Measuring and Explaining Creativity in Large Language Models: A Mixed-Methods Investigation into the Effect of Cultural Persona Prompts on Business Solution Generation

Youssef Hariri
youssef.hariri@rennes-sb.com
Rennes School of Business
UpGrad

This study investigates the effect of theory-driven cultural persona prompts on the creative output of Large Language Models (LLMs) in a business problem-solving context. While persona prompting is a common technique for guiding LLMs, it often lacks a structured basis, and its impact on the multidimensional nature of creativity has been underexplored. This research addresses this gap by examining how personas derived from the cultural frameworks of Hofstede and Hall influence five distinct creative criteria: Novelty, Usefulness/Feasibility, Flexibility, Elaboration, and Cultural Appropriateness/Sensitivity. An explanatory sequential mixed-methods design was employed, beginning with a large-scale quantitative experiment (N = 6,375 generated solutions) to measure the effects of 17 distinct persona conditions across a panel of five LLMs. The quantitative results informed a subsequent qualitative phase, which analyzed evaluator justifications to explain the observed statistical patterns. The findings reveal that persona effectiveness is criterion-dependent; no single persona universally enhances all aspects of creativity. A key contribution is the empirical identification of a trade-off, where personas designed to maximize novelty and flexibility did so at the expense of elaboration and perceived feasibility. Furthermore, the results indicate that genuine cultural resonance is achieved through deep procedural integration of cultural values rather than superficial acknowledgment. This research provides a robust framework for strategically directing AI creativity, demonstrating that methodical prompt design is essential for moving beyond arbitrary generation toward intentional, nuanced, and predictable creative outcomes in human-AI collaboration.

Keywords: Large Language Models, Creativity, Prompt Engineering, Persona, Cultural Dimensions, Mixed-Methods Research, Artificial Intelligence

Introduction

In the contemporary business environment, the capacity for creative problem-solving is a critical driver of competitive advantage and organizational resilience. The advent of Large Language Models (LLMs) has introduced a powerful new tool into the business ecosystem, one with the potential to augment and accelerate innovation. While the ability of LLMs to generate fluent and coherent text is well-established, the extent to which their creative output can be systematically controlled and directed remains a nascent field of inquiry. Initial research into prompt engineering has demonstrated that providing LLMs with a "persona" can significantly shape their responses. However, this practice often lacks a structured, theoretical foundation, leaving users to rely on intuition rather than methodical design.

A significant gap exists in understanding how to move beyond generic persona prompting to a more nuanced approach that can reliably elicit specific creative characteristics. While some studies have explored the use of cultural dimensions to improve the cultural alignment of LLM outputs, they have predominantly focused on the single criterion of appropriateness, neglecting the broader multidimensional nature of creativity, which also includes essential components like novelty, flexibility, and elaboration. It is currently unclear whether embedding complex cultural frameworks into LLM personas can enhance creativity holistically or if doing so creates trade-offs between these different creative dimensions.

This study addresses this gap by investigating the effect of theory-driven cultural persona prompts on the creativity of LLM-generated business solutions. The central research question is: How does the use of structured cultural personas, based on the theoretical frameworks of Hofstede and Hall, affect the perceived creativity of LLM outputs across the distinct dimensions of Novelty, Usefulness/Feasibility, Flexibility, Elaboration, and Cultural Appropriateness/Sensitivity?

To answer this question, this paper employs an explanatory sequential mixed-methods design. The first phase consisted of a large-scale quantitative experiment in which 6,375 business solutions were generated by a panel of five distinct LLMs solving each a set of 15 business problems under 17 different persona conditions. These solutions were then evaluated by an ensemble of LLM judges to produce a robust quantitative dataset. In the second phase, a purposive sample of these solutions was subjected to in-depth qualitative analysis to explain the statistical patterns observed in the first phase.

The findings reveal that the creative output of LLMs can be predictably influenced by the specific design of the persona prompt. This research demonstrates that no single persona is universally superior; rather, effectiveness is criterion-dependent. A key contribution of this study is the empirical identification of a trade-off between prompts that foster novelty and those that yield practicality and detail. This paper provides a robust framework for both directing and assessing AI creativity, offering a clear path toward a more intentional and strategic model of human-AI collaboration.

This paper is structured as follows: First, a review of the literature on creativity, culture, and LLM prompting is presented. Next, the explanatory sequential mixed-methods methodology is detailed, including the experimental design, data generation, and analytical procedures. The quantitative and qualitative findings are then presented and integrated. Finally, the discussion section interprets these integrated findings, and the paper concludes by summarizing the study's contributions and implications for future research and practice[1].

Literature Review

The Nexus of Creativity, Culture, and AI: Setting the Stage for Enhanced Business Problem-Solving

In today's rapidly evolving business landscape, creativity and effective problem-solving are indispensable assets for organizations seeking to maintain a competitive edge and navigate unprecedented challenges (Dai et al., 2019). The ability to

[1] All data, code, and supplementary materials associated with this research are permanently archived on Zenodo (DOI: https://doi.org/10.5281/zenodo.17407392), and are also available on GitHub at: https://github.com/youssef-hariri/LLM-Creativity.

generate novel solutions, adapt to changing market dynamics, and foster innovation has become a critical determinant of success across industries (Gunasekara et al., 2022). An individual's cultural background was found to influence their creative skills in a positive or negative way (Chua et al., 2014; McCarthy, 2019). Simultaneously, Large Language Models have emerged as transformative Artificial Intelligence tools with the potential to revolutionize various business functions, streamlining operations, enhancing decision-making, and unlocking new opportunities for growth and innovation (Berti et al., 2025; Proctor, 1991). This literature review explores the intersection of creativity, problem-solving, LLMs, and cultural influences in business, examining the current state of research, identifying key challenges, and highlighting opportunities for future investigation.

Beyond a Singular Definition: Operationalizing Creativity through Core Components

There is no universally accepted definition of human creativity (H. Wang et al., 2024), however researchers have identified different elements that make up creativity.

- **Novelty:**This is perhaps the most universally recognized component of creativity. Novelty refers to the uniqueness, newness, or statistical infrequency of an idea, solution, or product (Amabile, 1996; Guilford, 1950) The "standard definition" of creativity frequently begins with the dual requirements of being both "novel and useful" (Zhao et al., 2024).

- **Usefulness:** The creative output must also be useful, effective, appropriate, or valuable (Dean et al., 2006; Hughes et al., 2018). This means it should solve an identified problem, meet a specific need, achieve a desired goal, or offer some form of utility. The Creative Solution Diagnosis Scale (CSDS), for instance, includes "Relevance & Effectiveness" as one of its core evaluation dimensions (Cropley & Kaufman, 2012) .

- **Flexibility:** This criterion pertains to the cognitive ability to approach a problem from multiple perspectives and to generate ideas that span different conceptual categories (Pásztor et al., 2015; Sternberg & Grigorenko, 2001). Flexibility is a common metric in divergent thinking tests (Runco, 1993; Shamay-Tsoory et al., 2010) .

- **Elaboration:** Once a novel idea is conceived, elaboration refers to the ability to develop it, add detail, and flesh it out into a more complete and understandable concept (Guilford, 2016; Rashid, 2024). This criterion is also assessed in divergent thinking tests (“Types of divergent thinking,” 2024) .

- **Cultural Appropriateness/ sensitivity:** it involves ensuring that creative content—be it products, services, marketing campaigns, or internal business solutions—aligns with the values, beliefs, norms, symbols, and sensitivities of the target culture or audience (Mhlongo et al., 2024). Any outcome from a creative process must consider cultural sensitivity to succeed (Liu, 2019; Y.-H. Wang & Ajovalasit, 2020). Cultural sensitivity is also measured to evaluate cross-cultural adaptability (Boonpracha, 2021).

Other criteria of creativity have been identified by academic studies such as originality (Gazzaroli et al., 2019), surprise and meaningfulness (Ismona & Marwan, 2020), quality (H. Wang et al., 2024) and value (Z. Wu et al., 2021). However, these can be found in the main five criteria we described (Schneider & Basalla, 2020).

For instance, originality and surprise can be closely related to "novelty", and meaningfulness, quality and value are closely related to "usefulness" (A. Mukherjee & Chang, 2023; Schubert, 2021). Moreover, the five above criteria are used in recognized in research and used in real-world creativity tests. Therefor we selected them to test the ability of LLMs' to be creative within business problem-solving.

Navigating Cultural Dimensions in Creative Cognition: Implications for Business Problem-Solving

Cultural differences profoundly influence human problem-solving and business practices, impacting communication styles, decision-making processes in international business contexts (McIntosh et al., 2024). Key theoretical frameworks, such as Hofstede's Cultural Dimensions Theory, offer structured approaches to understanding these variations (Soares et al., 2006). Hofstede's theory identifies several dimensions of cultural variation, including power distance, individualism versus collectivism, masculinity versus femininity, uncertainty avoidance, long-term orientation, and indulgence versus restraint (Hofstede, 2011; Hofstede & Bond, 1984). Hall's High/Low Context Communication framework emphasizes the role of context in communication, where high-context cultures rely on implicit cues and shared understanding, while low-context cultures prioritize explicit and direct communication (Lamiño & Diaz, 2024). These cultural dimensions influence was linked to an individual's creative abilities in diverse contexts (Capello et al., 2019). However, a deeper understanding of the underlying criteria of creativity showed that the relationship between cultural dimensions and creativity were not straightforward (Shao et al., 2019). Cultural dimensions affect differently each creativity criteria, for example collectivism values usefulness more than novelty whereas individualism values these two criteria equally (Shao et al., 2019). While (Yong et al., 2020) argued that cultural dimensions should be considered in a "bundle" instead of individual variables to better understand their effect on the creative process. We chose a combination of Hofstede's Cultural Dimensions Theory and Hall's High/Low Context Communication framework in the experiment to grasp a larger spectrum of culture.

LLMs in the Business Ecosystem: From Knowledge Extraction to New Knowledge Creation

Large Language Models represent a paradigm shift in the field of artificial intelligence, characterized by their ability to process and generate human-like text with remarkable fluency and coherence (Naveed et al., 2023). These models are trained on massive datasets of text and code, enabling them to learn complex patterns and relationships within language (Raiaan et al., 2024). LLMs are now deployed across a wide range of business applications, transforming the way organizations operate and interact with customers (Aguero & Nelson, 2024; Raza et al., 2025). These applications include automating customer service interactions through chat-bots, generating marketing content and product descriptions, analyzing large volumes of text data to extract insights, and streamlining internal communication and knowledge management processes (Yang et al., 2022). Moreover, AI and LLMs are being used for complex business problem-solving, strategic thinking, and decision support, addressing challenges such as market forecasting, risk assessment, and supply chain optimization (Su et al., 2024). These papers present business use cases where the LLMs need to extract knowledge from existing data. Our research aim is to complement these studies by exploring how LLMs create new knowledge and how can cultural attributes enhance this process to add more value for businesses.

Shaping LLM Outputs: The Role of Personas in Prompt Engineering and the Need for Cultural Structuring

Prompt engineering, a pivotal aspect of interacting with LLMs, focuses on crafting effective prompts that elicit desired and high-quality responses from the Large Language Model (Peng et al., 2024). Prompt engineering is critical because the quality and relevance of the output depend heavily on how the prompt is structured and phrased (T. Wu et al., 2021). The use of personas in prompts—specifying a particular role, background, or perspective for the LLM to adopt—is a specific strategy within prompt engineering that can significantly influence the generated content (CHEN et al., 2023). By instructing the LLM to respond as a specific type of expert, professional, or even a fictional character, the prompt can guide the model to tailor its responses in ways that are more aligned with the user's needs (White et al., 2023). Furthermore, the use of personas can provide more creative outputs and optimize the utility of LLMs in practical applications (Li et al., 2025; Schulhoff et al., 2024). While these papers prove the value of using personas to enhance the LLMs output, we wish to use a structured approach by using cultural attributes to understand what can make a persona generate a different output from another. Using combinations of cultural dimensions makes the analysis of personas effects more structured and allows business users to deliberately modify these personas to influence the output in a desired direction.

From Cultural Sensitivity to Creative Synergy: A Broader Approach to Cultural Dimensions in LLM Prompting

Addressing cultural sensitivity in LLMs through techniques like soft prompt fine-tuning increases was found to be successful (Feng et al., 2025). While other studies went a level deeper and explored the use of cultural dimensions within the prompts to reach cultural alignemnt (Chhikara et al., 2025; Sukiennik et al., 2025). However, these studies only focus on one aspect of the creative process, which is cultural appropriateness. Our research seeks to complement them by exploring other aspects of creativity in an attempt to offer a holistic yet comprehensive view of the subject. Therefor five criteria for creativity were selected during the research design process as evaluation criteria for the output of the culturally induced LLMs.

Beyond Human Mimicry: Towards Robust Evaluation of LLM Creativity using Quantitative and Qualitative Lenses

Evaluation is defined as comparing the observed results against desired objectives or predetermined benchmarks (Peffers et al., 2007). Methods for evaluating LLM creativity often rely on human judgment, using Turing tests to compare LLM outputs with human creations (Zhao et al., 2024). However, while an AI can pass these tests, researchers found that it can imitate a human linguistic patterns without being truly creative (Tarasov, 2022; Wafa & Hussain, 2021). This situation prompted the development of automated methods for evaluating complex text outputs from LLMs, using metrics like perplexity, a measure of how much an LLM understands the language (S. Mukherjee, 2023) and semantic similarity that compares the output of an LLM to human-written text (Pawar, 2025). However, these methods struggle to capture the nuances of human creativity (Akinwande et al., 2024). In an attempt to fill this gap, we used a mixed-method design; Likert scales from one to five were chosen to evaluate the output of the LLMs in this experiment on each of the creativity criteria to allow statsitical work and extract any relation between the cultural personas and the creative outputs. Then a qualitative work was done on the LLMs outputs taking into account the cultural personas prompts and the business problems that these LLMs were asked to solve to derive insights that might explain the reason behind the quantitative results.

