Theory	Erikson's Psychosocial Stages	0.0000	2.45996e-245	Yes
Dual-Process Theory	Flow Theory	0.0000	4.9423e-239	Yes
Dual-Process Theory	Freudian Psychoanalysis	0.0000	8.17198e-244	Yes
Dual-Process Theory	Mental Models	316134.0000	2.01474e-246	Yes
Dual-Process Theory	Myers-Briggs Type Indicator	0.0000	1.49137e-245	Yes
Dual-Process Theory	Social Cognitive Theory	0.0000	1.10125e-239	Yes
Dual-Process Theory	User Design Persona	0.0000	4.0576e-245	Yes
Dual-Process Theory	User-Centered Design	0.0000	9.01892e-241	Yes
Enneagram of Personality Traits	Erikson's Psychosocial Stages	323162.0000	4.96743e-249	Yes
Enneagram of Personality Traits	Flow Theory	306835.0000	9.97809e-243	Yes
Enneagram of Personality Traits	Freudian Psychoanalysis	319221.0000	1.6501e-247	Yes
Enneagram of Personality Traits	Mental Models	325977.0000	4.06852e-250	Yes
Enneagram of Personality Traits	Myers-Briggs Type Indicator	323725.0000	3.01157e-249	Yes
Enneagram of Personality Traits	Social Cognitive Theory	308524.0000	2.22338e-243	Yes
Enneagram of Personality Traits	User Design Persona	0.0000	8.19353e-249	Yes
Enneagram of Personality Traits	User-Centered Design	311339.0000	1.82095e-244	Yes
Erikson's Psychosocial Stages	Flow Theory	312830.0000	4.0576e-245	Yes
Erikson's Psychosocial Stages	Freudian Psychoanalysis	325458.0000	6.71082e-250	Yes
Erikson's Psychosocial Stages	Mental Models	332346.0000	1.65471e-252	Yes
Erikson's Psychosocial Stages	Myers-Briggs Type Indicator	330050.0000	1.22482e-251	Yes
Erikson's Psychosocial Stages	Social Cognitive Theory	314552.0000	9.04155e-246	Yes
Erikson's Psychosocial Stages	User Design Persona	0.0000	3.33231e-251	Yes
Erikson's Psychosocial Stages	User-Centered Design	317422.0000	7.40516e-247	Yes
Flow Theory	Freudian Psychoanalysis	309015.0000	1.34794e-243	Yes
Flow Theory	Mental Models	315555.0000	3.32322e-246	Yes
Flow Theory	Myers-Briggs Type Indicator	313375.0000	2.45996e-245	Yes
Flow Theory	Social Cognitive Theory	298660.0000	1.81649e-239	Yes
Flow Theory	User Design Persona	0.0000	6.69287e-245	Yes
Flow Theory	User-Centered Design	301385.0000	1.48765e-240	Yes
Freudian Psychoanalysis	Mental Models	328293.0000	5.49647e-251	Yes
Freudian Psychoanalysis	Myers-Briggs Type Indicator	326025.0000	4.06852e-250	Yes
Freudian Psychoanalysis	Social Cognitive Theory	310716.0000	3.00359e-244	Yes
Freudian Psychoanalysis	User Design Persona	0.0000	1.10691e-249	Yes
Freudian Psychoanalysis	User-Centered Design	313551.0000	2.45995e-245	Yes
Mental Models	Myers-Briggs Type Indicator	0.0000	1.0032e-252	Yes
Mental Models	Social Cognitive Theory	0.0000	7.40517e-247	Yes
Mental Models	User Design Persona	0.0000	2.72935e-252	Yes
Mental Models	User-Centered Design	0.0000	6.06501e-248	Yes
Myers-Briggs Type Indicator	Social Cognitive Theory	315100.0000	5.48152e-246	Yes
Myers-Briggs Type Indicator	User Design Persona	0.0000	2.02027e-251	Yes
Myers-Briggs Type Indicator	User-Centered Design	317975.0000	4.48946e-247	Yes
Social Cognitive Theory	User Design Persona	0.0000	1.49137e-245	Yes
Social Cognitive Theory	User-Centered Design	303044.0000	3.31484e-241	Yes
User Design Persona	User-Centered Design	316869.0000	1.22145e-246	Yes

Table 14: Pairwise Weighted Mann–Whitney comparisons of description frameworks for Common Sense QA (Final Decision Maker).

Weighted Mann–Whitney for Final Decision Maker, Dataset: Last Letter Concat				
Framework 1	Framework 2	MW-stat	p-value	Sig?
Big Five Personality Traits	Cognitive Behavioral Theory	250000.0000	2.97418e-219	Yes
Big Five Personality Traits	Cognitive Load Theory	250000.0000	2.97418e-219	Yes
Big Five Personality Traits	Dual-Process Theory	0.0000	2.97418e-219	Yes
Big Five Personality Traits	Enneagram of Personality Traits	250000.0000	2.97418e-219	Yes
Big Five Personality Traits	Erikson's Psychosocial Stages	250000.0000	2.97418e-219	Yes
Big Five Personality Traits	Flow Theory	250000.0000	2.97418e-219	Yes
Big Five Personality Traits	Freudian Psychoanalysis	0.0000	2.97418e-219	Yes
Big Five Personality Traits	Mental Models	125000.0000	1	No
Big Five Personality Traits	Myers-Briggs Type Indicator	0.0000	2.97418e-219	Yes
Big Five Personality Traits	Social Cognitive Theory	0.0000	2.97418e-219	Yes
Big Five Personality Traits	User Design Persona	0.0000	2.97418e-219	Yes
Big Five Personality Traits	User-Centered Design	250000.0000	2.97418e-219	Yes
Cognitive Behavioral Theory	Cognitive Load Theory	250000.0000	2.97418e-219	Yes
Cognitive Behavioral Theory	Dual-Process Theory	0.0000	2.97418e-219	Yes
Cognitive Behavioral Theory	Enneagram of Personality Traits	250000.0000	2.97418e-219	Yes
Cognitive Behavioral Theory	Erikson's Psychosocial Stages	250000.0000	2.97418e-219	Yes
Cognitive Behavioral Theory	Flow Theory	250000.0000	2.97418e-219	Yes
Cognitive Behavioral Theory	Freudian Psychoanalysis	0.0000	2.97418e-219	Yes
Cognitive Behavioral Theory	Mental Models	0.0000	2.97418e-219	Yes
Cognitive Behavioral Theory	Myers-Briggs Type Indicator	0.0000	2.97418e-219	Yes
Cognitive Behavioral Theory	Social Cognitive Theory	0.0000	2.97418e-219	Yes
Cognitive Behavioral Theory	User Design Persona	0.0000	2.97418e-219	Yes
Cognitive Behavioral Theory	User-Centered Design	250000.0000	2.97418e-219	Yes
Cognitive Load Theory	Dual-Process Theory	0.0000	2.97418e-219	Yes
Cognitive Load Theory	Enneagram of Personality Traits	0.0000	2.97418e-219	Yes
Cognitive Load Theory	Erikson's Psychosocial Stages	250000.0000	2.97418e-219	Yes
Cognitive Load Theory	Flow Theory	0.0000	2.97418e-219	Yes
Cognitive Load Theory	Freudian Psychoanalysis	0.0000	2.97418e-219	Yes
Cognitive Load Theory	Mental Models	0.0000	2.97418e-219	Yes
Cognitive Load Theory	Myers-Briggs Type Indicator	0.0000	2.97418e-219	Yes
Cognitive Load Theory	Social Cognitive Theory	0.0000	2.97418e-219	Yes
Cognitive Load Theory	User Design Persona	0.0000	2.97418e-219	Yes
Cognitive Load Theory	User-Centered Design	250000.0000	2.97418e-219	Yes
Dual-Process Theory	Enneagram of Personality Traits	250000.0000	2.97418e-219	Yes
Dual-Process Theory	Erikson's Psychosocial Stages	250000.0000	2.97418e-219	Yes
Dual-Process Theory	Flow Theory	250000.0000	2.97418e-219	Yes
Dual-Process Theory	Freudian Psychoanalysis	250000.0000	2.97418e-219	Yes
Dual-Process Theory	Mental Models	250000.0000	2.97418e-219	Yes
Dual-Process Theory	Myers-Briggs Type Indicator	250000.0000	2.97418e-219	Yes
Dual-Process Theory	Social Cognitive Theory	250000.0000	2.97418e-219	Yes
Dual-Process Theory	User Design Persona	250000.0000	2.97418e-219	Yes
Dual-Process Theory	User-Centered Design	250000.0000	2.97418e-219	Yes
Enneagram of Personality Traits	Erikson's Psychosocial Stages	250000.0000	2.97418e-219	Yes
Enneagram of Personality Traits	Flow Theory	250000.0000	2.97418e-219	Yes
Enneagram of Personality Traits	Freudian Psychoanalysis	0.0000	2.97418e-219	Yes
Enneagram of Personality Traits	Mental Models	0.0000	2.97418e-219	Yes
Enneagram of Personality Traits	Myers-Briggs Type Indicator	0.0000	2.97418e-219	Yes
Enneagram of Personality Traits	Social Cognitive Theory	0.0000	2.97418e-219	Yes
Enneagram of Personality Traits	User Design Persona	0.0000	2.97418e-219	Yes
Enneagram of Personality Traits	User-Centered Design	250000.0000	2.97418e-219	Yes
Erikson's Psychosocial Stages	Flow Theory	0.0000	2.97418e-219	Yes
Erikson's Psychosocial Stages	Freudian Psychoanalysis	0.0000	2.97418e-219	Yes
Erikson's Psychosocial Stages	Mental Models	0.0000	2.97418e-219	Yes
Erikson's Psychosocial Stages	Myers-Briggs Type Indicator	0.0000	2.97418e-219	Yes
Erikson's Psychosocial Stages	Social Cognitive Theory	0.0000	2.97418e-219	Yes
Erikson's Psychosocial Stages	User Design Persona	0.0000	2.97418e-219	Yes
Erikson's Psychosocial Stages	User-Centered Design	0.0000	2.97418e-219	Yes
Flow Theory	Freudian Psychoanalysis	0.0000	2.97418e-219	Yes
Flow Theory	Mental Models	0.0000	2.97418e-219	Yes
Flow Theory	Myers-Briggs Type Indicator	0.0000	2.97418e-219	Yes
Flow Theory	Social Cognitive Theory	0.0000	2.97418e-219	Yes
Flow Theory	User Design Persona	0.0000	2.97418e-219	Yes
Flow Theory	User-Centered Design	250000.0000	2.97418e-219	Yes
Freudian Psychoanalysis	Mental Models	250000.0000	2.97418e-219	Yes
Freudian Psychoanalysis	Myers-Briggs Type Indicator	250000.0000	2.97418e-219	Yes
Freudian Psychoanalysis	Social Cognitive Theory	250000.0000	2.97418e-219	Yes
Freudian Psychoanalysis	User Design Persona	250000.0000	2.97418e-219	Yes
Freudian Psychoanalysis	User-Centered Design	250000.0000	2.97418e-219	Yes
Mental Models	Myers-Briggs Type Indicator	0.0000	2.97418e-219	Yes
Mental Models	Social Cognitive Theory	0.0000	2.97418e-219	Yes
Mental Models	User Design Persona	0.0000	2.97418e-219	Yes
Mental Models	User-Centered Design	250000.0000	2.97418e-219	Yes
Myers-Briggs Type Indicator	Social Cognitive Theory	250000.0000	2.97418e-219	Yes
Myers-Briggs Type Indicator	User Design Persona	125000.0000	1	No
Myers-Briggs Type Indicator	User-Centered Design	250000.0000	2.97418e-219	Yes
Social Cognitive Theory	User Design Persona	0.0000	2.97418e-219	Yes
Social Cognitive Theory	User-Centered Design	250000.0000	2.97418e-219	Yes
User Design Persona	User-Centered Design	250000.0000	2.97418e-219	Yes