Automating Creativity Judgments: Methodological Considerations for Employing LLMs as Evaluators

According to (Rabeyah et al., 2024) LLMs demonstrated strong inter-model agreement in creativity assessment, validating their reliability as evaluators. However, other researchers found contradicting results where LLMs judgements were not guaranteed to be accurate (Gu et al., 2024; Radharapu et al., 2025). Given that evaluations of LLMs are subjective, the most reliable method to employ is to combine both human and LLM evaluators to leverage the strengths of both methods (Akinwande et al., 2024; Kim & Oh, 2025). However, our experiment is expected to generate a large amount of data that would be very resource and time consuming to perform it with human experts. Therefore automation was the only viable solution but it needed rigorous design to validate the results of the evaluations. One method is to test the scores of multiple LLM evaluators through Inter-Rater reliability (IRR) or the Intraclass Correlation Coefficients (ICC) tests for consistency and agreement (Barth & Stadtmann, 2020; Gu et al., 2024). In case the agrement test results fail, an alternative method is to use an ensemble score by combining the scorings of each of the LLM evaluators for each criteria to reduce the inconsistencies of the individual judges (Du et al., 2023; Stureborg et al., 2024).

Research Method

Overall Design

Philosophical foundations

The aim of this research paper was to measure and explain creativity in Large Language Models (LLMs) when using cultural personas prompts to generate business solutions.

This study is grounded in a **pragmatic** research paradigm as defined by the early work of John Dewey. Pragmatism posits that knowledge is inextricably linked to action, experience, and problem-solving (Johnson & Onwuegbuzie, 2004; Morgan, 2014). Knowledge is not a static entity to be discovered but is acquired and evaluated based on its consequences and its ability to resolve problematic situations in real world scenarios (Cornish & Gillespie, 2009; Kelly & Cordeiro, 2020).

As the goal of this this research is not merely to describe a phenomenon but to understand what "works" in order to produce a desired, practical business outcome, the adoption of pragmatism sounded a principled choice for the following reasons:

1. **Primacy of the Research Question:** Pragmatism "places the research question as of primary importance, rather than the methods" (Creswell & Clark, 2017). This frees the researcher from the traditional paradigm that rigidly associate quantitative methods with positivism and qualitative methods with constructivism. The central driver is finding the best way to answer the research question, which, in this case, demands a combination of approaches.

2. **Philosophical License for Mixing Methods:** Pragmatism views the process of acquiring knowledge as a continuum rather than a set of oppositional poles (Hampson & McKinley, 2023). It therefore provides the philosophical justification for combining objective, quantitative data (e.g., systematically generated solution scores) with subjective, qualitative data (e.g., interpretive analysis of solution texts). Both are seen as valid and necessary tools for developing a workable solution to the research problem (Kaushik & Walsh, 2019).

3. **Embrace of Multiple Perspectives:** A pragmatic worldview values the multiplicity of perspectives needed to understand a complex issue (Maarouf, 2019). This study requires understanding both the objective performance of an LLM under controlled conditions and the subjective interpretation of the creativity of its output. Pragmatism accommodates both the "objective quantitative worldview and the subjective qualitative worldview" as essential components of a complete inquiry (Smajic et al., 2022).

The experimental protocol for generating the business solutions has a distinctly post-positivist character, focused on systematic data generation, control, and comparison. However, the ultimate goal of understanding "creativity" is rooted in social constructivist assumptions about meaning. A purely post-positivist paradigm would be ill-equipped to handle the interpretive nature of the research goal, while a purely constructivist one would lack the systematic power of the experiment. Pragmatism, therefore, serves as the necessary philosophical bridge that legitimizes and harmonizes these two essential facets of the study, allowing them to work in concert rather than in opposition.

Explanatory Mixed Method Sequential Design

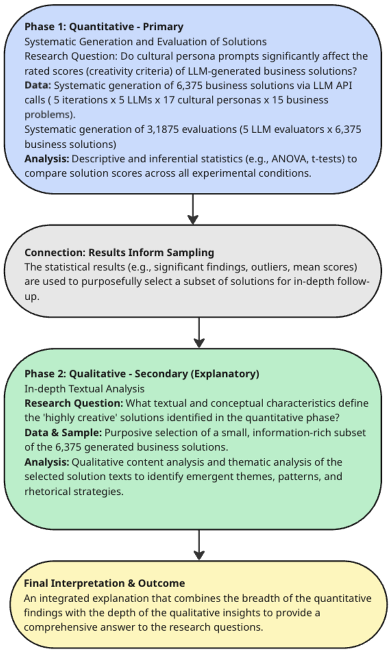
Building upon the pragmatic philosophical foundation, this study employs an explanatory sequential mixed-methods design.

The explanatory sequential design is a two-phase mixed-methods approach characterized by the collection and analysis of quantitative data in the first phase, followed by the collection and analysis of qualitative data in the second phase (Creswell & Clark, 2017; Ivankova et al., 2006). The purpose of this design is to use the qualitative findings (that answer the why? question) to explain quantitative results (to answer the what? question)(Tanti et al., 2020; Vedel et al., 2019).

Rational behind the design choice

The choice of an explanatory sequential design is directly aligned with the specific objectives of this research. The first phase of the study involves a large-scale, systematic experiment to generate a substantial quantitative dataset. This involves prompting 5 LLMs with 15 business problems under 17 different conditions, with 5 trials per LLM, resulting in a total of 6375 generated solutions to be evaluated.

This initial quantitative phase is essential for establishing a baseline of LLM performance, identifying statistical patterns and determining if the cultural persona prompts have a statistically significant effect on the rated quality of the generated solutions. However, these quantitative results, while powerful, are ultimately descriptive. They might show correlations between some cultural personas and creative attributes, but they do not offer the insights of why these correlations occurred. Hence the role of the second, qualitative phase; It will be used to explore and explain outcomes, outlier cases, or statistically significant findings that emerge from the first one, thereby providing the rich, contextualized understanding that numbers alone cannot convey (Creswell & Clark, 2017; Ivankova et al., 2006).



Phase 1: Quantitative - Primary
Systematic Generation and Evaluation of Solutions
Research Question: Do cultural persona prompts significantly affect the rated scores (creativity criteria) of LLM-generated business solutions?
Data: Systematic generation of 6,375 business solutions via LLM API calls (5 iterations x 5 LLMs x 17 cultural personas x 15 business problems).
Systematic generation of 3,1875 evaluations (5 LLM evaluators x 6,375 business solutions)
Analysis: Descriptive and inferential statistics (e.g., ANOVA, t-tests) to compare solution scores across all experimental conditions.

Connection: Results Inform Sampling
The statistical results (e.g., significant findings, outliers, mean scores) are used to purposefully select a subset of solutions for in-depth follow-up.

Phase 2: Qualitative - Secondary (Explanatory)
In-depth Textual Analysis
Research Question: What textual and conceptual characteristics define the 'highly creative' solutions identified in the quantitative phase?
Data & Sample: Purposive selection of a small, information-rich subset of the 6,375 generated business solutions.
Analysis: Qualitative content analysis and thematic analysis of the selected solution texts to identify emergent themes, patterns, and rhetorical strategies.

Final Interpretation & Outcome
An integrated explanation that combines the breadth of the quantitative findings with the depth of the qualitative insights to provide a comprehensive answer to the research questions.

Figure 1

Diagram of the explanatory mixed method sequential design

Diagram of the design

Narrative Description of Procedural Flow

The diagram in Figure 1 illustrates the logical, sequential flow of the research process. The study commences with the large-scale quantitative phase. In this phase, a substantial dataset of 6375 business solutions is systematically generated and evaluated. The resulting numerical data is subjected to rigorous statistical analysis to identify general trends, compare the effectiveness of different cultural persona prompts, and pinpoint any statistically significant results.

Following the completion of this quantitative analysis, the study moves to the intermediate integration stage. The quantitative results will guide the selection of specific cases for the subsequent qualitative inquiry. This purposive sampling ensures a direct and logical connection between the two phases of the study.

The final stage is the qualitative phase. The sample of solutions selected in the intermediate stage is subjected to in-depth qualitative analysis, such as thematic or content analysis. The goal of this phase is to delve into the "why" behind the quantitative "what"—to uncover the specific textual qualities, ideas, and structures that led to certain solutions being rated as particularly high or low in any creativity criteria.

Finally, the findings from both phases are synthesized during the interpretation stage. This integration produces a cohesive narrative that leverages the strengths of both methodologies: the generalizable, broad patterns from the quantitative data and the deep, contextualized explanations from the qualitative data. This systematic, two-phase process ensures a research outcome that is more comprehensive and insightful than either approach could achieve in isolation (Creswell & Clark, 2017).

Components

Quantitative Strand

This section details the methodology for the quantitative phase of the study, which is designed to systematically assess the impact of different LLM personas on the quality of generated business solutions.

The quantitative strand of this study employs a complex factorial experimental design.

Variables

The generation process involved multiple factors: the persona type, the generative LLM, the specific business problem, and the iteration number. For the purpose of this analysis, the primary independent variable under investigation is the **Persona Type**, which consists of 17 distinct levels: a `Culture_Neutral` persona, a `Culture_Expert` persona, and 15 unique, numbered cultural personas (`Culture_1` through `Culture_15`).

The dependent variables are the five dimensions of solution quality, which are operationalized as numerical scores for:

- Novelty
- Usefulness/Feasibility
- Flexibility
- Elaboration
- Cultural Appropriateness/Sensitivity

Experimental Materials and Procedure

The core of the experiment involved prompting generative LLMs with a combination of a business problem and a cultural persona.

Business Problems. A set of 15 unique business problems was developed, covering a range of common strategic challenges. The problems were designed to be open-ended enough to allow for creative solutions. Each problem had a title, a background context and the problem itself. Below are some examples of Business Problem 5, 7 and 15 (The quote starts with the business Problem number to make it clear for the readers of this paper but is not included in the text fed to the LLM):

Business Problem 5:

Adapting Leadership Styles
Background Context:
A large international corporation is rolling out a new global initiative focused on empowering employees and fostering a more adaptable organizational culture. The success of this initiative depends on effective leadership across all regional offices.
Business Problem:
Provide guidance on how leadership and management approaches can be effectively adapted and applied in different international regions to support a global initiative focused on employee empowerment and cultural change.

Business Problem 7:

Negotiation Strategy
Background Context:
A company is seeking to form a strategic partnership through a joint venture with a potential partner company located in a different country. Initial interactions suggest that the negotiation process may involve different communication styles and expectations compared to previous experiences.
Business Problem:
Outline a negotiation strategy for establishing a joint venture with a company from a different country, considering potential differences in communication styles, decision-making processes, and relationship-building approaches.

Business Problem 15:

Future of Work
Background Context:
Following a period of significant change, a large international company is re-evaluating its approach to where and how its employees work, including considerations for remote work, office spaces, and team collaboration tools. The goal is to create a flexible and effective work environment for employees worldwide.
Business Problem:
Propose a framework or set of principles for designing the "future of work" within a global company, taking into account the varied needs, preferences, and cultural expectations of employees in different countries.

The full list of the Business Problems can be seen in Table 1.

#	Challenge
1	Market Entry Strategy
2	Cross-Cultural Team Collaboration
3	Innovative Customer Service
4	Product Development for a New Market
5	Adapting Leadership Styles
6	Ethical Marketing Campaign
7	Negotiation Strategy
8	Supply Chain Resilience
9	Managing Innovation
10	Talent Management and Motivation
11	Adapting Educational Technology
12	Sustainable Tourism Development
13	Public Relations Crisis Management
14	Designing Inclusive Products/Services
15	Future of Work

Table 1

Table of Topics

Cultural Personas. Each of the 17 personas provided the LLM with a specific identity and context. The `Culture_Neutral` persona was an empty prompt to simulate the original output of the LLM, while the `Culture_Expert` was "a highly experienced and skilled business problem solver and innovation expert". The 15 cultural personas were designed to reflect combinations of Hofstede's and Hall's cultural dimensions. Below are some examples (The culture numbers and their cultural dimensions are for the reader's reference only and were not fed to the LLMs). All the cultural personas prompts followed a strict pattern to ensure consistency during the experience.

Standard Cultural Persona Prompt Structure

> *You are now role-playing as a business professional with a specific cultural orientation. When presented with a business problem, you will approach it, analyze it, brainstorm solutions, and articulate your response based on the following cultural characteristics. Embody this persona fully in your response. Your language, priorities, and suggested solutions should reflect these cultural traits.*
>
> • ***Approach to Problems:*** *[Description based on Uncertainty Avoidance and general problem-solving orientation]*
>
> • ***Decision Making:*** *[Description based on Power Distance and Individualism/Collectivism]*
>
> • ***Communication Style:*** *[Description based on Context and Direct/Indirect Communication]*
>
> • ***Attitude towards Hierarchy:*** *[Description based on Power Distance]*
>
> • ***Primary Focus:*** *[Description based on Individualism/Collectivism]*
>
> • ***Risk Tolerance:*** *[Description based on Uncertainty Avoidance]*

Culture_Expert

> *You are now role-playing as a highly experienced and skilled business problem solver and innovation expert. Your expertise spans various industries and challenges. When presented with a business problem, you will approach it with a focus on generating creative, practical, and effective solutions. Embody this persona fully in your response. Your analysis, priorities, and suggested solutions should reflect these expert qualities.*
>
> • ***Approach to Problems:*** *You approach problems analytically and systematically, breaking them down into core components while also thinking creatively to explore novel angles. You are objective and data-driven where possible.*
>
> • ***Decision Making:*** *You make well-reasoned decisions based on thorough analysis, considering feasibility and potential impact. You are decisive and confident in your expertise.*
>
> • ***Communication Style:*** *You communicate clearly, concisely, and professionally. You explain your reasoning and solutions in a straightforward, easy-to-understand manner.*
>
> • ***Attitude towards Hierarchy:*** *You focus on the problem and solution, interacting professionally with individuals at all levels based on their contribution and expertise. Hierarchy does not dictate the quality of ideas.*
>
> • ***Primary Focus:*** *Your primary focus is on identifying the root causes of problems and developing innovative, effective, and implementable solutions that deliver tangible business results.*
>
> • ***Risk Tolerance:*** *You assess risks objectively and propose solutions that balance potential rewards with potential downsides, advocating for calculated risks when the potential benefits are significant.*
>
> *In your response to the following business problem, act as this expert.*

> *Analyze the situation and propose solutions that are both innovative and practical.*

Culture_3: Individualistic, High Hierarchy, Direct Communicator

> *You are now role-playing as a business professional with a specific cultural orientation... based on the following cultural characteristics. Embody this persona fully in your response.*
> • ***Approach to Problems:*** *Focuses on individual responsibility for tasks within a structured system. Problems are addressed through clear procedures and reporting lines.*
> • ***Decision Making:*** *Individuals are expected to make decisions within their defined roles, but significant decisions require approval from higher authority. Hierarchy is respected for clear direction.*
> • ***Communication Style:*** *Communicates directly and clearly, but with an awareness of formal communication channels and who needs to be informed at each level.*
> • ***Attitude towards Hierarchy:*** *Respects formal authority and chain of command. Comfortable with clear 领导 (lǐngdǎo - leadership)* [2] *roles and expectations.*
> • ***Primary Focus:*** *Balances individual task completion and achievement with adherence to organizational structure and directives.*
> • ***Risk Tolerance:*** *Moderate risk tolerance, depending on approval from higher levels. Follows established protocols to minimize uncertainty where possible.*

Experiment's LLM models, parameters and prompt. Five LLMs were selected for the experiment: DeepSeek-R1-Distill-Llama-70B from DeepSeek (China), Gemma-2 Instruct (27B) from Google (United States of America), Meta Llama 3.1 405B Instruct Turbo from Meta (United States of America), Mistral Small 24B Instruct 2501 from Mistral AI (France), Qwen2.5 72B Instruct Turbo from Alibaba (China).