Table 15: Pairwise Weighted Mann–Whitney comparisons of description frameworks for Last Letter Concat (Final Decision Maker).

Weighted Mann–Whitney for Majority Vote, Dataset: Last Letters				
Framework 1	Framework 2	MW-stat	p-value	Sig?
Big Five Personality Traits	Cognitive Behavioral Theory	125000.0000	1	No
Big Five Personality Traits	Cognitive Load Theory	250000.0000	2.97418e-219	Yes
Big Five Personality Traits	Dual-Process Theory	0.0000	2.97418e-219	Yes
Big Five Personality Traits	Enneagram of Personality Traits	250000.0000	2.97418e-219	Yes
Big Five Personality Traits	Erikson's Psychosocial Stages	250000.0000	2.97418e-219	Yes
Big Five Personality Traits	Flow Theory	250000.0000	2.97418e-219	Yes
Big Five Personality Traits	Freudian Psychoanalysis	0.0000	2.97418e-219	Yes
Big Five Personality Traits	Mental Models	250000.0000	2.97418e-219	Yes
Big Five Personality Traits	Myers-Briggs Type Indicator	0.0000	2.97418e-219	Yes
Big Five Personality Traits	Social Cognitive Theory	250000.0000	2.97418e-219	Yes
Big Five Personality Traits	User Design Persona	0.0000	2.97418e-219	Yes
Big Five Personality Traits	User-Centered Design	250000.0000	2.97418e-219	Yes
Cognitive Behavioral Theory	Cognitive Load Theory	250000.0000	2.97418e-219	Yes
Cognitive Behavioral Theory	Dual-Process Theory	0.0000	2.97418e-219	Yes
Cognitive Behavioral Theory	Enneagram of Personality Traits	250000.0000	2.97418e-219	Yes
Cognitive Behavioral Theory	Erikson's Psychosocial Stages	250000.0000	2.97418e-219	Yes
Cognitive Behavioral Theory	Flow Theory	250000.0000	2.97418e-219	Yes
Cognitive Behavioral Theory	Freudian Psychoanalysis	0.0000	2.97418e-219	Yes
Cognitive Behavioral Theory	Mental Models	250000.0000	2.97418e-219	Yes
Cognitive Behavioral Theory	Myers-Briggs Type Indicator	0.0000	2.97418e-219	Yes
Cognitive Behavioral Theory	Social Cognitive Theory	250000.0000	2.97418e-219	Yes
Cognitive Behavioral Theory	User Design Persona	0.0000	2.97418e-219	Yes
Cognitive Behavioral Theory	User-Centered Design	250000.0000	2.97418e-219	Yes
Cognitive Load Theory	Dual-Process Theory	0.0000	2.97418e-219	Yes
Cognitive Load Theory	Enneagram of Personality Traits	0.0000	2.97418e-219	Yes
Cognitive Load Theory	Erikson's Psychosocial Stages	250000.0000	2.97418e-219	Yes
Cognitive Load Theory	Flow Theory	0.0000	2.97418e-219	Yes
Cognitive Load Theory	Freudian Psychoanalysis	0.0000	2.97418e-219	Yes
Cognitive Load Theory	Mental Models	0.0000	2.97418e-219	Yes
Cognitive Load Theory	Myers-Briggs Type Indicator	0.0000	2.97418e-219	Yes
Cognitive Load Theory	Social Cognitive Theory	250000.0000	2.97418e-219	Yes
Cognitive Load Theory	User Design Persona	0.0000	2.97418e-219	Yes
Cognitive Load Theory	User-Centered Design	250000.0000	2.97418e-219	Yes
Dual-Process Theory	Enneagram of Personality Traits	250000.0000	2.97418e-219	Yes
Dual-Process Theory	Erikson's Psychosocial Stages	250000.0000	2.97418e-219	Yes
Dual-Process Theory	Flow Theory	250000.0000	2.97418e-219	Yes
Dual-Process Theory	Freudian Psychoanalysis	250000.0000	2.97418e-219	Yes
Dual-Process Theory	Mental Models	250000.0000	2.97418e-219	Yes
Dual-Process Theory	Myers-Briggs Type Indicator	250000.0000	2.97418e-219	Yes
Dual-Process Theory	Social Cognitive Theory	250000.0000	2.97418e-219	Yes
Dual-Process Theory	User Design Persona	250000.0000	2.97418e-219	Yes
Dual-Process Theory	User-Centered Design	250000.0000	2.97418e-219	Yes
Enneagram of Personality Traits	Erikson's Psychosocial Stages	250000.0000	2.97418e-219	Yes
Enneagram of Personality Traits	Flow Theory	0.0000	2.97418e-219	Yes
Enneagram of Personality Traits	Freudian Psychoanalysis	0.0000	2.97418e-219	Yes
Enneagram of Personality Traits	Mental Models	0.0000	2.97418e-219	Yes
Enneagram of Personality Traits	Myers-Briggs Type Indicator	0.0000	2.97418e-219	Yes
Enneagram of Personality Traits	Social Cognitive Theory	250000.0000	2.97418e-219	Yes
Enneagram of Personality Traits	User Design Persona	0.0000	2.97418e-219	Yes
Enneagram of Personality Traits	User-Centered Design	250000.0000	2.97418e-219	Yes
Erikson's Psychosocial Stages	Flow Theory	0.0000	2.97418e-219	Yes
Erikson's Psychosocial Stages	Freudian Psychoanalysis	0.0000	2.97418e-219	Yes
Erikson's Psychosocial Stages	Mental Models	0.0000	2.97418e-219	Yes
Erikson's Psychosocial Stages	Myers-Briggs Type Indicator	0.0000	2.97418e-219	Yes
Erikson's Psychosocial Stages	Social Cognitive Theory	250000.0000	2.97418e-219	Yes
Erikson's Psychosocial Stages	User Design Persona	0.0000	2.97418e-219	Yes
Erikson's Psychosocial Stages	User-Centered Design	250000.0000	2.97418e-219	Yes
Flow Theory	Freudian Psychoanalysis	0.0000	2.97418e-219	Yes
Flow Theory	Mental Models	0.0000	2.97418e-219	Yes
Flow Theory	Myers-Briggs Type Indicator	0.0000	2.97418e-219	Yes
Flow Theory	Social Cognitive Theory	250000.0000	2.97418e-219	Yes
Flow Theory	User Design Persona	0.0000	2.97418e-219	Yes
Flow Theory	User-Centered Design	250000.0000	2.97418e-219	Yes
Freudian Psychoanalysis	Mental Models	250000.0000	2.97418e-219	Yes
Freudian Psychoanalysis	Myers-Briggs Type Indicator	125000.0000	1	No
Freudian Psychoanalysis	Social Cognitive Theory	250000.0000	2.97418e-219	Yes
Freudian Psychoanalysis	User Design Persona	0.0000	2.97418e-219	Yes
Freudian Psychoanalysis	User-Centered Design	250000.0000	2.97418e-219	Yes
Mental Models	Myers-Briggs Type Indicator	0.0000	2.97418e-219	Yes
Mental Models	Social Cognitive Theory	250000.0000	2.97418e-219	Yes
Mental Models	User Design Persona	0.0000	2.97418e-219	Yes
Mental Models	User-Centered Design	250000.0000	2.97418e-219	Yes
Myers-Briggs Type Indicator	Social Cognitive Theory	250000.0000	2.97418e-219	Yes
Myers-Briggs Type Indicator	User Design Persona	0.0000	2.97418e-219	Yes
Myers-Briggs Type Indicator	User-Centered Design	250000.0000	2.97418e-219	Yes
Social Cognitive Theory	User Design Persona	0.0000	2.97418e-219	Yes
Social Cognitive Theory	User-Centered Design	250000.0000	2.97418e-219	Yes
User Design Persona	User-Centered Design	250000.0000	2.97418e-219	Yes

Table 16: Pairwise Weighted Mann–Whitney comparisons of description frameworks for Last Letter Concat (Majority Vote).

Weighted Mann–Whitney for Final Decision Maker, Dataset: SocialIQa				
Framework 1	Framework 2	MW-stat	p-value	Sig?
Big Five Personality Traits	Cognitive Behavioral Theory	0.0000	0	Yes
Big Five Personality Traits	Cognitive Load Theory	2794800.0000	0	Yes
Big Five Personality Traits	Dual-Process Theory	2762500.0000	0	Yes
Big Five Personality Traits	Enneagram of Personality Traits	2839000.0000	0	Yes
Big Five Personality Traits	Erikson's Psychosocial Stages	2952900.0000	0	Yes
Big Five Personality Traits	Flow Theory	2811800.0000	0	Yes
Big Five Personality Traits	Freudian Psychoanalysis	2927400.0000	0	Yes
Big Five Personality Traits	Mental Models	2907000.0000	0	Yes
Big Five Personality Traits	Myers-Briggs Type Indicator	2881500.0000	0	Yes
Big Five Personality Traits	Social Cognitive Theory	2759100.0000	0	Yes
Big Five Personality Traits	User Design Persona	0.0000	0	Yes
Big Five Personality Traits	User-Centered Design	2832200.0000	0	Yes
Cognitive Behavioral Theory	Cognitive Load Theory	2706024.0000	0	Yes
Cognitive Behavioral Theory	Dual-Process Theory	2674750.0000	0	Yes
Cognitive Behavioral Theory	Enneagram of Personality Traits	2748820.0000	0	Yes
Cognitive Behavioral Theory	Erikson's Psychosocial Stages	2859102.0000	0	Yes
Cognitive Behavioral Theory	Flow Theory	2722484.0000	0	Yes
Cognitive Behavioral Theory	Freudian Psychoanalysis	2834412.0000	0	Yes
Cognitive Behavioral Theory	Mental Models	2814660.0000	0	Yes
Cognitive Behavioral Theory	Myers-Briggs Type Indicator	2789970.0000	0	Yes
Cognitive Behavioral Theory	Social Cognitive Theory	2671458.0000	0	Yes
Cognitive Behavioral Theory	User Design Persona	0.0000	0	Yes
Cognitive Behavioral Theory	User-Centered Design	2742236.0000	0	Yes
Cognitive Load Theory	Dual-Process Theory	2671500.0000	0	Yes
Cognitive Load Theory	Enneagram of Personality Traits	0.0000	0	Yes
Cognitive Load Theory	Erikson's Psychosocial Stages	2855628.0000	0	Yes
Cognitive Load Theory	Flow Theory	0.0000	0	Yes
Cognitive Load Theory	Freudian Psychoanalysis	0.0000	0	Yes
Cognitive Load Theory	Mental Models	2811240.0000	0	Yes
Cognitive Load Theory	Myers-Briggs Type Indicator	2786580.0000	0	Yes
Cognitive Load Theory	Social Cognitive Theory	0.0000	0	Yes
Cognitive Load Theory	User Design Persona	0.0000	0	Yes
Cognitive Load Theory	User-Centered Design	2738904.0000	0	Yes
Dual-Process Theory	Enneagram of Personality Traits	0.0000	0	Yes
Dual-Process Theory	Erikson's Psychosocial Stages	0.0000	0	Yes
Dual-Process Theory	Flow Theory	0.0000	0	Yes
Dual-Process Theory	Freudian Psychoanalysis	0.0000	0	Yes
Dual-Process Theory	Mental Models	0.0000	0	Yes
Dual-Process Theory	Myers-Briggs Type Indicator	2754375.0000	0	Yes
Dual-Process Theory	Social Cognitive Theory	0.0000	0	Yes
Dual-Process Theory	User Design Persona	0.0000	0	Yes
Dual-Process Theory	User-Centered Design	0.0000	0	Yes
Enneagram of Personality Traits	Erikson's Psychosocial Stages	2900790.0000	0	Yes
Enneagram of Personality Traits	Flow Theory	0.0000	0	Yes
Enneagram of Personality Traits	Freudian Psychoanalysis	0.0000	0	Yes
Enneagram of Personality Traits	Mental Models	2855700.0000	0	Yes
Enneagram of Personality Traits	Myers-Briggs Type Indicator	2830650.0000	0	Yes
Enneagram of Personality Traits	Social Cognitive Theory	0.0000	0	Yes
Enneagram of Personality Traits	User Design Persona	0.0000	0	Yes
Enneagram of Personality Traits	User-Centered Design	2782220.0000	0	Yes
Erikson's Psychosocial Stages	Flow Theory	0.0000	0	Yes
Erikson's Psychosocial Stages	Freudian Psychoanalysis	0.0000	0	Yes
Erikson's Psychosocial Stages	Mental Models	2970270.0000	0	Yes
Erikson's Psychosocial Stages	Myers-Briggs Type Indicator	2944215.0000	0	Yes
Erikson's Psychosocial Stages	Social Cognitive Theory	0.0000	0	Yes
Erikson's Psychosocial Stages	User Design Persona	0.0000	0	Yes
Erikson's Psychosocial Stages	User-Centered Design	2893842.0000	0	Yes
Flow Theory	Freudian Psychoanalysis	0.0000	0	Yes
Flow Theory	Mental Models	2828340.0000	0	Yes
Flow Theory	Myers-Briggs Type Indicator	2803530.0000	0	Yes
Flow Theory	Social Cognitive Theory	2684442.0000	0	Yes
Flow Theory	User Design Persona	0.0000	0	Yes
Flow Theory	User-Centered Design	2755564.0000	0	Yes
Freudian Psychoanalysis	Mental Models	2944620.0000	0	Yes
Freudian Psychoanalysis	Myers-Briggs Type Indicator	2918790.0000	0	Yes
Freudian Psychoanalysis	Social Cognitive Theory	2794806.0000	0	Yes
Freudian Psychoanalysis	User Design Persona	0.0000	0	Yes
Freudian Psychoanalysis	User-Centered Design	2868852.0000	0	Yes
Mental Models	Myers-Briggs Type Indicator	2898450.0000	0	Yes
Mental Models	Social Cognitive Theory	0.0000	0	Yes
Mental Models	User Design Persona	0.0000	0	Yes
Mental Models	User-Centered Design	2848860.0000	0	Yes
Myers-Briggs Type Indicator	Social Cognitive Theory	0.0000	0	Yes
Myers-Briggs Type Indicator	User Design Persona	0.0000	0	Yes
Myers-Briggs Type Indicator	User-Centered Design	0.0000	0	Yes
Social Cognitive Theory	User Design Persona	0.0000	0	Yes
Social Cognitive Theory	User-Centered Design	2703918.0000	0	Yes
User Design Persona	User-Centered Design	2857190.0000	0	Yes