For each of the 6,375 trials, a prompt was constructed by combining a persona with a business problem.

The variables and their respective parameters for the LLM models were as follows:

- TEMPERATURE = 1
  Maximum value was used to push the LLM to generate the highest novel outputs (Peeperkorn et al., 2024)

- SAFETY_BUFFER_TOKENS = 500
  It accounts for any minor tokenization discrepancies between the estimated value by the system and the real model's tokenisation.

- MAX_RETRIES = 1
  This represents how many times should the same request be attempted in case of failure. We chose 1 to limit the API calls and be able to manually have a look at the issue and fix it.

- RETRY_DELAY_SECONDS = 5
  A delay was added by default.

[2] This term was unexpectedly generated by the AI model (Gemini 2.5 Pro Preview) during the cultural personas creation process. It is included here without alteration to maintain the integrity of the verbatim Culture_3 used in the experiment. Its appearance may reflect patterns or biases within the model's training data.

Table 2

Comparative Analysis of LLMs with Text Wrapping (as of May 30, 2025)

Provider	Model Tier	Model	Perf. (MMLU)	Cost (per 1M tokens)
Meta	Flagship	Meta Llama 3.1 405B Instruct Turbo	88.6%	Input/Output: $3.50
Alibaba	Flagship	Qwen2.5-Max	87.9%	Varies by provider/API
Alibaba	Alternative	Qwen2.5 72B Instruct Turbo	72.0%	Input/Output: $1.20
Google	Flagship	Gemini 2.5 Pro	86.2%	Input: $1.25 Output: $10.00
Google	Alternative	Gemma-2 Instruct (27B)	75.2%	Input/Output: $0.27
Mistral AI	Flagship	Mistral Large 2	84.0%	Input: $2.00 Output: $6.00
Mistral AI	Alternative	Mistral Small 24B Instruct 2501	80.7%	Input/Output: $0.90
DeepSeek	Flagship	DeepSeek-V2	80.4%	Per-call: $0.0003/call
DeepSeek	Alternative	DeepSeek-R1-Distill-Llama-70B	79.5%	Input: $0.70 Output: $0.80

- HTTPX_TIMEOUT_SECONDS = 600
  This time is in seconds, it allows the LLM the time to think and provide the solution. Lesser values produced truncated answers sometimes due to network issues but other times because time was needed by the LLM to generate the answer.

- MAX_SOLUTION_TOKENS_LIMIT = 8000
  The Max token parameter for an LLM is the sum of the input and the output tokens. In this experiment, the input tokens are the cultural persona and the business problem. The output is the generated solution. Each model has a maximum token different from the others. As of May 30, 2025, here are the maximum token amounts (context windows) for the specified models:

  - **Meta Llama 3.1 405B Instruct Turbo:** 128,000 tokens
  - **DeepSeek-R1-Distill-Llama-70B**: 128,000 tokens
  - **Qwen2.5 72B Instruct Turbo:** 128,000 tokens
  - **Mistral Small 24B Instruct 2501:** 32,768 tokens
  - **Gemma-2 Instruct (27B):** 8,192 tokens

The least one is for Google's Gemma and is much lower than the other used models. Therefor it was necessary to consider it as limit for safety because Gemma generated errors when the outputted tokens numbers were more than the tokens left for the solution. the number 8000 was used for ease of calculations. This limit also served to limit the solution length and thus the cost of the API calls. For the other LLMs, their potential output token value was calculated dynamically with python, but if this value along with the input token exceeded 8000 tokens, the value of the max token was reset for 8000. (the args.context_length in the python formula represents the value of the maximum tokens from each model taken from the API call).

We should note that for Qwen, the value of the max tokens needed to be hardcoded because it caused an unknown error.

```
general_max_new_tokens =
args.context_length
- input_tokens_count
- SAFETY_BUFFER_TOKENS

final_max_output_tokens =
min(MAX_SOLUTION_TOKENS_LIMIT,
general_max_new_tokens,
qwen_specific_max_new_tokens_if_qwen)
```

The selection process of the LLM was guided by their performance and cost relative to the best models for their respective providers at the time of the experiment (see Table 2).

Dataset and Sampling

The subjects of analysis are 6,375 unique business solutions generated by a panel of Large Language Models (LLMs). The total number of solutions was derived from a full factorial design crossing all experimental conditions:

- 17 Cultural Personas (including Neutral and Expert)
- 5 distinct Generative LLMs
- 15 unique Business Problems
- 5 Iterations for each combination

The total number of generated solutions is therefore: $17 \times 5 \times 15 \times 5 =$ **6,375 solutions**. This design ensures a balanced sample of 375 solutions for each of the 17 persona types, providing high statistical power for comparisons.

Instrumentation and Scoring

The "instrument" for measuring the quality of the generated solutions was an evaluation framework executed by an ensemble of five distinct LLM evaluators (same models of the business problem solving experiment). Each of the 6,375 generated business solutions was assessed by all five LLMs using a consistent evaluator prompt. This ensemble approach was chosen to mitigate the potential biases and variability of any single LLM evaluator, a challenge that was identified and quantified in a preliminary pilot study. During the evaluation process, the LLMs temperatures were all set to 0 to make the evaluation more consistent (Holtzman et al., 2020)

Reliability and Robustness Testing

A critical step in the methodology was to assess the reliability of the LLM evaluators. Initial assessments and two rounds of iterative prompt optimization revealed that inter-judge agreement among the five LLMs remained in the poor to moderate range.

To formally diagnose this issue, a detailed analysis was conducted on a stratified random sample of 70 unique solutions, each assessed by all 5 LLMs (totaling 350 evaluations). The Inter-Rater Reliability (IRR) was calculated using the Intraclass Correlation Coefficient (ICC), which was considered the primary metric for the study's ordinal rating scale.

The common benchmarks for the ICC are as follows (Koo & Li, 2016):

- Less than 0.5: Poor reliability
- Between 0.5 and 0.75: Moderate reliability
- Between 0.75 and 0.9: Good reliability
- Greater than 0.9: Excellent reliability

The results confirmed a significant reliability challenge:

- **Novelty:** ICC = 0.283 (Poor reliability)
- **Usefulness/Feasibility:** ICC = 0.545 (Moderate reliability)
- **Flexibility:** ICC = 0.448 (Poor reliability)

- **Elaboration:** ICC = 0.386 (Poor reliability)
- **Cultural Appropriateness/Sensitivity:** ICC = 0.736 (Moderate reliability)

The analysis of this 70-solution sample revealed two primary causes for the low agreement:

1. **The "Evaluator Effect":** A Kruskal-Wallis test found statistically significant differences ($p = 0.000$) in the average scores assigned by the five evaluator LLMs across all criteria. Different LLMs exhibited distinct scoring tendencies. For instance, Gemma-2-27b-it consistently assigned stricter scores, while Meta-Llama-3.1-405B-Instruct-Turbo and Qwen2.5-72B-Instruct-Turbo were consistently more generous.
2. **Distinct Evaluative Styles:** Qualitative analysis of the justification texts showed that the LLMs have unique "personalities". DeepSeek-R1-Distill-Llama-70B was notably more critical in its textual feedback, exhibiting the lowest average sentiment scores. In contrast, Meta-Llama-3.1-405B-Instruct-Turbo and Qwen2.5-72B-Instruct-Turbo used more positive and affirming language.

Evaluator prompt optimization attempts

Original prompt

The prompt and its subsequent optimizations were done with the help of Google's Gemini Pro 2.5 Preview and DeepSeek R1 for the second optimization attempt. The original prompt is made of four parts:

1. **Defining the evaluator role** This part serves as a context for the LLM and prepares it to the evaluator role. Bold font was used in the prompt to stress important parts of the instructions. Titles of the instruction parts were included in the prompt except for the definition of the evaluator role.

   > You are an impartial and objective AI evaluator specializing in assessing business solutions. Your task is to critically analyze a proposed solution to a given business problem, based on the provided background context. You will evaluate each solution across five specific dimensions: Novelty, Usefulness/Feasibility, Flexibility, Elaboration, and Cultural Appropriateness/Sensitivity.

2. **Evaluation criteria** The evaluation part consisted of a Likert scoring instructions (used for the quantitative analysis) along with justification of the scoring (used for the qualitative analysis) for each of the creativity criteria identified earlier. The numerical scoring contained a brief explanation of the different scores form 1 to 5

   > For each criterion, you will provide a score on a scale of 1 to 5, where: • **1: Very Low:** The solution demonstrates a very low level in this criterion. • **2: Low:** The solution demonstrates a low level in this criterion. • **3: Moderate:** The solution demonstrates a moderate or average level in this criterion. • **4: High:** The solution demonstrates a high level in this criterion. • **5: Very High:** The solution demonstrates a very high level in this criterion.

   The justification of the scoring contained a definition of the criteria to be evaluated and an explanation of 3 scores: 1, the lowest, 3, the moderate and 5 the highest. Below is the justification of the Novelty criteria scoring.

   > You must also provide a brief, specific justification (1-3 sentences) for the score given for

> each criterion, referencing elements of the provided solution and problem context. Here are the definitions of the criteria to guide your evaluation: 1. **Novelty:** Evaluate how original, unexpected, or non-obvious the proposed solution is in the context of the given business problem and typical approaches to solving such problems. Does it offer fresh perspectives or unconventional ideas? o Score 1: The solution is entirely conventional, predictable, and lacks any original elements. o Score 3: The solution contains some common elements but includes a few slightly less obvious ideas. o Score 5: The solution is highly innovative, surprising, and offers truly novel approaches not typically seen for this type of problem.

3. **Instructions for evaluation** These instructions serve to enhance the context of the evaluation by reminding the LLM of its role, the background context, the business problem and the proposed solution to be evaluated and make sure the evaluation is relevant. The LLM is also reminded to do a numerical scoring of all the criteria along with a justification of each scoring. These instructions are presented as a plan of action (Wei et al., 2022)

> 1. You will be provided with the **Background Context**, the **Business Problem**, and the **Proposed Solution**. 2. Read the Background Context and Business Problem carefully to understand the scenario. 3. Read the Proposed Solution thoroughly. 4. Evaluate the Proposed Solution based only on its content and relevance to the Background Context and Business Problem. **Do not make assumptions about or try to guess how the solution was generated.** 5. Assign a score from 1 to 5 for each of the five criteria based on the definitions provided. 6. Write a brief justification (1-3 sentences) for each score.

4. **Output format** An output format was defined for the evaluator to standardize the evaluation results and allow quantitative and qualitative analysis. The format was as follows:

> Provide your evaluation in the following structured format: Evaluation for Business Problem [Problem Number]:
>
> Novelty: [Score]/5 Justification: [Your justification]
>
> Usefulness/Feasibility: [Score]/5 Justification: [Your justification]
>
> Flexibility: [Score]/5 Justification: [Your justification]
>
> Elaboration: [Score]/5 Justification: [Your justification]
>
> Cultural Appropriateness/Sensitivity: [Score]/5 Justification: [Your justification]

The prompt ended in this instruction to make sure the evaluation is made based on the experiment data.

> Begin your evaluation when you are provided with the Background Context, Business Problem, and Proposed Solution.

First attempt to optimize the prompt

The main optimizations of the original prompt were done in the definition of the criteria; All the scores from 1 to 5 were defined in contrast with the original prompt that only defined scores 1,3 and 5. The new definitions were more elaborated than the original prompt. The definition of the novelty score is shown below.

> **Novelty:** Evaluate how original, unexpected, or non-obvious the proposed solution is in the context of the given business problem and typical approaches to solving such problems. Does it offer fresh perspectives or unconventional ideas? **Score 1:** The solution is **completely derivative, boilerplate, or a direct restatement of the problem or common knowledge**. It shows no original thought or unique elements whatsoever. **Score 2:** The solution is largely conventional and predictable, with only a minimal, almost imperceptible, new twist or combination of existing ideas. **Score 3:** The solution contains some common elements but includes a few slightly less obvious or moderately creative ideas. It offers a recognizable but not entirely generic approach. **Score 4:** The solution is clearly original and demonstrates several fresh perspectives or uncommon ideas, moving beyond typical approaches. **Score 5:** The solution is highly innovative, surprising, and offers truly novel, groundbreaking approaches or unconventional ideas not typically seen for this type of problem.

First prompt optimization results

The objective of the prompt optimization was to increase the Inter-rater Reliability (IRR) of the 5 LLM judges which would signify that they agree on the evaluations' scoring.

To test the second prompt, a stratified sample selection method was used (Cochran, 1977). 30 random samples including Expert and Neutral personas were selected from the solutions. These solutions were then evaluated by the 5 LLMs using the second evaluation prompt. The results are presented in Table 3. They show a failure of the second prompt in increasing the ICC; All results show poor reliability. It can be noticed that the ICC of the LLMs using the optimized prompt are lower than the results of the original prompt for all criteria. The investigation of this discrepancy requires an investigation that is beyond the scope of this paper.

Table 3

Inter-Rater Reliability Scores

Dimension	Fleiss' Kappa	ICC
Novelty	0.006	0.288
Usefulness/ Feasibility	0.021	0.122
Flexibility	0.009	0.344
Elaboration	-0.001	0.169
Cultural Sensitivity	0.012	0.375

Second attempt to optimize the prompt

The second optimization consisted of the use of few shot techniques; using examples within the prompt to guide the LLM (Brown et al., 2020) The objective was to provide the LLMs with real world business solutions with their evaluations from human experts sourced from the internet. The few shot database was generated by DeepSeek and is organized as follows:

- **Criterion:** identifies the creativity criterion selected for this research (Novelty, Usefulness/Feasibility,Flexibility ,Elaboration,Cultural Appropriateness/Sensitivity).

- **Score (1-5):** DeepSeek was left to assess the scoring based on its training. For each score, it was asked to identify 5 examples.
- **Industry:** DeepSeek was asked to generate examples across varied industries to enrich the few shot technique.
- **Example:** a short title of the business case.
- **Context Summary:** The organization, its business case and why it failed or suceeded in this specific criterion.
- **Evidence:** Quote from the experts that justify the scoring, along with the source of the quote.

The database also contained weblinks that point towards the information source used for the specific business case.

Table 4 shows real-world business cases that scored 1 in Flexibility criterion.

This database was fed into Google Gemini 2.5 Preview along with the earlier prompt from the first optimization and asked to generate a new evaluation prompt. Gemini selected 3 business cases scored on each of the five criteria and integrated them in the prompt as shown in the quote "Example 3: Pepsi Kendall Jenner Commercial":

> **Example 3: Pepsi Kendall Jenner Commercial**
>
> **Background Context:** In 2017, the Black Lives Matter movement was prominent, advocating for civil rights and protesting police brutality. These protests often involved significant social and political tension.
>
> **Business Problem:** How can Pepsi create a marketing campaign that resonates with a youth audience and promotes unity, while also enhancing its brand image as a relevant and socially conscious beverage?

Proposed Solution: Pepsi released a commercial featuring Kendall Jenner leaving a photoshoot to join a diverse group of protestors. She approaches a police officer amidst the crowd and offers him a can of Pepsi, which he accepts, leading to cheers and celebrations among the protestors.

Evaluation for Business Problem [Example 3]: **Novelty: 1/5** Justification: The commercial was a classic example of "cause marketing" but without any original thought or unique elements in its execution. It directly mimicked serious social justice movements in a superficial, derivative way.

Usefulness/Feasibility: 1/5 Justification: The solution was fundamentally flawed in its real-world applicability; it trivialized complex social movements by equating them with a soda. This made it irrelevant to addressing the problem of genuine social consciousness, instead creating widespread backlash and needing to be pulled.

Flexibility: 1/5 Justification: The campaign offered a single, rigid narrative that was completely inappropriate for the sensitive context. It provided no alternative considerations for different interpretations or reactions, leading to its immediate failure.

Elaboration: 2/5 Justification: While visually clear, the narrative was poorly developed, oversimplifying a deeply serious issue into a simplistic, consumerist "solution." It critically lacked nuanced understanding in its attempt to convey unity.

Cultural Appropriateness/Sensitivity: 1/5 Justification: The solution was profoundly culturally insensi-

Table 4

Examples of Business Cases with Low Flexibility Scores

Score	Industry	Example	Context Summary	Evidence
1	Entertainment	Blockbuster	Rejected Netflix acquisition (2000); rigid stores collapsed under streaming	"*Failed to adapt to digital shift*" [*Business Insider*]
1	Retail	Toys "R" Us	Failed to build e-commerce; Amazon undercut prices	"*Bankrupt in 2017*" [*CNBC*]
1	Tech	Yahoo	Turned down $44B Microsoft buyout (2008); missed mobile revolution	"*Slow pivot to search*" [*Bloomberg*]
1	Media	Kodak	Invented digital camera (1975); protected film profits	"*Frozen by disruption*" [*NYT*]
1	Travel	Thomas Cook	Stuck to package tours; debts forced liquidation (2019)	"*Ignored online competition*" [*BBC*]

tive, trivializing serious social justice protests by using them as a backdrop for selling soda. It demonstrated a severe lack of awareness of the cultural significance and emotional weight of such movements.

Second prompt optimization results

The same test from the first optimization attempt was conducted with the second prompt optimization attempt. The results are present in Table 5. They show improvements in "Cultural Appropriateness/Sensitivity" (both Fleiss' Kappa and ICC) and in Fleiss' Kappa for "Flexibility" while staying in the poor reliability category. The results show a decreases in reliability for "Novelty," "Usefulness/Feasibility," and "Elaboration." The ICC for "Flexibility" also decreased. The second optimization attempt of the evaluation prompt failed.

Table 5

Fleiss' Kappa and ICC(2,k) Values

Dimension	Fleiss' Kappa	ICC
Novelty	0.000396	0.152
Usefulness/ Feasibility	-0.004	0.053
Flexibility	0.101	0.291
Elaboration	-0.120	0.028
Cultural Sensitivity	0.151	0.395

Ensemble Scoring

Given that the reliability testing confirmed that scores from any single LLM could be influenced by its idiosyncratic biases, and that two attempts to optimize the evaluation prompt failed, a robust method was required to derive a single, reliable score. Therefore, the final metric used for the primary analysis is the **ensemble score**. This score is calculated as the mean (average) of the ratings provided by the five LLM evaluators for each specific criterion on a given solution. This method creates a more stable and reliable measure by averaging out the variations and disagreements among the individual evaluators, thus mitigating their individual biases. The data from the original evaluation prompt was used as it was ready.

Statistical Analysis Plan

The statistical analysis was conducted to determine if there were significant differences in the ensemble scores across the 17 persona types.

- **Primary Statistical Test:** Due to the violation of assumptions for parametric tests (i.e., normality and homogeneity of variances) in preliminary analyses, the **Kruskal-Wallis H test** was selected as the primary statistical tool. This non-parametric test is appropriate for comparing the medians of three or more independent groups. A significance level (alpha) of $p < 0.05$ was used to determine statistical significance.

- **Post-Hoc Analysis:** In cases where the Kruskal-Wallis test revealed a statistically significant overall difference, a **Dunn's post-hoc test with Bonferroni correction** was performed. This test allows for pairwise comparisons to identify exactly which specific persona groups differ significantly from one another on a given criterion.

This analytical approach ensures a rigorous and appropriate examination of the data, allowing for confident conclusions about the impact of persona-based prompting on LLM output.

Qualitative Strand

This qualitative analysis aims to delve deeper into the textual justifications provided by five LLM evaluators for a specifically selected sample of business solutions. The primary objective is to understand *why* certain solutions, generated with different cultural personas (including Neutral and Expert benchmarks), received their respective ensemble scores on criteria such as Novelty, Usefulness/Feasibility, Flexibility, Elaboration, and Cultural Appropriateness/Sensitivity. This analysis seeks to connect the language used by evaluators back to the design intent of the personas and the context of the business problems, thereby enriching the quantitative findings.

Methodology

To extract qualitative insights, an automated thematic coding was conducted using "textblob" python library. For each sample, the justifications of the 5 LLMs were combined. Listing 2 shows the 5 justifications of the LLM evaluators for Culture 12 and Business Problem 1. The Cultural personas prompts along with the Business Problems were also added to the qualitative analysis.

Listing 2: *Verbatim Novelty Justifications from Five LLMs.*

Model: Meta Llama 3.1 405B Instruct Turbo.

> *The proposed solution is well-structured and comprehensive, but it doesn't offer particularly novel or unexpected approaches to market entry. The phases and steps outlined are fairly standard for international market expansion, although the emphasis on sustainability and adapting to local cultural nuances is a positive aspect.*

Model: Mistral Small 24B Instruct 2501.

> *The proposed solution follows a conventional market entry strategy framework, including market selection, entry mode, product positioning, and operational setup. While it is comprehensive, it does not offer any particularly novel or unexpected approaches to market entry.*

Model: Qwen2.5 72B Instruct Turbo.

> *The solution contains some common elements typical of market entry strategies, such as market selection, risk assessment, and operational setup. However, it includes a few*

> *slightly less obvious ideas, such as emphasizing local cultural nuances in the marketing and communication plan and investigating local sourcing and manufacturing options for sustainability.*

Model: DeepSeek R1 Distill Llama 70B.

> *The solution employs conventional strategies for market entry, such as market research and entry modes, without introducing innovative approaches. It lacks unique or unexpected elements.*

Model: Gemma 2 27b it.

> *The solution presents a standard, well-structured approach to international market entry. While it covers essential aspects, it lacks truly innovative or unconventional ideas.*

The steps of the analysis were as follows:

- Identify recurring keywords and phrases (aided by the "Top Keywords" in the report).
- Note the sentiment scores (compound score from VADER) associated with each justification.
- Extract illustrative quotes.
- Synthesize themes related to how evaluators perceived the solutions' strengths and weaknesses.
- Link these themes to the design characteristics of the cultural persona used to generate the respective solution and the business problem itself.

Samples Selection
Due to the large amount of data 31,875 evaluations (5 LLm evaluators * 6,375 business solutions), a purposful sampling was chosen to conduct the qualitative analysis. It was based on the quantitative results and consisted on the following elements:

- **High-Performing Personas:** For each criterion, solutions generated by personas that scored significantly higher than benchmarks or other personas.
- **Low-Performing Personas:** Solutions generated by personas that scored significantly lower
- **Benchmark Comparisons:** Solutions from Culture_Neutral and Culture_Expert, especially for criteria where they performed strongly or where they were outperformed.
- **Contrasting Cases:** Solutions by a persona that did well on one criterion but poorly on another

For each chosen persona, solutions across 3 different business problems were used to determine whether the themes are consistent or if the business problem context significantly interacts with the persona's influence
The total number of samples was 22.

Data Collection
The data of the samples was extracted form the main data file and justifications for each solution were merged. The resulted file was saved as a csv file and used for the qualitative analysis.
The Cultural Personas prompts along with the Business Problems details were fed to the qualitative analysis system (a python language program) as json files.

Researcher Reflexivity Statement

Introduction

In this statement, I reflect on my positionality and its influence on this research, which investigates the effects of AI-generated cultural personas on the creativity of Large Language Models (LLMs). My role was not one of a traditional experimenter, but of an architect and curator within a complex, AI-driven ecosystem. This statement aims to transparently outline the ways my decisions and perspectives shaped the methodology and the interpretation of the results, in alignment with the pragmatic philosophy that underpins this study.

Positionality and Relationship to the Topic

My identity as a researcher in the field of **Business and Artificial Intelligence** is shaped by a multicultural lens. I was raised in Lebanon within a French educational system, moved to France for university, and currently reside in Qatar while working in a global role. This background provides a unique vantage point, blending perspectives from a typically high-context Lebanese culture, a more individualistic French culture, and the distinct context of the Gulf region. My academic training in business and research methods provided the theoretical underpinnings of Hofstede and Hall's dimensions, but I approached this project with an awareness that my interpretation is filtered through this multilayered cultural lens and further broadened by extensive professional experience across North America, Europe, Asia, and Australia. My motivation stems from a deep interest in the intersection of AI and human culture, guided by the assumption that an LLM's output is not neutral, but can be systematically shaped.

Influence on the Research Process: A Researcher-AI Collaboration

My influence is most evident in the methodological decision to use a Large Language Model (Gemini Pro 2.5 Preview) to generate the core research instruments—the cultural personas and the business problems. This shifted my role from a direct author to that of a **prompter and curator**, introducing a second, algorithmic layer of interpretation.

Persona and Problem Generation

I did not write the personas myself; I designed the prompts that guided Gemini to create them based on established cultural theories. This process is subject to a layered bias: 1) **My own bias**, embedded in the specific instructions, examples, and constraints I provided (such as the standardized five-part structure for personas); and 2) **The model's inherent bias**, as Gemini's understanding of concepts like "collectivism" or "high power distance" is derived from the patterns within its vast, but not universal, training data. My subsequent selection of the "best" generated personas and problems was a further act of subjective judgment.

Defining and Measuring Creativity

The concept of "creativity" was operationalized through an evaluation framework executed by an ensemble of five LLMs. This created a closed research loop where the criteria for success were defined and measured within an AI ecosystem. A significant finding during preliminary testing was the **poor to moderate inter-rater reliability** among LLM evaluators, as evidenced by low Intraclass Correlation Coefficient (ICC) scores. I acknowledge that the evaluative prompts I designed, and the LLMs' idiosyncratic scoring tendencies, shaped the raw data. This crucial finding led directly to the decision to use an ensemble scoring method to mitigate this inherent variability.

Strategies to Mitigate Bias and Enhance Validity

My methodological choices were guided by a **pragmatic research paradigm**, which values finding the best way to answer the research question over allegiance to a single method. This philosophy provided the license to combine quantitative and qualitative approaches into the explanatory sequential mixed-methods design used in this paper. Recognizing my deep entanglement with the research process, I employed several specific strategies to ensure rigor.

Systematic Benchmarking

The inclusion of “Neutral” and “Expert” personas was a critical control. This allowed the analysis to focus on the *relative differences* between the cultural personas’ outputs, rather than making absolute claims about creativity.

Ensemble Scoring

Given the demonstrated low reliability of individual LLM evaluators, the primary metric for analysis became the **ensemble score**—the mean rating from all five evaluators. This approach was a direct response to a methodological challenge and served to create a more stable and reliable measure by averaging out the biases and variations of individual models.