Table 17: Pairwise Weighted Mann–Whitney comparisons of description frameworks for Social IQa (Final Decision Maker).

Weighted Mann–Whitney for Majority Vote, Dataset: SocialIQa				
Framework 1	Framework 2	MW-stat	p-value	Sig?
Big Five Personality Traits	Cognitive Behavioral Theory	0.0000	0	Yes
Big Five Personality Traits	Cognitive Load Theory	2794800.0000	0	Yes
Big Five Personality Traits	Dual-Process Theory	2762500.0000	0	Yes
Big Five Personality Traits	Enneagram of Personality Traits	2839000.0000	0	Yes
Big Five Personality Traits	Erikson's Psychosocial Stages	2952900.0000	0	Yes
Big Five Personality Traits	Flow Theory	2811800.0000	0	Yes
Big Five Personality Traits	Freudian Psychoanalysis	2927400.0000	0	Yes
Big Five Personality Traits	Mental Models	2907000.0000	0	Yes
Big Five Personality Traits	Myers-Briggs Type Indicator	2881500.0000	0	Yes
Big Five Personality Traits	Social Cognitive Theory	2759100.0000	0	Yes
Big Five Personality Traits	User Design Persona	0.0000	0	Yes
Big Five Personality Traits	User-Centered Design	2832200.0000	0	Yes
Cognitive Behavioral Theory	Cognitive Load Theory	2706024.0000	0	Yes
Cognitive Behavioral Theory	Dual-Process Theory	2674750.0000	0	Yes
Cognitive Behavioral Theory	Enneagram of Personality Traits	2748820.0000	0	Yes
Cognitive Behavioral Theory	Erikson's Psychosocial Stages	2859102.0000	0	Yes
Cognitive Behavioral Theory	Flow Theory	2722484.0000	0	Yes
Cognitive Behavioral Theory	Freudian Psychoanalysis	2834412.0000	0	Yes
Cognitive Behavioral Theory	Mental Models	2814660.0000	0	Yes
Cognitive Behavioral Theory	Myers-Briggs Type Indicator	2789970.0000	0	Yes
Cognitive Behavioral Theory	Social Cognitive Theory	2671458.0000	0	Yes
Cognitive Behavioral Theory	User Design Persona	0.0000	0	Yes
Cognitive Behavioral Theory	User-Centered Design	2742236.0000	0	Yes
Cognitive Load Theory	Dual-Process Theory	0.0000	0	Yes
Cognitive Load Theory	Enneagram of Personality Traits	0.0000	0	Yes
Cognitive Load Theory	Erikson's Psychosocial Stages	2855628.0000	0	Yes
Cognitive Load Theory	Flow Theory	0.0000	0	Yes
Cognitive Load Theory	Freudian Psychoanalysis	0.0000	0	Yes
Cognitive Load Theory	Mental Models	2811240.0000	0	Yes
Cognitive Load Theory	Myers-Briggs Type Indicator	2786580.0000	0	Yes
Cognitive Load Theory	Social Cognitive Theory	0.0000	0	Yes
Cognitive Load Theory	User Design Persona	0.0000	0	Yes
Cognitive Load Theory	User-Centered Design	0.0000	0	Yes
Dual-Process Theory	Enneagram of Personality Traits	0.0000	0	Yes
Dual-Process Theory	Erikson's Psychosocial Stages	2822625.0000	0	Yes
Dual-Process Theory	Flow Theory	1343875.0000	1	No
Dual-Process Theory	Freudian Psychoanalysis	2798250.0000	0	Yes
Dual-Process Theory	Mental Models	2778750.0000	0	Yes
Dual-Process Theory	Myers-Briggs Type Indicator	2754375.0000	0	Yes
Dual-Process Theory	Social Cognitive Theory	0.0000	0	Yes
Dual-Process Theory	User Design Persona	0.0000	0	Yes
Dual-Process Theory	User-Centered Design	2707250.0000	0	Yes
Enneagram of Personality Traits	Erikson's Psychosocial Stages	2900790.0000	0	Yes
Enneagram of Personality Traits	Flow Theory	2762180.0000	0	Yes
Enneagram of Personality Traits	Freudian Psychoanalysis	2875740.0000	0	Yes
Enneagram of Personality Traits	Mental Models	2855700.0000	0	Yes
Enneagram of Personality Traits	Myers-Briggs Type Indicator	2830650.0000	0	Yes
Enneagram of Personality Traits	Social Cognitive Theory	0.0000	0	Yes
Enneagram of Personality Traits	User Design Persona	0.0000	0	Yes
Enneagram of Personality Traits	User-Centered Design	2782220.0000	0	Yes
Erikson's Psychosocial Stages	Flow Theory	0.0000	0	Yes
Erikson's Psychosocial Stages	Freudian Psychoanalysis	0.0000	0	Yes
Erikson's Psychosocial Stages	Mental Models	2970270.0000	0	Yes
Erikson's Psychosocial Stages	Myers-Briggs Type Indicator	2944215.0000	0	Yes
Erikson's Psychosocial Stages	Social Cognitive Theory	0.0000	0	Yes
Erikson's Psychosocial Stages	User Design Persona	0.0000	0	Yes
Erikson's Psychosocial Stages	User-Centered Design	0.0000	0	Yes
Flow Theory	Freudian Psychoanalysis	2848188.0000	0	Yes
Flow Theory	Mental Models	2828340.0000	0	Yes
Flow Theory	Myers-Briggs Type Indicator	2803530.0000	0	Yes
Flow Theory	Social Cognitive Theory	0.0000	0	Yes
Flow Theory	User Design Persona	0.0000	0	Yes
Flow Theory	User-Centered Design	2755564.0000	0	Yes
Freudian Psychoanalysis	Mental Models	2944620.0000	0	Yes
Freudian Psychoanalysis	Myers-Briggs Type Indicator	2918790.0000	0	Yes
Freudian Psychoanalysis	Social Cognitive Theory	0.0000	0	Yes
Freudian Psychoanalysis	User Design Persona	0.0000	0	Yes
Freudian Psychoanalysis	User-Centered Design	0.0000	0	Yes
Mental Models	Myers-Briggs Type Indicator	0.0000	0	Yes
Mental Models	Social Cognitive Theory	0.0000	0	Yes
Mental Models	User Design Persona	0.0000	0	Yes
Mental Models	User-Centered Design	0.0000	0	Yes
Myers-Briggs Type Indicator	Social Cognitive Theory	0.0000	0	Yes
Myers-Briggs Type Indicator	User Design Persona	0.0000	0	Yes
Myers-Briggs Type Indicator	User-Centered Design	0.0000	0	Yes
Social Cognitive Theory	User Design Persona	0.0000	0	Yes
Social Cognitive Theory	User-Centered Design	2703918.0000	0	Yes
User Design Persona	User-Centered Design	2857190.0000	0	Yes

Table 18: Pairwise Weighted Mann–Whitney comparisons of description frameworks for Social IQa (Majority Vote).