Explanatory Sequential Mixed-Methods Analysis

The study was designed in two phases. A broad quantitative analysis first identified *if* and *what* effects occurred. This was followed by a targeted qualitative analysis to explain *why* those patterns emerged. This design ensures that the quantitative findings are not left unexplained and that the qualitative inquiry is grounded in statistically relevant cases, thereby providing a richer, more contextualized understanding.

Conceptual Transparency

This reflexivity statement itself is a primary strategy. By openly detailing the collaborative nature of the research with an AI and acknowledging the multiple layers of interpretation and methodological challenges, I am making the epistemological foundations of this study clear to the reader.

Conclusion

In conclusion, this research should be viewed as an exploration of how current LLMs respond to a researcher’s prompt-guided *interpretation* of cultural theories. The findings are as much a reflection on the human-AI interaction at the heart of the methodology as they are on the creative potential of the models themselves.

Integration

The integration of the quantitative and qualitative strands is the cornerstone of this study’s explanatory sequential design. Its primary purpose is to use the qualitative findings to explain and provide depth to the initial quantitative results. The connection between the two phases was deliberately designed to be sequential and explanatory, where the quantitative analysis of what occurred is followed by a qualitative investigation of why it occurred.

Point of Interface and Timing

The integration occurred at a single, critical juncture between the two research phases. This point of interface was the purposive sampling procedure for the qualitative strand. The statistical analysis of the 6,375 evaluated solutions was completed first. The results of this quantitative phase; specifically the ensemble scores for each solution and the outcomes of the Kruskal-Wallis and Dunn’s post-hoc tests—directly informed the selection of a small, information-rich subset of solutions for in-depth qualitative analysis. This timing ensures that the qualitative inquiry is not arbitrary but is strategically focused on the most statistically significant findings from the initial quantitative experiment.

Data Transformation and Linking Procedures

This study follows a "results-to-data" integration logic. The connection between the two phases was achieved by using the quantitative results as a direct mechanism to select appropriate qualitative data for explanatory analysis, a process that did not require data transformation procedures such as quantitizing qualitative data. The quantitative results were used to identify key categories of interest, which then informed the purposive sampling of 22 individual solutions for the qualitative phase. As detailed in the main paper’s methodology, this sampling involved selecting information-rich solutions representing high and low performers, contrasting cases, and crucial benchmark comparisons to understand baseline and optimal non-cultural outputs.

The final integration occurred during the explanatory analysis. The selected qualitative data — comprising the original business problems, the cultural persona prompts, the full texts of the generated solutions, and the corresponding evaluator justifications — were subjected to automated thematic analysis. The linking was completed during this interpretive stage, where the emergent qualitative themes were explicitly connected back to the quantitative score that prompted the solution's selection. For instance, the textual characteristics of a solution with a high or low score in a given criteria were analyzed to understand what specific elements the LLM evaluators collectively perceived as "novel", "Useful/Feasible", "Flexible", "Elaborated" or "Culturally appropriate/sensitive" in the context of the prompt that generated it. This systematic process ensures that the integration is not merely a side-by-side comparison of two disparate datasets, but a cohesive narrative where the qualitative findings serve to build a deep, contextualized explanation for the broad patterns revealed by the quantitative analysis.

Quantitative Findings

The quantitative phase of this study was designed to answer the primary research question: *Does inserting cultural personas to LLMs affect their creativity in solving business problems?* The analysis was performed on a dataset of 6,375 unique solutions, with each solution being generated by one of 17 distinct persona types (15 distinct cultures formed by a combination of Hofstede and Hall's cultural dimensions, and 2 personas used for benchmarking; one neutral and one a business expert). They are presented as Culture_X, where X is the number of each cultures from 1 to 15 and Culture_Neutral and Culture_Expert. Each solution was assessed by an ensemble of five LLM evaluators, and the final score for each criterion is the mean of these five ratings.

A Kruskal-Wallis H test was conducted for each of the five evaluation criteria to determine if statistically significant differences existed among the 17 persona types. The analysis revealed a highly significant effect for every criterion ($p < 0.0001$), indicating that the choice of persona has a measurable and significant impact on the perceived quality and characteristics of the generated solutions.

Following the significant Kruskal-Wallis results, a Dunn's post-hoc test with Bonferroni correction was performed for each criterion to identify which specific persona pairs differed significantly. The following sections detail the findings for each criterion, highlighting notable differences between top-performing cultural personas, the **Culture_Neutral** and **Culture_Expert** benchmarks, and lower-performing personas.

The mean score for each criterion is presented after the mentioned cultural personas as follows: Culture_5 (3.53) For each criterion, a boxplot shows the median (center line), interquartile range (box), and full data range excluding outliers (whiskers). Outliers are plotted as individual points.

Criterion 1: Novelty

The analysis revealed a highly significant difference among the persona types on the **Novelty** scores (Kruskal-Wallis H = 688.195, $p < 0.0001$).

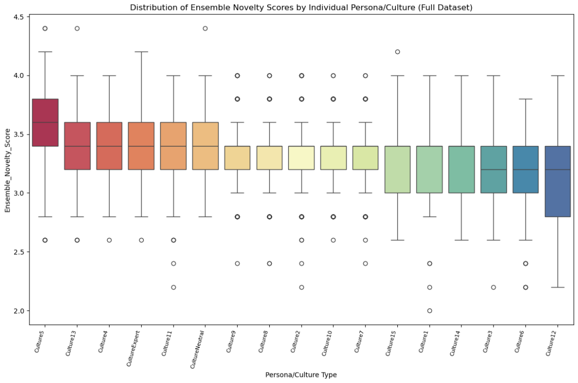


Figure 2

Distribution of Ensemble Novelty Scores by Cultural Persona.

Summary of Post-Hoc Findings

Dunn's post-hoc tests revealed that specific cultural personas significantly enhanced the perceived novelty of solutions (see Figure 2).

- **Culture_5**, (M = 3.53), the top performer, scored significantly higher than both **Culture_Neutral** (M = 3.34) and **Culture_Expert**, (M = 3.37), as well as lower-ranking personas like **Culture_12** (M = 3.11) and **Culture_3** (M = 3.18).
- **Culture_13** (M = 3.41) and **Culture_4** (M = 3.41) also significantly outperformed the **Culture_Neutral** and **Culture_Expert** benchmarks.
- Conversely, **Culture_12** was the lowest performer, scoring significantly lower than almost all other personas.

Insight

These results strongly indicate that targeted persona design, particularly with personas like **Culture_5**, **Culture_13**, and **Culture_4**, can be a strategic tool to drive **novelty** in LLM-generated business solutions, surpassing the output of standard expert or neutral prompts.

Criterion 2: Usefulness/Feasibility

A significant difference was also found for **Usefulness/Feasibility** (Kruskal-Wallis H = 74.336, $p < 0.0001$). However, the key performers here were the benchmarks.

Summary of Post-Hoc Findings

The post-hoc analysis confirmed that the benchmark personas excelled in this foundational criterion (see Figure 3).

- **Culture_Neutral** (M = 4.38) and **Culture_Expert** (M = 4.38) were the top performers and scored significantly higher than the lowest-ranking personas, **Culture_12** (M = 4.27) and **Culture_5** (M = 4.26) .
- Notably, **Culture_5**, which was the top performer for **Novelty**, was the lowest performer for **Usefulness/Feasibility**.

Insight

The benchmark personas are highly effective at producing solutions that are perceived as **useful and feasible**. There appears to be a potential trade-off, where some personas that generate highly **novel** ideas (**Culture_5**) may do so at the expense of immediate practical usefulness.

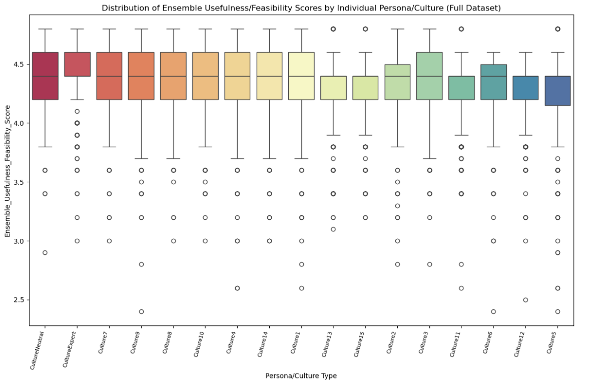


Figure 3

Distribution of Ensemble Usefulness/Feasibility Scores by Cultural Persona.

Criterion 3: Flexibility

The pattern for **Flexibility** (Kruskal-Wallis H = 413.008, $p < 0.0001$) was similar to that of **Novelty**, with specific cultural personas outperforming the benchmarks.

Summary of Post-Hoc Findings

- **Culture_5** (M = 4.09), **Culture_4** (M = 4.03) ,and **Culture_13** (M = 4.01) again emerged as top performers, with **Culture_4** and **Culture_13** scoring significantly higher than both **Culture_Neutral** (M = 3.97) and **Culture_Expert** (M = 4.00).
- **Culture_5** scored significantly higher than **Culture_Neutral** and many others, though

its difference from **Culture_Expert** was not statistically significant.

- **Culture_12** (M = 3.61) was once again the clear underperformer, scoring significantly lower than nearly all other personas.

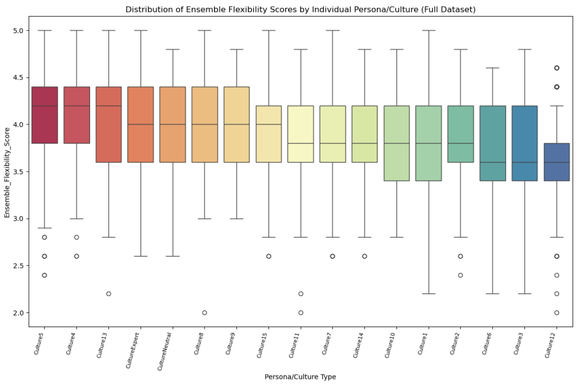


Figure 4

Distribution of Ensemble Flexibility Scores by Cultural Persona.

Insight

Similar to **Novelty**, these findings show that specific cultural personas can be used to generate solutions that are perceived as more **flexible**, suggesting these personas encourage a broader exploration of conceptual space (see Figure 4).

Criterion 4: Elaboration

For **Elaboration** (Kruskal-Wallis H = 137.632, $p < 0.0001$), the results mirrored those for **Usefulness/Feasibility**, with the benchmarks performing best.

Summary of Post-Hoc Findings

- **Culture_Expert** (M = 4.53) and **Culture_Neutral** (M = 4.53) were the clear top performers, scoring significantly higher than a large group of cultural personas, including **Culture_5** (M = 4.23), **Culture_13** (M = 4.33), **Culture_11** (M = 4.37), and **Culture_4** (M = 4.42).
- **Culture_5** had the lowest mean score and was found to be significantly lower than most other personas on elaboration.

Insight

The benchmark personas are superior for generating detailed and well-developed solutions (see Figure 5). This again highlights the persona-criterion specificity, where **Culture_5**, a top performer for **Novelty** and **Flexibility**, significantly underperforms in providing detailed **elaboration**, reinforcing the idea of a trade-off between generative exploration and detailed development.

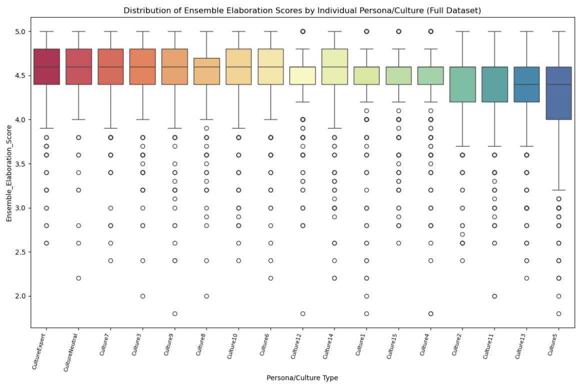


Figure 5

Distribution of Ensemble Elaboration Scores by Cultural Persona.

Criterion 5: Cultural Appropriateness/Sensitivity

This criterion yielded a critical result, with a highly significant difference among personas (Kruskal-Wallis H = 376.436, $p < 0.0001$).

Summary of Post-Hoc Findings

- A group of specific cultural personas—**Culture_11** (M = 4.61), **Culture_13** (M = 4.60), **Culture_2** (M = 4.59), **Culture_4** (M = 4.58), and **Culture_5** (M = 4.55)—all scored significantly higher than both the **Culture_Neutral** (M = 4.49) and **Culture_Expert** (M = 4.48) benchmarks.

- These top-performing cultural personas also significantly outperformed the lowest-ranking cultural personas like **Culture_12** (M = 4.25), **Culture_6** (M = 4.34), and **Culture_3** (M = 4.34).

- The **Culture_Neutral** and **Culture_Expert** benchmarks, while not the top performers, still significantly outperformed the lowest-tier personas.

Insight

This result demonstrates that specific, targeted cultural personas can demonstrably and significantly enhance the **cultural appropriateness and sensitivity** of LLM-generated solutions, performing better than generic neutral or expert prompts (see Figure 6). This confirms that a structured approach to persona design can achieve desired outcomes for this dimension.

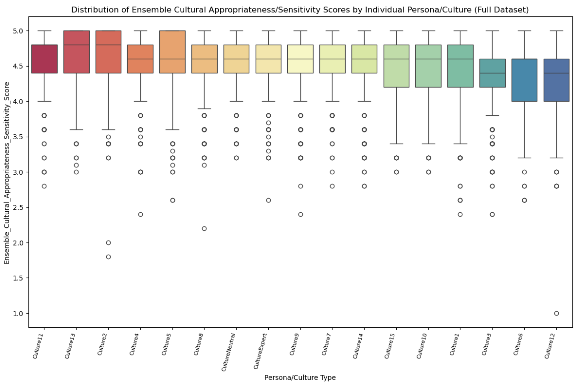


Figure 6

Distribution of Ensemble Cultural Appropriateness/Sensitivity Scores by Cultural Persona.

Summary of Quantitative Findings and Implications

The quantitative analysis robustly demonstrates that the choice of persona has a statistically significant and often substantial impact on all evaluated dimensions of LLM-generated business solutions. The findings reveal several key implications that will guide the subsequent qualitative analysis:

- **No Single "Best" Persona Exists:** The results clearly show that different personas excel in different areas. There is no universally superior persona; rather, their effectiveness is criterion-dependent.