Weighted Mann–Whitney for Final Decision Maker, Dataset: Social Support				
Framework 1	Framework 2	MW-stat	p-value	Sig?
Big Five Personality Traits	Cognitive Behavioral Theory	0.0000	0	Yes
Big Five Personality Traits	Cognitive Load Theory	402304.5000	1	No
Big Five Personality Traits	Dual-Process Theory	0.0000	0	Yes
Big Five Personality Traits	Enneagram of Personality Traits	0.0000	0	Yes
Big Five Personality Traits	Erikson's Psychosocial Stages	0.0000	0	Yes
Big Five Personality Traits	Flow Theory	0.0000	0	Yes
Big Five Personality Traits	Freudian Psychoanalysis	0.0000	0	Yes
Big Five Personality Traits	Mental Models	0.0000	0	Yes
Big Five Personality Traits	Myers-Briggs Type Indicator	0.0000	0	Yes
Big Five Personality Traits	Social Cognitive Theory	0.0000	0	Yes
Big Five Personality Traits	User Design Persona	0.0000	0	Yes
Big Five Personality Traits	User-Centered Design	0.0000	0	Yes
Cognitive Behavioral Theory	Cognitive Load Theory	804609.0000	0	Yes
Cognitive Behavioral Theory	Dual-Process Theory	804609.0000	0	Yes
Cognitive Behavioral Theory	Enneagram of Personality Traits	804609.0000	0	Yes
Cognitive Behavioral Theory	Erikson's Psychosocial Stages	804609.0000	0	Yes
Cognitive Behavioral Theory	Flow Theory	804609.0000	0	Yes
Cognitive Behavioral Theory	Freudian Psychoanalysis	804609.0000	0	Yes
Cognitive Behavioral Theory	Mental Models	804609.0000	0	Yes
Cognitive Behavioral Theory	Myers-Briggs Type Indicator	804609.0000	0	Yes
Cognitive Behavioral Theory	Social Cognitive Theory	402304.5000	1	No
Cognitive Behavioral Theory	User Design Persona	804609.0000	0	Yes
Cognitive Behavioral Theory	User-Centered Design	804609.0000	0	Yes
Cognitive Load Theory	Dual-Process Theory	0.0000	0	Yes
Cognitive Load Theory	Enneagram of Personality Traits	0.0000	0	Yes
Cognitive Load Theory	Erikson's Psychosocial Stages	0.0000	0	Yes
Cognitive Load Theory	Flow Theory	0.0000	0	Yes
Cognitive Load Theory	Freudian Psychoanalysis	0.0000	0	Yes
Cognitive Load Theory	Mental Models	0.0000	0	Yes
Cognitive Load Theory	Myers-Briggs Type Indicator	0.0000	0	Yes
Cognitive Load Theory	Social Cognitive Theory	0.0000	0	Yes
Cognitive Load Theory	User Design Persona	0.0000	0	Yes
Cognitive Load Theory	User-Centered Design	0.0000	0	Yes
Dual-Process Theory	Enneagram of Personality Traits	0.0000	0	Yes
Dual-Process Theory	Erikson's Psychosocial Stages	804609.0000	0	Yes
Dual-Process Theory	Flow Theory	0.0000	0	Yes
Dual-Process Theory	Freudian Psychoanalysis	0.0000	0	Yes
Dual-Process Theory	Mental Models	0.0000	0	Yes
Dual-Process Theory	Myers-Briggs Type Indicator	0.0000	0	Yes
Dual-Process Theory	Social Cognitive Theory	0.0000	0	Yes
Dual-Process Theory	User Design Persona	804609.0000	0	Yes
Dual-Process Theory	User-Centered Design	0.0000	0	Yes
Enneagram of Personality Traits	Erikson's Psychosocial Stages	804609.0000	0	Yes
Enneagram of Personality Traits	Flow Theory	804609.0000	0	Yes
Enneagram of Personality Traits	Freudian Psychoanalysis	804609.0000	0	Yes
Enneagram of Personality Traits	Mental Models	804609.0000	0	Yes
Enneagram of Personality Traits	Myers-Briggs Type Indicator	0.0000	0	Yes
Enneagram of Personality Traits	Social Cognitive Theory	0.0000	0	Yes
Enneagram of Personality Traits	User Design Persona	804609.0000	0	Yes
Enneagram of Personality Traits	User-Centered Design	0.0000	0	Yes
Erikson's Psychosocial Stages	Flow Theory	0.0000	0	Yes
Erikson's Psychosocial Stages	Freudian Psychoanalysis	0.0000	0	Yes
Erikson's Psychosocial Stages	Mental Models	0.0000	0	Yes
Erikson's Psychosocial Stages	Myers-Briggs Type Indicator	0.0000	0	Yes
Erikson's Psychosocial Stages	Social Cognitive Theory	0.0000	0	Yes
Erikson's Psychosocial Stages	User Design Persona	0.0000	0	Yes
Erikson's Psychosocial Stages	User-Centered Design	0.0000	0	Yes
Flow Theory	Freudian Psychoanalysis	804609.0000	0	Yes
Flow Theory	Mental Models	804609.0000	0	Yes
Flow Theory	Myers-Briggs Type Indicator	0.0000	0	Yes
Flow Theory	Social Cognitive Theory	0.0000	0	Yes
Flow Theory	User Design Persona	804609.0000	0	Yes
Flow Theory	User-Centered Design	0.0000	0	Yes
Freudian Psychoanalysis	Mental Models	0.0000	0	Yes
Freudian Psychoanalysis	Myers-Briggs Type Indicator	0.0000	0	Yes
Freudian Psychoanalysis	Social Cognitive Theory	0.0000	0	Yes
Freudian Psychoanalysis	User Design Persona	804609.0000	0	Yes
Freudian Psychoanalysis	User-Centered Design	0.0000	0	Yes
Mental Models	Myers-Briggs Type Indicator	0.0000	0	Yes
Mental Models	Social Cognitive Theory	0.0000	0	Yes
Mental Models	User Design Persona	804609.0000	0	Yes
Mental Models	User-Centered Design	0.0000	0	Yes
Myers-Briggs Type Indicator	Social Cognitive Theory	0.0000	0	Yes
Myers-Briggs Type Indicator	User Design Persona	804609.0000	0	Yes
Myers-Briggs Type Indicator	User-Centered Design	0.0000	0	Yes
Social Cognitive Theory	User Design Persona	804609.0000	0	Yes
Social Cognitive Theory	User-Centered Design	804609.0000	0	Yes
User Design Persona	User-Centered Design	0.0000	0	Yes

Table 19: Pairwise Weighted Mann–Whitney comparisons of description frameworks for Social Support (Final Decision Maker).