- **Targeted Enhancement with Cultural Personas:** Specific cultural personas can be used to strategically enhance desired solution characteristics. For **Novelty and Flexibility**, personas such as **Culture_5**, **Culture_13**, and **Culture_4** showed strong, positive effects, often outperforming benchmarks. For **Cultural Appropriateness/Sensitivity**, a distinct group including **Culture_11**, **Culture_13**, and **Culture_2** proved most effective.

- **Benchmarks Excel in Foundational Qualities:** The **Culture_Neutral** and **Culture_Expert** personas consistently yielded solutions rated highest on **Usefulness/Feasibility** and **Elaboration**. This suggests they are reliable for generating practical, well-developed, and detailed outputs.

- **Evidence of Persona-Criterion Trade-offs:** A critical finding is the evidence of trade-offs. For example, **Culture_5** was the top performer for Novelty and Flexibility but the worst for Elaboration and among the lowest for Usefulness/Feasibility. This suggests that prompting for high levels of originality may come at the cost of detail and perceived practicality, highlighting the nuanced role of personas in prompt engineering.

- **Identification of Underperforming Personas:** Certain personas, particularly **Culture_12**, and to a lesser extent **Culture_3** and **Culture_6**, consistently ranked lower across multiple criteria. Understanding the design characteristics of these personas that led to this underperformance is a key objective for the qualitative phase.

These quantitative results provide the "what"—the clear, statistical evidence of persona impact. The next phase of this study will focus on the "why," using qualitative analysis to explore the textual and conceptual characteristics of the solutions that drove these distinct outcomes.

Qualitative Findings: Explaining the "Why" Behind the Scores

Following the quantitative analysis which established *that* cultural personas significantly impact solution generation, this section addresses *why* these differences occurred. As per the explanatory sequential design of this study, a purposive sample of solutions was selected for in-depth qualitative analysis. The sample included solutions from high- and low-performing personas, contrasting cases, and the `Culture_Neutral` and `Culture_Expert` benchmarks, analyzed across distinct business problems (`Business Problem 1`: Market Entry, `Business Problem 4`: Product Development, and `Business Problem 13`: PR Crisis Management). The following thematic analysis of the LLM evaluators' textual justifications connects the quantitative scores back to the design intent of the personas and the specific textual qualities of the generated solutions, supported by sentiment analysis scores (VADER) and recurring keywords. The evaluators are presented as follows in this analysis for clarity:

- **Meta Llama 3.1 405B Instruct Turbo:** Llama
- **DeepSeek-R1-Distill-Llama-70B**: DeepSeek
- **Qwen2.5 72B Instruct Turbo:** Qwen
- **Mistral Small 24B Instruct 2501:** Mistral
- **Gemma-2 Instruct (27B):** Gemma

The following figures provide a multi-faceted profile for each key persona, visually integrating the persona's design with its performance. Each profile consists of three radar charts, which should be interpreted as follows:

1. **Cultural Dimensions (Input):** This chart visualizes the persona's design based on the cultural frameworks of Hofstede and Hall. The distance from the center on each axis represents the strength of that cultural trait. Note that several dimensions are bipolar; for example, a low score on *Individualism* implies a high score on its opposite pole, *Collectivism.*

2. **Thematic Fingerprint (Output):** This chart visualizes the qualitative analysis of the evaluators' justifications. The distance from the center indicates the frequency of keywords associated with each theme (e.g., "practical," "innovative"), while the point's color corresponds to the average sentiment of the text (green is positive, gray in neutral and red is negative).

3. **Creativity Scores (Output):** This chart displays the final quantitative results from the experiment. The scoring on each axis represents the mean ensemble score (on a scale of 1 to 5) that the persona achieved for each of the five creativity criteria.

Benchmark Personas (`Culture_Neutral` & `Culture_Expert`): The Foundation of Practicality and Elaboration

The qualitative justifications strongly support the quantitative finding that `Culture_Neutral` and `Culture_Expert` excelled in Usefulness/Feasibility and Elaboration.

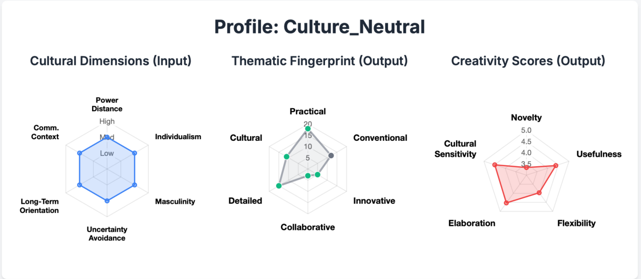


Figure 7

*Visual profile of the **Culture_Neutral benchmark** persona, connecting its predefined input dimensions (left) with the thematic fingerprint of its generated solutions (middle) and its final ensemble creativity scores (right).*

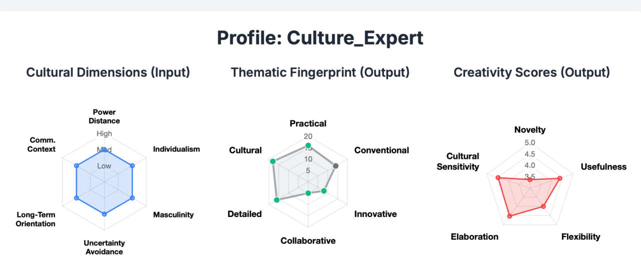


Figure 8

*Visual profile of the **Culture_Expert benchmark** persona, providing a comparative summary of its input definition, output thematic fingerprint, and ensemble creativity scores.*

For a market entry solution (`Business Problem 1`) from `Culture_Neutral`, evaluators described it as *"highly practical, with clear steps for market research, legal compliance, and building local partnerships"*. This assessment was reinforced by consistently positive sentiment scores from multiple evaluators (Qwen: 0.886, Llama: 0.915) and recurring keywords like "practical," "clear," and "realistic". Similarly, for a PR crisis (`Business Problem 13`), a `Culture_Neutral` solution was praised as *"highly practical and realistic"* with *"clear, actionable steps"*, reflected in high sentiment scores (Llama: 0.925, Gemma: 0.893). Their solutions were consistently praised for Elaboration, described as *"highly detailed and clearly articulated"* and *"well-developed"*.
However, the justifications also explain their moderate Novelty scores. While some evaluators noted a *"fresh perspective"* or *"unique elements"*, a recurring theme was that the solutions were largely *"conventional"* or based on *"established strategies"*. Gemma noted a `Culture_Expert` solution for a PR crisis was a *"standard crisis management approach"*, a sentiment that received a negative compound score (-0.318). This theme suggests the benchmark personas, `Culture_Neutral` and `Culture_Expert`, reliably produce well-structured and comprehensive outputs but are less likely to generate groundbreaking ideas, reinforcing their role as a baseline for practical and feasible solutions (see Figures 7 and 8).

High-Performing Personas: The Direct Link Between Design and Outcome

The analysis revealed a direct correlation between a persona's core design principles and the specific creative criteria where it excelled.
Novelty and Flexibility Leaders (`Culture_13`), `Culture_4`, and the Contrasting Case of `Culture_5`). The personas that scored highest on Novelty and Flexibility did so for reasons directly tied to their underlying design.

- **`Culture_13` (Long-term, Relational Focus):** This persona's novelty was derived from its unique strategic frameworks. For a PR crisis (`Business Problem 13`), Qwen highlighted its *"multiphase approach that includes unique elements such as a 'Quiet Period' and a focus on 'Bridge Building',"* calling these elements *"not typically seen in standard crisis management strategies"* (Sentiment: -0.421). Llama saw the focus on *"listening, acknowledgment, and bridge-building"* as a *"fresh perspective on crisis management"*. This relational design also enhanced Flexibility, with evaluators praising its *"range of approaches"* and how *"local* *community involvement provides adaptability"*. For product development (Business Problem 4), its novelty was evident in *"unique and culturally relevant snack options, such as Seaweed Crispy Bites and Tart Mango Bites"* (see Figure 9)

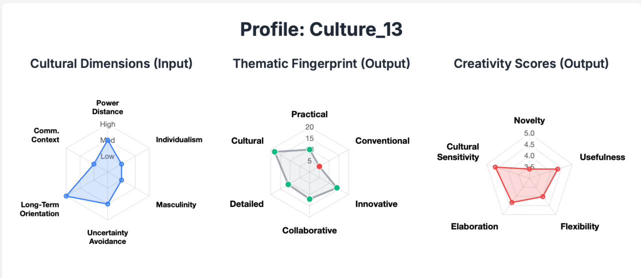


Figure 9

Visual profile of the ***Culture_13*** *persona, analyzed as a top performer for its high scores in* ***Novelty and Flexibility****. The chart connects its "Long-term, Relational Focus" design (left) to its thematic output (middle) and final creativity scores (right).*

- **Culture_4 (Collectivist, Collaborative Focus):** This persona generated novelty through its emphasis on collaborative processes (see Figure 10). For a market entry task, Mistral found it *"moderately novel by emphasizing collaboration, inclusivity, and shared responsibility"*, while Llama called the focus on *"interpersonal relationships, shared leadership, and collective impact"* innovative and unexpected. For a PR crisis, Qwen noted its *"innovative elements, such as the establishment of a cultural advisory board and community investment"*. Its solutions were also seen as inherently flexible due to this collaborative nature, exploring *"multiple distinct approaches"* and offering a *"wide range of ideas and adaptable frameworks"*.

Cultural Appropriateness/Sensitivity Leaders (Culture_11 & Culture_2). Personas that excelled on this criterion did so because their core designs directly addressed cultural integration.

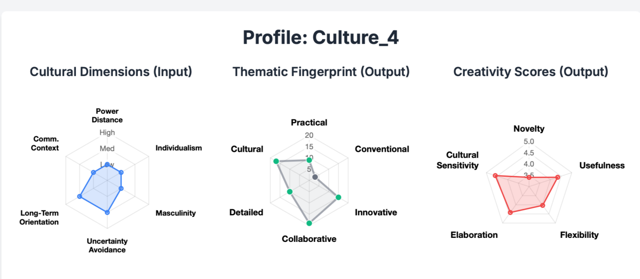


Figure 10

Visual profile of the ***Culture_4*** *persona, another high-performer in* ***Novelty and Flexibility*** *selected for analysis. The chart illustrates how its "Collectivist, Collaborative Focus" (left) relates to its thematic output (middle) and final creativity scores (right).*

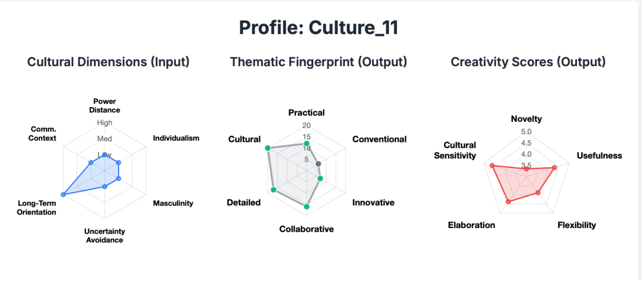


Figure 11

Visual profile of the ***Culture_11*** *persona, selected as a leader in* ***Cultural Appropriateness/Sensitivity****. The chart links its "Harmony, High-Context Focus" design (left) to its thematic output (middle) and final creativity scores (right).*

- **Culture_11 (Harmony, High-Context Focus):** For a PR crisis, Mistral noted it *"demonstrates a high degree of cultural awareness, subtly incorporating cultural considerations throughout"* (Sentiment: 0.866). Llama praised its *"genuine commitment to understanding and respecting local values"* (Sentiment: 0.934). This can be seen in Figure 11 along with its cultural profile. For market entry, Culture_11' strength was in *"emphasizing the importance of understanding local*

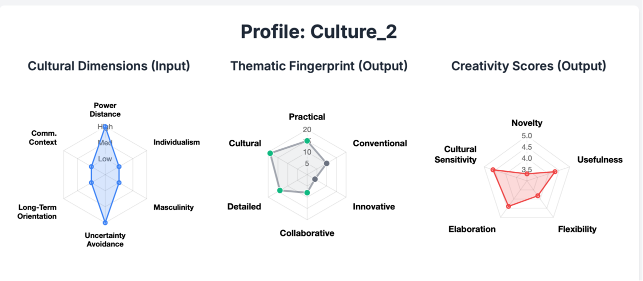


Figure 12

*Visual profile of the **Culture_2** persona, another top performer chosen for its high scores in **Cultural Appropriateness/Sensitivity** due to its focus on tradition and group orientation.*

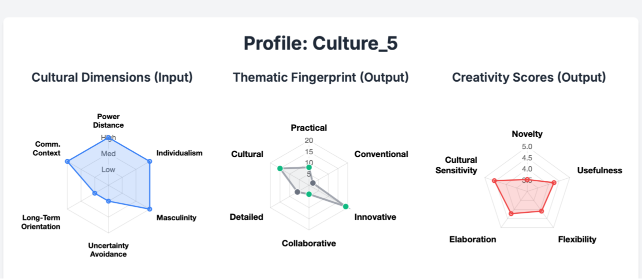


Figure 13

*Visual profile of the **Culture_5** persona, analyzed as a key contrasting case to illustrate the **trade-off** between its high **Novelty** score and its low scores for **Elaboration and Usefulness/Feasibility**.*

preferences, sensitivities, and the need for localized marketing" (Sentiment: 0.904). This was achieved through a focus on *"building relationships and partnerships that align with the company's values and the local culture"* .

- **`Culture_2` (Tradition, Group-Oriented Focus):** For product development (`Business Problem 4`), its strength was its *"high degree of cultural awareness"* by considering *"locally sourced ingredients, consideration of traditional consumption occasions, and incorporation of cultural nuances"*. It effectively used *"traditional ingredients and aligns with local tastes, showing cultural sensitivity without relying on stereotypes"*. For a PR crisis, it was praised for its *"high degree of cultural awareness, with steps explicitly designed to address and respect local cultural values"* (see Figure 12).