Weighted Mann–Whitney for Majority Vote, Dataset: Social Support				
Framework 1	Framework 2	MW-stat	p-value	Sig?
Big Five Personality Traits	Cognitive Behavioral Theory	0.0000	0	Yes
Big Five Personality Traits	Cognitive Load Theory	0.0000	0	Yes
Big Five Personality Traits	Dual-Process Theory	0.0000	0	Yes
Big Five Personality Traits	Enneagram of Personality Traits	0.0000	0	Yes
Big Five Personality Traits	Erikson's Psychosocial Stages	0.0000	0	Yes
Big Five Personality Traits	Flow Theory	0.0000	0	Yes
Big Five Personality Traits	Freudian Psychoanalysis	0.0000	0	Yes
Big Five Personality Traits	Mental Models	0.0000	0	Yes
Big Five Personality Traits	Myers-Briggs Type Indicator	0.0000	0	Yes
Big Five Personality Traits	Social Cognitive Theory	0.0000	0	Yes
Big Five Personality Traits	User Design Persona	0.0000	0	Yes
Big Five Personality Traits	User-Centered Design	0.0000	0	Yes
Cognitive Behavioral Theory	Cognitive Load Theory	804609.0000	0	Yes
Cognitive Behavioral Theory	Dual-Process Theory	804609.0000	0	Yes
Cognitive Behavioral Theory	Enneagram of Personality Traits	402304.5000	1	No
Cognitive Behavioral Theory	Erikson's Psychosocial Stages	804609.0000	0	Yes
Cognitive Behavioral Theory	Flow Theory	402304.5000	1	No
Cognitive Behavioral Theory	Freudian Psychoanalysis	804609.0000	0	Yes
Cognitive Behavioral Theory	Mental Models	804609.0000	0	Yes
Cognitive Behavioral Theory	Myers-Briggs Type Indicator	804609.0000	0	Yes
Cognitive Behavioral Theory	Social Cognitive Theory	0.0000	0	Yes
Cognitive Behavioral Theory	User Design Persona	804609.0000	0	Yes
Cognitive Behavioral Theory	User-Centered Design	804609.0000	0	Yes
Cognitive Load Theory	Dual-Process Theory	804609.0000	0	Yes
Cognitive Load Theory	Enneagram of Personality Traits	0.0000	0	Yes
Cognitive Load Theory	Erikson's Psychosocial Stages	0.0000	0	Yes
Cognitive Load Theory	Flow Theory	0.0000	0	Yes
Cognitive Load Theory	Freudian Psychoanalysis	0.0000	0	Yes
Cognitive Load Theory	Mental Models	0.0000	0	Yes
Cognitive Load Theory	Myers-Briggs Type Indicator	0.0000	0	Yes
Cognitive Load Theory	Social Cognitive Theory	0.0000	0	Yes
Cognitive Load Theory	User Design Persona	0.0000	0	Yes
Cognitive Load Theory	User-Centered Design	0.0000	0	Yes
Dual-Process Theory	Enneagram of Personality Traits	0.0000	0	Yes
Dual-Process Theory	Erikson's Psychosocial Stages	0.0000	0	Yes
Dual-Process Theory	Flow Theory	0.0000	0	Yes
Dual-Process Theory	Freudian Psychoanalysis	0.0000	0	Yes
Dual-Process Theory	Mental Models	0.0000	0	Yes
Dual-Process Theory	Myers-Briggs Type Indicator	0.0000	0	Yes
Dual-Process Theory	Social Cognitive Theory	0.0000	0	Yes
Dual-Process Theory	User Design Persona	0.0000	0	Yes
Dual-Process Theory	User-Centered Design	0.0000	0	Yes
Enneagram of Personality Traits	Erikson's Psychosocial Stages	804609.0000	0	Yes
Enneagram of Personality Traits	Flow Theory	402304.5000	1	No
Enneagram of Personality Traits	Freudian Psychoanalysis	804609.0000	0	Yes
Enneagram of Personality Traits	Mental Models	804609.0000	0	Yes
Enneagram of Personality Traits	Myers-Briggs Type Indicator	804609.0000	0	Yes
Enneagram of Personality Traits	Social Cognitive Theory	0.0000	0	Yes
Enneagram of Personality Traits	User Design Persona	804609.0000	0	Yes
Enneagram of Personality Traits	User-Centered Design	804609.0000	0	Yes
Erikson's Psychosocial Stages	Flow Theory	0.0000	0	Yes
Erikson's Psychosocial Stages	Freudian Psychoanalysis	804609.0000	0	Yes
Erikson's Psychosocial Stages	Mental Models	0.0000	0	Yes
Erikson's Psychosocial Stages	Myers-Briggs Type Indicator	0.0000	0	Yes
Erikson's Psychosocial Stages	Social Cognitive Theory	0.0000	0	Yes
Erikson's Psychosocial Stages	User Design Persona	0.0000	0	Yes
Erikson's Psychosocial Stages	User-Centered Design	0.0000	0	Yes
Flow Theory	Freudian Psychoanalysis	804609.0000	0	Yes
Flow Theory	Mental Models	804609.0000	0	Yes
Flow Theory	Myers-Briggs Type Indicator	804609.0000	0	Yes
Flow Theory	Social Cognitive Theory	0.0000	0	Yes
Flow Theory	User Design Persona	804609.0000	0	Yes
Flow Theory	User-Centered Design	804609.0000	0	Yes
Freudian Psychoanalysis	Mental Models	0.0000	0	Yes
Freudian Psychoanalysis	Myers-Briggs Type Indicator	0.0000	0	Yes
Freudian Psychoanalysis	Social Cognitive Theory	0.0000	0	Yes
Freudian Psychoanalysis	User Design Persona	0.0000	0	Yes
Freudian Psychoanalysis	User-Centered Design	0.0000	0	Yes
Mental Models	Myers-Briggs Type Indicator	0.0000	0	Yes
Mental Models	Social Cognitive Theory	0.0000	0	Yes
Mental Models	User Design Persona	0.0000	0	Yes
Mental Models	User-Centered Design	0.0000	0	Yes
Myers-Briggs Type Indicator	Social Cognitive Theory	0.0000	0	Yes
Myers-Briggs Type Indicator	User Design Persona	0.0000	0	Yes
Myers-Briggs Type Indicator	User-Centered Design	0.0000	0	Yes
Social Cognitive Theory	User Design Persona	804609.0000	0	Yes
Social Cognitive Theory	User-Centered Design	804609.0000	0	Yes
User Design Persona	User-Centered Design	804609.0000	0	Yes

Table 20: Pairwise Weighted Mann–Whitney comparisons of description frameworks for Social Support (Majority Vote).

Permutation test for Final Decision Maker, Dataset: Social Support				
Framework 1	Framework 2	Ob. diff.	p-value	Sig?
Big Five Personality Traits	Cognitive Behavioral Theory	0.07478720098940464	0.0	Yes
Big Five Personality Traits	Cognitive Load Theory	0.00018706392265849248	0.978	No
Big Five Personality Traits	Dual-Process Theory	0.004288782590265855	0.436	No
Big Five Personality Traits	Enneagram of Personality Traits	0.06990413270040416	0.0	Yes
Big Five Personality Traits	Erikson's Psychosocial Stages	0.001208191025936306	0.827	No
Big Five Personality Traits	Flow Theory	0.06977930844300786	0.0	Yes
Big Five Personality Traits	Freudian Psychoanalysis	0.06478344650610132	0.0	Yes
Big Five Personality Traits	Mental Models	0.06484863387346404	0.0	Yes
Big Five Personality Traits	Myers-Briggs Type Indicator	0.1738772620351862	0.0	Yes
Big Five Personality Traits	Social Cognitive Theory	0.014690781269527331	0.025	Yes
Big Five Personality Traits	User Design Persona	0.0028452629964489107	0.631	No
Big Five Personality Traits	User-Centered Design	0.07536115789968492	0.0	Yes
Cognitive Behavioral Theory	Cognitive Load Theory	0.07497426491206313	0.0	Yes
Cognitive Behavioral Theory	Dual-Process Theory	0.07049841839913878	0.0	Yes
Cognitive Behavioral Theory	Enneagram of Personality Traits	0.004883068289000481	0.473	No
Cognitive Behavioral Theory	Erikson's Psychosocial Stages	0.07357900996346833	0.0	Yes
Cognitive Behavioral Theory	Flow Theory	0.005007892546396775	0.561	No
Cognitive Behavioral Theory	Freudian Psychoanalysis	0.010003754483303318	0.311	No
Cognitive Behavioral Theory	Mental Models	0.0099385671159406	0.347	No
Cognitive Behavioral Theory	Myers-Briggs Type Indicator	0.09909006104578155	0.0	Yes
Cognitive Behavioral Theory	Social Cognitive Theory	0.06009641971987731	0.0	Yes
Cognitive Behavioral Theory	User Design Persona	0.07194193799295573	0.0	Yes
Cognitive Behavioral Theory	User-Centered Design	0.0005739569102802822	0.941	No
Cognitive Load Theory	Dual-Process Theory	0.004475846512924347	0.425	No
Cognitive Load Theory	Enneagram of Personality Traits	0.07009119662306265	0.001	Yes
Cognitive Load Theory	Erikson's Psychosocial Stages	0.0013952549485947985	0.822	No
Cognitive Load Theory	Flow Theory	0.06996637236566636	0.0	Yes
Cognitive Load Theory	Freudian Psychoanalysis	0.06497051042875981	0.0	Yes
Cognitive Load Theory	Mental Models	0.06503569779612253	0.0	Yes
Cognitive Load Theory	Myers-Briggs Type Indicator	0.17406432595784468	0.0	Yes
Cognitive Load Theory	Social Cognitive Theory	0.014877845192185823	0.059	No
Cognitive Load Theory	User Design Persona	0.003032326919107403	0.561	No
Cognitive Load Theory	User-Centered Design	0.07554822182234341	0.0	Yes
Dual-Process Theory	Enneagram of Personality Traits	0.0656153501101383	0.0	Yes
Dual-Process Theory	Erikson's Psychosocial Stages	0.003080591564329549	0.566	No
Dual-Process Theory	Flow Theory	0.06549052585274201	0.0	Yes
Dual-Process Theory	Freudian Psychoanalysis	0.06049466391583547	0.001	Yes
Dual-Process Theory	Mental Models	0.060559851283198185	0.001	Yes
Dual-Process Theory	Myers-Briggs Type Indicator	0.16958847944492034	0.0	Yes
Dual-Process Theory	Social Cognitive Theory	0.010401998679261476	0.191	No
Dual-Process Theory	User Design Persona	0.0014435195938169443	0.797	No
Dual-Process Theory	User-Centered Design	0.07107237530941907	0.0	Yes
Enneagram of Personality Traits	Erikson's Psychosocial Stages	0.06869594167446785	0.001	Yes
Enneagram of Personality Traits	Flow Theory	0.00012482425739629432	0.987	No
Enneagram of Personality Traits	Freudian Psychoanalysis	0.0051206861943028 37	0.607	No
Enneagram of Personality Traits	Mental Models	0.0050554988269401 19	0.587	No
Enneagram of Personality Traits	Myers-Briggs Type Indicator	0.10397312933478203	0.0	Yes
Enneagram of Personality Traits	Social Cognitive Theory	0.05521335143087683	0.009	Yes
Enneagram of Personality Traits	User Design Persona	0.06705886970395525	0.0	Yes
Enneagram of Personality Traits	User-Centered Design	0.005457025199280763	0.5	No
Erikson's Psychosocial Stages	Flow Theory	0.06857111741707156	0.0	Yes
Erikson's Psychosocial Stages	Freudian Psychoanalysis	0.06357525548016502	0.0	Yes
Erikson's Psychosocial Stages	Mental Models	0.06364044284752773	0.0	Yes
Erikson's Psychosocial Stages	Myers-Briggs Type Indicator	0.17266907100924989	0.0	Yes
Erikson's Psychosocial Stages	Social Cognitive Theory	0.013482590243591025	0.089	No
Erikson's Psychosocial Stages	User Design Persona	0.0016370719705126047	0.73	No
Erikson's Psychosocial Stages	User-Centered Design	0.07415296687374862	0.0	Yes
Flow Theory	Freudian Psychoanalysis	0.004995861936906543	0.489	No
Flow Theory	Mental Models	0.0049306745695438 25	0.461	No
Flow Theory	Myers-Briggs Type Indicator	0.10409795359217833	0.0	Yes
Flow Theory	Social Cognitive Theory	0.055088527173480534	0.001	Yes
Flow Theory	User Design Persona	0.06693404544655895	0.0	Yes
Flow Theory	User-Centered Design	0.0055818494566770 57	0.394	No
Freudian Psychoanalysis	Mental Models	6.518736736271791e-05	0.984	No
Freudian Psychoanalysis	Myers-Briggs Type Indicator	0.10909381552908487	0.0	Yes
Freudian Psychoanalysis	Social Cognitive Theory	0.05009266523657399	0.009	Yes
Freudian Psychoanalysis	User Design Persona	0.06193818350965241	0.0	Yes
Freudian Psychoanalysis	User-Centered Design	0.0105777113935836	0.165	No
Mental Models	Myers-Briggs Type Indicator	0.10902862816172215	0.0	Yes
Mental Models	Social Cognitive Theory	0.05015785260393671	0.002	Yes
Mental Models	User Design Persona	0.06200337087701513	0.0	Yes
Mental Models	User-Centered Design	0.010512524026220882	0.12	No
Myers-Briggs Type Indicator	Social Cognitive Theory	0.15918648076565886	0.0	Yes
Myers-Briggs Type Indicator	User Design Persona	0.17103199903873728	0.0	Yes
Myers-Briggs Type Indicator	User-Centered Design	0.09851610413550127	0.0	Yes
Social Cognitive Theory	User Design Persona	0.01184551827307842	0.167	No
Social Cognitive Theory	User-Centered Design	0.06067037663015759	0.0	Yes
User Design Persona	User-Centered Design	0.07251589490323601	0.0	Yes

Table 21: Pairwise Permutational Test comparisons of description frameworks for Social Support (Final Decision Maker).

Dataset	*F-statistic*	*p-value*
Strategy QA	4.233278	0.002217
Common Sense QA	0.653576	0.660868
Last Letter Concat	0.504182	0.872550
Social IQa	8.901237	0.000003
Social Support	2.380994	0.045282

Table 22: Weighted ANOVA (Welch's ANOVA) Test Results Within Each Dataset.

Dataset	*H-statistic*	*p-value*
Strategy QA	6.695412	0.461267
Common Sense QA	5.464388	0.361867
Last Letter Concat	5.247787	0.874023
Social IQa	14.857131	0.094936
Social Support	14.967613	0.036417

Table 23: Weighted Kruskal-Wallis Test Results Within Each Dataset.

Dataset	*Expert Role*	*Count*
Social IQa	Behavioral Analysis	1
Social IQa	Behavioral Science	1
Social IQa	Cognitive Behavioral Analysis	1
Social IQa	Communication Studies	1
Social IQa	Emotional Intelligence	10
Social IQa	Interpersonal Communication	6
Social IQa	Interpersonal Development	1
Social IQa	Psychology	5
Social IQa	Social Psychology	9
Social IQa	Sociology	4
Common Sense QA	Artificial Intelligence	13
Common Sense QA	Cognitive Psychology	4
Common Sense QA	Cognitive Science	8
Common Sense QA	Knowledge Representation and Reasoning	1
Common Sense QA	Linguistics	1
Common Sense QA	Natural Language Processing	12
Last Letter Concat	Algorithmic Problem Solving	1
Last Letter Concat	Coding Theory	1
Last Letter Concat	Computational Linguistics	8
Last Letter Concat	Cryptography	4
Last Letter Concat	Linguistics	1
Last Letter Concat	Morphology	7
Last Letter Concat	Natural Language Processing	4
Last Letter Concat	Orthography	2
Last Letter Concat	Phonetics	7
Last Letter Concat	Phonology	2
Last Letter Concat	String Manipulation	2
Social Support	Artificial Intelligence and Machine Learning	1
Social Support	Computational Linguistics	4
Social Support	Linguistics	3
Social Support	Machine Learning	5
Social Support	Natural Language Processing	13
Social Support	Psycholinguistics	1
Social Support	Psychology	1
Social Support	Social Psychology	11
Strategy QA	Artificial Intelligence	12
Strategy QA	Cognitive Psychology	6
Strategy QA	Cognitive Science	6
Strategy QA	Formal Logic	5
Strategy QA	Logic and Formal Reasoning	3
Strategy QA	Machine Learning	1
Strategy QA	Mathematical Logic	2
Strategy QA	Natural Language Processing	4

Table 24: Expert Role Distribution Within Each Dataset by summing the number of times a role is used among the description frameworks.