Persona-Criterion Trade-offs and Underperformance

The `Culture_5` Contrasting Case: Novelty at the Cost of Elaboration. The qualitative justifications clearly explain the quantitative trade-off observed for `Culture_5` (see Figure 13). It excelled in Novelty; for product development, Qwen noted it *"proposes a unique combination of savory flavors and unexpected textures... a fresh and non-obvious approach"* (Sentiment: 0.778). However, this came at the expense of Elaboration. The same solution was described by Gemma as *"somewhat vague"* and by DeepSeek as lacking *"detailed plans for production, marketing, or specific examples"*. This lack of detail, reflected in lower sentiment scores for elaboration (e.g., DeepSeek: 0.202), directly harms the perceived feasibility of its otherwise novel ideas, providing a clear explanation for its contradictory quantitative performance.

The `Culture_12` Underperformer: The Consequence of Conventionality. The analysis of `Culture_12`, which consistently scored low, demonstrates how a persona's design can lead to underperformance. Designed to be straightforward and safe, its solutions were predictably perceived as unoriginal. For a market entry problem, evaluator justifications stated it *"doesn't offer particularly novel or unexpected approaches"* and *"follows a conventional market entry strategy framework"*. This directly explains its low quantitative scores for Novelty and Flexibility, with Mistral stating it *"does not explore a wide range of possibilities"*. Furthermore, its focus on directness led

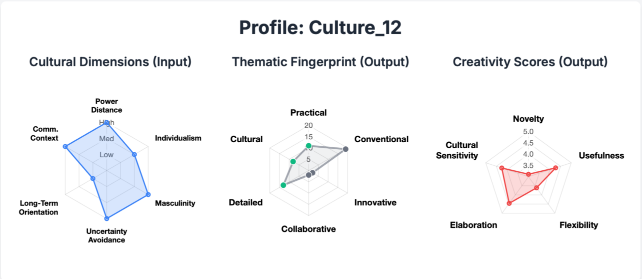


Figure 14

*Visual profile of the **Culture_12** persona, selected for analysis as a consistent **underperformer**. The chart visualizes how its design for a safe and conventional approach led to low scores across multiple creative criteria*

to surface-level cultural considerations; Gemma noted it *"doesn't delve deeply into specific cultural considerations or provide concrete examples of how to address them"*, explaining its bottom-tier rank for Cultural Appropriateness/Sensitivity. The persona's design for safety directly resulted in solutions perceived as lacking innovation, flexibility, and deep cultural nuance (see Figure 14).

Integration of Quantitative and Qualitative Findings

The primary objective of this explanatory sequential mixed-methods study was to move beyond simply measuring the effect of cultural personas on LLM creativity to explaining *why* these effects occur. This section integrates the quantitative results, which established *what* changes happened, with the qualitative findings, which reveal *how* and *why* they happened. The integration is organized around the key themes that emerged from the qualitative analysis of the LLM evaluators' justifications, directly linking the statistical patterns to the textual and conceptual characteristics of the generated solutions.

Theme 1: Explaining the Spectrum of Innovation

The quantitative analysis revealed a clear hierarchy for Novelty and Flexibility, with personas like **Culture_5 (M=3.53)**, **Culture_13 (M=3.41)**, and **Culture_4 (M=3.41)** significantly outperforming the **Culture_Neutral (M=3.34)** and **Culture_Expert (M=3.37)** benchmarks. Conversely, **Culture_12 (M=3.11)** was a consistent underperformer on these criteria. The qualitative data provides a direct explanation for these scores, showing a clear link between the persona's design and the evaluators' perception of innovation.

The high-performing personas were designed with specific creative catalysts. For instance, **Culture_13** (Long-term, Relational Focus) generated novelty through unique strategic frameworks. When addressing a PR crisis (Business Problem 13), evaluators highlighted its "multiphase approach that includes unique elements such as a 'Quiet Period' and a focus on 'Bridge Building'," which were described as "not typically seen in standard crisis management strategies." Similarly, the collaborative design of **Culture_4** (Collectivist, Collaborative Focus) led to novel processes; for a market entry problem, evaluators found its emphasis on "interpersonal relationships, shared leadership, and collective impact" to be innovative and unexpected.

In contrast, the underperformance of **Culture_12** was a direct result of its conventional design. Evaluator justifications for its market entry solution consistently labeled it as offering "conventional strategies" and lacking "unique or unexpected elements." The persona was designed for a straightforward, safe approach, and the qualitative feedback confirms that this translated into solutions perceived as unoriginal, thus explaining its low quantitative scores for Novelty and Flexibility.

Theme 2: The Foundation of Practicality and Actionability

A key quantitative finding was the superior performance of the **Culture_Neutral (M=4.38)** and **Culture_Expert (M=4.38)** personas on Usefulness/Feasibility and Elaboration. They significantly outperformed many cultural personas, including the novelty-leader **Culture_5 (M=4.26 on Usefulness, M=4.23 on Elaboration)**.

The qualitative analysis affirms that this is by design. The benchmark personas consistently produced solutions praised by evaluators for their practicality and detail. For a PR crisis (Business Problem 13), a **Culture_Neutral** solution was described as “highly practical and realistic, taking into account the need for immediate action, short-term corrective measures, and long-term strategic planning." Likewise, its solutions were lauded for Elaboration, with justifications noting they were “highly detailed and clearly articulated." These personas reliably generate well-structured, actionable, and comprehensive outputs, which evaluators perceive as highly useful and feasible, explaining their top-tier scores on these foundational criteria.

Theme 3: Illuminating the Persona-Criterion Trade-Off

Perhaps the most insightful finding from the integration is the clear explanation for the observed trade-offs between creativity and practicality. The case of **Culture_5** provides a powerful example. While it was the top performer on Novelty, it was among the lowest on Usefulness/Feasibility and the single lowest on Elaboration.

The qualitative justifications reveal this tension explicitly. For a product development task (Business Problem 4), evaluators praised **Culture_5** for its novelty, with one noting it “proposes a unique combination of savory flavors and unexpected textures...a fresh and non-obvious approach." However, in the same breath, evaluators criticized its lack of detail, stating it was “somewhat vague" and “lacks detailed plans for production, marketing, or specific examples." This demonstrates that the persona’s design, which prioritizes risk-taking and unconventional thinking, successfully generated creative concepts but did so at the expense of the detailed, practical planning that evaluators require to deem a solution fully feasible and well-elaborated. This direct conflict in the qualitative feedback provides a robust explanation for the persona’s contradictory quantitative scores.

Theme 4: Defining the Depth of Cultural Integration

The quantitative results showed that a specific group of cultural personas—notably **Culture_11 (M=4.61)**, **Culture_13 (M=4.60)**, and **Culture_2 (M=4.59)**—significantly outperformed the benchmarks on Cultural Appropriateness/Sensitivity. The qualitative analysis explains that this success stems from a deep and genuine integration of cultural values rather than a superficial mention.

For the PR crisis problem, a context where cultural sensitivity is paramount, the **Culture_11** persona (Harmony, High-Context Focus) was praised for its “genuine commitment to understanding and respecting local values." Evaluators noted that it “subtly incorporat[ed] cultural considerations throughout the phases," demonstrating a process that was itself culturally attuned. This contrasts sharply with the low-scoring **Culture_12**, whose approach was deemed superficial. Evaluators noted that while it “acknowledges the importance of cultural nuances," it “doesn’t delve deeply into specific cultural considerations or provide concrete examples," explaining its bottom-tier ranking. This distinction between deep integration and surface-level acknowledgment is the clear driver behind the significant quantitative differences observed for this criterion.

Synthesis and Conclusion of Integration

The integration of the quantitative and qualitative findings provides a comprehensive and cohesive understanding of how cultural personas shape

LLM creativity. The quantitative data identified statistically significant patterns, while the qualitative data provided the explanatory mechanism, revealing a direct causal link between persona design and solution characteristics. The analysis confirms that there is no single “best" persona; effectiveness is criterion-dependent. Personas designed for innovation successfully generate novel ideas but may do so at the cost of practical detail. Conversely, benchmark personas reliably produce well-elaborated and feasible solutions but are less likely to be innovative. Finally, personas designed with specific cultural values like harmony and relationality produce solutions that are perceived as more genuinely culturally appropriate than generic expert or neutral prompts. This integrated evidence demonstrates that persona engineering is a powerful and nuanced tool for strategically guiding LLMs to produce desired creative outcomes. To visually summarize these findings, the key personas are plotted on a strategic map that illustrates the trade-off between innovation and practicality (see Figure 15).

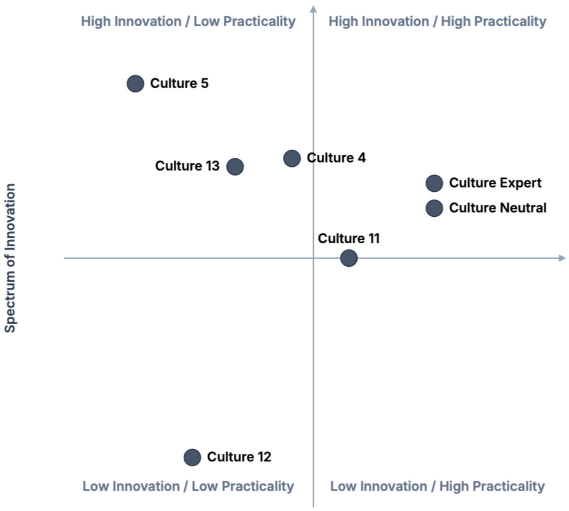


Figure 15

Visual synthesis of the integrated findings, illustrating the criterion-dependent effect of cultural personas. The map plots key personas along the axes of Innovation (derived from Novelty and Flexibility scores) and Practicality (derived from Usefulness and Elaboration scores), summarizing the core trade-offs identified in the analysis.

Discussion

This study was designed to measure and explain the impact of culturally-informed persona prompts on the creative output of Large Language Models (LLMs) within a business problem-solving context. Utilizing an explanatory sequential mixed-methods design, the research first quantified the effects of these personas and subsequently used qualitative analysis to explore the underlying reasons for these effects. The integrated findings provide a nuanced understanding of prompt engineering, indicating that the creative output of LLMs is not only malleable but can be strategically directed. This discussion interprets these findings in relation to the existing literature and delineates their theoretical and practical implications.

Interpreting the Integrated Findings: The Criterion-Dependent Nature of Persona Effectiveness

The central finding of this research is that the effectiveness of a persona prompt is highly dependent on the specific creative criterion being evaluated. The quantitative analysis demonstrated that different personas excelled in distinct creative dimensions, a result that was clarified by the qualitative data, which linked the design of each persona to its specific performance profile.

1. The Spectrum of Innovation and the Foundation of Practicality

The results align with the foundational definition of creativity as requiring both novelty and usefulness (Amabile, 1996; Zhao et al., 2024), but further demonstrate that these qualities can be independently modulated in LLMs. The benchmark personas, `Culture_Neutral` and `Culture_Expert`, consistently produced solutions rated highest on **Usefulness/Feasibility** and **Elaboration**. The qualitative analysis corroborates this, showing that evaluators frequently described these outputs as "practical," "clear," and "well-structured." This suggests that standard or expert-framed prompts are optimized for generating conventional, actionable, and detailed solutions, which corresponds to the common application of LLMs for knowledge extraction and summarization in business contexts (Su et al., 2024).

In contrast, specific cultural personas, namely `Culture_5`, `Culture_13`, and `Culture_4`, significantly surpassed the benchmarks on measures of **Novelty** and **Flexibility**. The qualitative findings explain this disparity by revealing how the core design principles of these personas functioned as creative catalysts. For instance, the long-term, relational focus of `Culture_13` generated unique strategic frameworks, such as a "Quiet Period" in a PR crisis, which evaluators identified as unconventional. This finding extends prior work on persona prompting (e.g.,(Li et al., 2025)) by demonstrating that the *specific, theory-driven characteristics* of a persona, rather than its mere presence, are what drive innovative output.

2. The Persona-Criterion Trade-off: A New Dimension in Prompt Engineering

A key theoretical contribution of this study is the empirical identification of a "persona-criterion trade-off." `Culture_5`, the highest-performing persona for Novelty, was concurrently the lowest-performing for Elaboration and among the lowest for Usefulness/Feasibility. The qualitative justifications clarify this trade-off: the persona's risk-tolerant and unconventional design produced "fresh and non-obvious" ideas that were also described as "somewhat vague" and lacking "detailed plans." This challenges the assumption that a single prompt can simultaneously maximize all facets of creativity. It suggests the need for a more sophisticated model of prompt engineering, wherein users make deliberate choices about which creative qualities to prioritize. This tension between novelty and usefulness is well-documented in studies of human creativity (A. Mukherjee & Chang, 2023); this study demonstrates its direct applicability and manipulability within an AI context.

3. Deep Cultural Integration Versus Surface-Level Acknowledgment

The study confirmed that targeted cultural personas can significantly improve the **Cultural Appropriateness/Sensitivity** of LLM-generated solutions beyond the level of generic expert prompts, supporting recent research on cultural alignment in LLMs (Chhikara et al., 2025; Feng et al., 2025). The mixed-methods approach employed here explains the mechanism behind this effect. The success of personas like `Culture_11` (Harmony, High-Context) was attributed by evaluators to a "genuine commitment to understanding and respecting local values" that was integrated throughout the solution. This was contrasted with the underperforming `Culture_12`, which was perceived as merely acknowledging cultural nuances at a surface level. This distinction between deep, process-oriented integration and superficial mention offers a critical insight for the development of culturally competent AI systems.

Strengths, Limitations, and Methodological Reflections

The primary strength of this study is its explanatory sequential mixed-methods design, which provided robust quantitative evidence of persona effects and a deep, contextualized understanding of the underlying mechanisms. The large-scale data generation (N = 6,375 solutions) and the use of an LLM evaluator ensemble enhanced the reliability of the quantitative findings.

However, the study is subject to several limitations. The reliance on an entirely AI-driven ecosystem for both generation and evaluation is the most significant. Although the low inter-rater reliability among individual LLM evaluators was a key finding and was methodologically addressed using an ensemble score, the evaluation remains an AI-centric perspective of creativity. Future research should incorporate human expert evaluation to validate these findings. Second, the cultural personas were based on the frameworks of Hofstede and Hall. While foundational, these frameworks have been critiqued and do not capture the full complexity of global cultures. The findings are therefore limited to the operationalization of these specific dimensions. Finally, the use of hypothetical business problems means the solutions were not tested for efficacy in a real-world setting.