	Group 1	Group 2	Mean Diff.	p-value	Lower	Upper	Significant
1	Artificial Intelligence	Cognitive Psychology	1408.1692	0.0052	260.0531	2556.2852	True
2	Artificial Intelligence	Cognitive Science	84.5792	1.0000	-1063.5369	1232.6952	False
3	Artificial Intelligence	Formal Logic	345.3855	0.9893	-876.8774	1567.6484	False
4	Artificial Intelligence	Logic and Formal Reasoning	2407.1308	0.0000	924.9194	3889.3423	True
5	Artificial Intelligence	Machine Learning	3558.8175	0.0002	1168.8234	5948.8116	True
6	Artificial Intelligence	Mathematical Logic	-725.5825	0.9132	-2479.3587	1028.1937	False
7	Artificial Intelligence	Natural Language Processing	1418.9850	0.0262	93.2548	2744.7152	True
8	Cognitive Psychology	Cognitive Science	-1323.5900	0.0507	-2649.3202	2.1402	False
9	Cognitive Psychology	Formal Logic	-1062.7837	0.2812	-2453.2212	327.6539	False
10	Cognitive Psychology	Logic and Formal Reasoning	998.9617	0.5704	-624.7196	2622.6429	False
11	Cognitive Psychology	Machine Learning	2150.6483	0.1444	-329.5658	4630.8624	False
12	Cognitive Psychology	Mathematical Logic	-2133.7517	0.0134	-4008.6173	-258.8860	True
13	Cognitive Psychology	Natural Language Processing	10.8158	1.0000	-1471.3956	1493.0273	False
14	Cognitive Science	Formal Logic	260.8063	0.9992	-1129.6312	1651.2439	False
15	Cognitive Science	Logic and Formal Reasoning	2322.5517	0.0004	698.8704	3946.2329	True
16	Cognitive Science	Machine Learning	3474.2383	0.0006	994.0242	5954.4524	True
17	Cognitive Science	Mathematical Logic	-810.1617	0.8929	-2685.0273	1064.7040	False
18	Cognitive Science	Natural Language Processing	1334.4058	0.1131	-147.8056	2816.6173	False
19	Formal Logic	Logic and Formal Reasoning	2061.7453	0.0050	384.8145	3738.6761	True
20	Formal Logic	Machine Learning	3213.4320	0.0029	698.0358	5728.8282	True
21	Formal Logic	Mathematical Logic	-1070.9680	0.6894	-2992.1336	850.1976	False
22	Formal Logic	Natural Language Processing	1073.5995	0.4024	-466.7598	2613.9588	False
23	Logic and Formal Reasoning	Machine Learning	1151.6867	0.8903	-1499.7737	3803.1471	False
24	Logic and Formal Reasoning	Mathematical Logic	-3132.7133	0.0002	-5228.8768	-1036.5498	True
25	Logic and Formal Reasoning	Natural Language Processing	-988.1458	0.6775	-2741.9220	765.6304	False
26	Machine Learning	Mathematical Logic	-4284.4000	0.0001	-7096.6985	-1472.1015	True
27	Machine Learning	Natural Language Processing	-2139.8325	0.1824	-4707.0980	427.4330	False
28	Mathematical Logic	Natural Language Processing	2144.5675	0.0242	155.9722	4133.1628	True

Table 25: Tukey's HSD Pairwise Comparison Results for Strategy QA. The columns "Upper" and "Lower" refer to the confidence interval bounds for the difference between the means of the two groups.

	Group 1	Group 2	Mean Diff.	p-value	Lower	Upper	Significant
1	Behavioral Analysis	Behavioral Science	8151.6300	0.0000	4760.6515	11542.6085	True
2	Behavioral Analysis	Cognitive Behavioral Analysis	8241.9300	0.0000	4850.9515	11632.9085	True
3	Behavioral Analysis	Communication Studies	9021.0300	0.0000	5630.0515	12412.0085	True
4	Behavioral Analysis	Emotional Intelligence	4399.5480	0.0000	1884.7310	6914.3650	True
5	Behavioral Analysis	Interpersonal Communication	4109.0417	0.0000	1519.1391	6698.9443	True
6	Behavioral Analysis	Interpersonal Development	11538.6900	0.0000	8147.7115	14929.6685	True
7	Behavioral Analysis	Psychology	4738.2040	0.0000	2111.5633	7364.8447	True
8	Behavioral Analysis	Social Psychology	5968.3100	0.0000	3440.8239	8495.7961	True
9	Behavioral Analysis	Sociology	5467.3250	0.0000	2786.5211	8148.1289	True

Table 26: Tukey's HSD Pairwise Comparison Results for Social IQa. The columns "Upper" and "Lower" refer to the confidence interval bounds for the difference between the means of the two groups.

	Group 1	Group 2	Mean Diff.	p-value	Lower	Upper	Significant
1	Artificial Intelligence and Machine Learning	Computational Linguistics	-1141.4325	0.9839	-5261.5507	2978.6857	False
2	Artificial Intelligence and Machine Learning	Linguistics	-2146.8200	0.7247	-6402.0597	2108.4197	False
3	Artificial Intelligence and Machine Learning	Machine Learning	-1365.2340	0.9525	-5402.1089	2671.6409	False
4	Artificial Intelligence and Machine Learning	Natural Language Processing	-2370.8400	0.4902	-6195.0964	1453.4164	False
5	Artificial Intelligence and Machine Learning	Psycholinguistics	-1139.1900	0.9961	-6350.7730	4072.3930	False
6	Artificial Intelligence and Machine Learning	Psychology	-1542.8400	0.9767	-6754.4230	3668.7430	False
7	Artificial Intelligence and Machine Learning	Social Psychology	-680.9045	0.9990	-4529.9137	3168.1046	False
8	Computational Linguistics	Linguistics	-1005.3875	0.9373	-3819.9640	1809.1890	False
9	Computational Linguistics	Machine Learning	-223.8015	1.0000	-2695.8724	2248.2694	False
10	Computational Linguistics	Natural Language Processing	-1229.4075	0.5648	-3336.4702	877.6552	False
11	Computational Linguistics	Psycholinguistics	2.2425	1.0000	-4117.8757	4122.3607	False
12	Computational Linguistics	Psychology	-401.4075	1.0000	-4521.5257	3718.7107	False
13	Computational Linguistics	Social Psychology	460.5280	0.9965	-1691.1336	2612.1895	False
14	Linguistics	Machine Learning	781.5860	0.9790	-1909.6639	3472.8359	False
15	Linguistics	Natural Language Processing	-224.0200	1.0000	-2584.4023	2136.3623	False
16	Linguistics	Psycholinguistics	1007.6300	0.9936	-3247.6097	5262.8697	False
17	Linguistics	Psychology	603.9800	0.9998	-3651.2597	4859.2197	False
18	Linguistics	Social Psychology	1465.9155	0.5088	-934.3635	3866.1944	False
19	Machine Learning	Natural Language Processing	-1005.6060	0.6973	-2944.8580	933.6460	False
20	Machine Learning	Psycholinguistics	226.0440	1.0000	-3810.8309	4262.9189	False
21	Machine Learning	Psychology	-177.6060	1.0000	-4214.4809	3859.2689	False
22	Machine Learning	Social Psychology	684.3295	0.9479	-1303.2903	2671.9492	False
23	Natural Language Processing	Psycholinguistics	1231.6500	0.9632	-2592.6064	5055.9064	False
24	Natural Language Processing	Psychology	828.0000	0.9963	-2996.2564	4652.2564	False
25	Natural Language Processing	Social Psychology	1689.9355	0.0197	180.2299	3199.6411	True
26	Psycholinguistics	Psychology	-403.6500	1.0000	-5615.2330	4807.9330	False
27	Psycholinguistics	Social Psychology	458.2855	0.9999	-3390.7237	4307.2946	False
28	Psychology	Social Psychology	861.9355	0.9954	-2987.0737	4710.9446	False

Table 27: Tukey's HSD Pairwise Comparison Results for Social Support. The columns "Upper" and "Lower" refer to the confidence interval bounds for the difference between the means of the two groups.

Expert	Successes	Failures
Expert 0	49 607.86	18 485.14
Expert 1	49 797.19	18 295.81
Expert 2	49 820.12	18 272.88

Table 28: Global Contingency Table of Successes and Failures by Expert Position (from 0 to n).