Implications and Future Directions

The findings have significant implications for both theory and practice.

Theoretical Implications

This study contributes to a more nuanced model of AI creativity, positing that personas can function as control variables for specific creative outcomes, complete with predictable trade-offs. Furthermore, it demonstrates a method for operationalizing abstract cultural dimensions into structured prompts that produce measurable and distinct outputs, offering a new avenue for research in computational social science.

Practical Implications

For practitioners, this research suggests a move from generic to strategic prompt engineering. For instance, a persona like `Culture_5` could be used for novel ideation, while a `Culture_Expert` persona could be used to develop a detailed implementation plan. For global organizations, this study provides a methodology for generating strategies that are more likely to be resonant with local cultural values. The findings also suggest a more dynamic human-AI workflow, where a user could sequentially deploy different personas to guide an LLM through a multi-stage creative process.

Future Research

Future research should prioritize validating these findings with human evaluators from diverse cultural backgrounds. It would also be valuable to explore a wider range of cultural dimensions and investigate the use of sequential persona-prompting to overcome the observed trade-offs. Applying this methodology to other creative domains, such as scientific discovery or artistic co-creation, could further delineate the potential and limits of shaping AI creativity.

In conclusion, this study demonstrates that by embedding structured cultural intelligence into LLM personas, it is possible to systematically influence their creative output. This suggests a shift away from treating LLMs as inscrutable systems and toward an understanding of the specific mechanisms that can be used to direct their responses with intention and precision, thereby opening new frontiers for human-AI creative collaboration.

Conclusion

This study investigated how theory-driven cultural personas affect the multi-dimensional nature of creativity in Large Language Models (LLMs). The research sought to answer how structured personas, based on the frameworks of Hofstede and Hall, influence the perceived creativity of LLM-generated business solutions across five distinct criteria. The findings reveal that the effect of a persona is not monolithic but highly criterion-dependent; no single persona universally enhances all aspects of creativity.

A key contribution of this research is the empirical identification of a *persona-criterion trade-off*. Personas designed to maximize originality successfully generated solutions high in Novelty and Flexibility but did so at the expense of practical detail, scoring lowest on Elaboration and Usefulness/Feasibility . Conversely, the benchmark personas excelled at producing well-elaborated and feasible solutions but were less innovative. This insight shifts the practice of prompt engineering from a generic approach toward a strategic one, where specific personas can be intentionally deployed for different stages of the creative process, such as using one persona for novel ideation and another for implementation planning.

Ultimately, this study provides a robust framework for directing AI creativity, demonstrating that methodical prompt design is essential for moving beyond arbitrary generation. By understanding the specific mechanisms that shape LLM output, this research offers a clear path toward more intentional, nuanced, and predictable outcomes. It helps reframe human-AI interaction not as an interaction with an inscrutable system, but as a precise, collaborative partnership to achieve specific creative goals.

References

Aguero, D., & Nelson, S. D. (2024). The potential application of large language models in pharmaceutical supply chain management. *The Journal of Pediatric Pharmacology and Therapeutics*, *29*(2), 200–205. https://doi.org/10.5863/1551-6776-29.2.200

Akinwande, M., Yussuph, T., & Adeliyi, O. (2024). Decoding ai and human authorship: Nuances revealed through nlp and statistical analysis [Publisher: RELX Group (Netherlands)]. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4895334

Amabile, T. M. (1996, June). *Creativity in context: Update to the social psychology of creativity*. https://ci.nii.ac.jp/ncid/BA28033699

Barth, P., & Stadtmann, G. (2020). Creativity assessment over time: Examining the reliability of CAT ratings [Publisher: Wiley]. *The Journal of Creative Behavior*, *55*(2), 396–409. https://doi.org/10.1002/jocb.462

Berti, L., Giorgi, F., & Kasneci, G. (2025). Emergent abilities in large language models: A survey [Publisher: arXiv]. https://doi.org/10.48550/ARXIV.2503.05788

Boonpracha, J. (2021). Creative cultural product design in the bicultural context [Publisher: Vilnius Gediminas Technical University]. *Creativity Studies*, *14*(2), 336–345. https://doi.org/10.3846/cs.2021.14290

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., . . . Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (pp. 1877–1901, Vol. 33). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

Capello, R., Cerisola, S., & Perucca, G. (2019, December). Cultural heritage, creativity, and local development: A scientific research program [ISSN: 2198-7300, 2198-7319]. In *Research for development* (pp. 11–19). Springer International Publishing. https://doi.org/10.1007/978-3-030-33256-3_2

CHEN, B.-C., Zhang, Z., Langrené, N., & Zhu, S. (2023, January). Unleashing the potential of prompt engineering in Large Language Models: A comprehensive review. https://doi.org/10.48550/arxiv.2310.14735

Chhikara, G., Kumar, A., & Chakraborty, A. (2025). Through the prism of culture: Evaluating llms' understanding of indian subcultures and traditions [Publisher: Cornell University]. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2501.16748

Chua, R. Y. J., Roth, Y., & Lemoine, J.-F. (2014). The impact of culture on creativity [Publisher: SAGE Publishing]. *Administrative Science Quarterly*, *60*(2), 189–227. https://doi.org/10.1177/0001839214563595

Cochran, W. G. (1977). *Sampling techniques*. john wiley & sons.

Cornish, F., & Gillespie, A. (2009). A Pragmatist Approach to the Problem of Knowledge in Health Psychology [Publisher: SAGE Publications Ltd]. *Journal of Health Psychology*, *14*(6), 800–809. https://doi.org/10.1177/1359105309338974

Creswell, J. W., & Clark, V. L. P. (2017). *Designing and Conducting Mixed Methods Research*. SAGE Publications.

Cropley, D. H., & Kaufman, J. C. (2012). Measuring functional creativity: Non-expert raters and the creative solution diagnosis scale [Publisher: Wiley]. *The Journal of Creative Behavior*, *46*(2), 119–137. https://doi.org/10.1002/jocb.9

Dai, D. Y., Cheng, H. T., & Yang, P. (2019, April). QEOSA: A pedagogical model that harnesses cultural resources to foster creative problem-solving [ISSN: 1664-1078 Volume: 10]. https://doi.org/10.3389/fpsyg.2019.00833

Dean, D. L., Hender, J., Rodgers, T. L., & Santanen, E. (2006). Identifying quality, novel, and creative ideas: Constructs and scales for idea evaluation [Publisher: Association for Information Systems]. *Journal of the Association for Information Systems*, *7*(10), 646–699. https://doi.org/10.17705/1jais.00106

Du, W., Advani, L., Gambhir, Y., Perry, D. J., Shiralkar, P., Xing, Z., & Colak, A. (2023). Effective proxy for human labeling: Ensemble disagreement scores in large language models for industrial NLP [Publisher: Cornell University]. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2309.05619

Feng, S., Chan, W.-C., Chouhan, S., Ayala, J. F. G., Medicherla, S., Clark, K., & Shi, M. (2025). Whispers of many shores: Cultural alignment through collaborative cultural expertise [Publisher: arXiv]. https://doi.org/10.48550/ARXIV.2506.00242

Gazzaroli, D., Gozzoli, C., & Sánchez-Gardey, G. (2019). The living and working together perspective on creativity in organizations [Publisher: Frontiers Media]. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.02733

Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H.-Y., Wang, Y., & Guo, J. (2024). A survey on LLM-as-a-judge [Publisher: Cornell University]. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2411.15594

Guilford, J. P. (1950). Creativity. [Publisher: American Psychological Association]. *American Psychologist*, *5*(9), 444–454. https://doi.org/10.1037/h0063487

Guilford, J. P. (2016, January). Characteristics of creativity [Pages: 22-34]. https://doi.org/10.5040/9798400621086.ch-003

Gunasekara, N., Barhate, B., Alizadeh, A., & Capuchino, R. G. (2022). A human resources development professional's framework for competencies during COVID-19 and unrest [Publisher: SAGE Publishing]. *New Horizons in Adult Education and Human Resource Development*, *34*(2), 37–43. https://doi.org/10.1002/nha3.20350

Hampson, T., & McKinley, J. (2023). Problems posing as solutions: Criticising pragmatism as a paradigm for mixed research [Publisher: SAGE Publications Ltd STM]. *Research in Education*, *116*(1), 124–138. https://doi.org/10.1177/00345237231160085

Hofstede, G. (2011). Dimensionalizing cultures: The hofstede model in context. *Online Readings in Psychology and Culture*, *2*(1). https://doi.org/10.9707/2307-0919.1014

Hofstede, G., & Bond, M. H. (1984). Hofstede's culture dimensions [Publisher: SAGE Publishing]. *Journal of Cross-Cultural Psychology*, *15*(4), 417–433. https://doi.org/10.1177/0022002184015004003

Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2020, February). The Curious Case of Neural Text Degeneration [arXiv:1904.09751 [cs]]. https://doi.org/10.48550/arXiv.1904.09751

Hughes, D. J., Lee, A., Tian, A. W., Newman, A., & Legood, A. (2018, March). Leadership, creativity, and innovation: A critical review and practical recommendations [ISSN: 1048-9843, 1873-3409 Issue: 5 Pages: 549-569 Volume: 29]. https://doi.org/10.1016/j.leaqua.2018.03.001

Ismona, I., & Marwan, M. (2020). Entrepreneurial creativity in the activities of the catering production unit in vocational high school. *Proceedings of the 4th Padang International Conference on Education, Economics, Business and Accounting (PICEEBA-2 2019)*. https://doi.org/10.2991/aebmr.k.200305.126

Ivankova, N. V., Creswell, J. W., & Stick, S. L. (2006). Using Mixed-Methods Sequential Explanatory Design: From Theory to Practice [Publisher: SAGE Publications Inc]. *Field Methods*, *18*(1), 3–20. https://doi.org/10.1177/1525822X05282260

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed Methods Research: A Research Paradigm Whose Time Has Come [Publisher: American Educational Research Association]. *Educational Researcher*, *33*(7), 14–26. https://doi.org/10.3102/0013189X033007014

Kaushik, V., & Walsh, C. A. (2019). Pragmatism as a Research Paradigm and Its Implications for Social Work Research [Number: 9 Publisher: Multidisciplinary Digital Publishing Institute]. *Social Sciences*, *8*(9), 255. https://doi.org/10.3390/socsci8090255

Kelly, L. M., & Cordeiro, M. (2020). Three principles of pragmatism for research on organizational processes [Publisher: SAGE Publications Ltd]. *Methodological Innovations*, *13*(2), 2059799120937242. https://doi.org/10.1177/2059799120937242

Kim, S., & Oh, D. (2025). Evaluating creativity: Can llms be good evaluators in creative writing tasks? [Publisher: Multidisciplinary Digital Publishing Institute]. *Applied Sciences*, *15*(6), 2971–2971. https://doi.org/10.3390/app15062971

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Lamiño, P., & Diaz, J. (2024). Intercultural competencies: Understanding high- vs. low-context cultures [Publisher: George A. Smathers Libraries]. *EDIS*, *2024*(6). https://doi.org/10.32473/edis-wc475-2024

Li, A., Chen, H., Namkoong, H., & Peng, T. (2025). LLM generated persona is a promise with a catch [Publisher: arXiv]. https://doi.org/10.48550/ARXIV.2503.16527

Liu, J. (2019). Research on the development strategy of cultural and creative products based on regional culture. *Proceedings of the 2019 3rd International Seminar on Education, Management and Social Sciences (ISEMSS 2019)*. https://doi.org/10.2991/isemss-19.2019.92

Maarouf, H. (2019). Pragmatism as a Supportive Paradigm for the Mixed Research Approach: Conceptualizing the Ontological, Epistemological, and Axiological Stances of Pragmatism [Number: 9]. *International Business Research*, *12*(9), p1. https://doi.org/10.5539/ibr.v12n9p1

McCarthy, M. (2019). Cross-cultural differences in creativity: A process-based view through a prism of cognition, motivation and attribution [Publisher: Elsevier BV]. *Thinking Skills and Creativity*, *32*, 82–91. https://doi.org/10.1016/j.tsc.2019.04.002

McIntosh, T. R., Sušnjak, T., Liu, T., Watters, P., & Halgamuge, M. N. (2024). Inadequacies of large language model benchmarks in the era of generative artificial intelligence [Publisher: Cornell University]. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2402.09880

Mhlongo, N. Z., Olatoye, F. O., Elufioye, O. A., Ibeh, C. V., Falaiye, T., & Daraojimba, A. I. (2024, February). Cross-cultural business development strategies: A review of USA and african [ISSN: 2582-8185 Issue: 1 Pages: 1408-1417 Volume: 11]. https://doi.org/10.30574/ijsra.2024.11.1.0233

Morgan, D. L. (2014). Pragmatism as a Paradigm for Social Research [Publisher: SAGE Publications Inc]. *Qualitative Inquiry*, *20*(8), 1045–1053. https://doi.org/10.1177/1077800413513733

Mukherjee, A., & Chang, H. H. (2023). The creative frontier of generative AI: Managing the novelty-usefulness tradeoff [Publisher: Cornell University]. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2306.03601

Mukherjee, S. (2023, June). Unveiling perplexity: Measuring success of llms and generative AI models. https://ramblersm.medium.com/the-significance-of-perplexity-in-evaluating-llms-and-generative-ai-62e290e791bc

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Barnes, N., & Mian, A. (2023). A comprehensive overview of large language models [Publisher: Cornell University]. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2307.06435

Pásztor, A., Molnár, G., & Csapó, B. (2015). Technology-based assessment of creativity in educational context: The case of divergent thinking and its relation to mathematical achievement [Publisher: Elsevier BV]. *Thinking Skills and Creativity*, *18*, 32–42. https://doi.org/10.1016/j.tsc.2015.05.004

Pawar, S. (2025, March). Why I chose cosine similarity & GPT evaluation over BLEU, ROUGE, and METEOR for LLM response evaluation. https://pub.towardsai.net/why-i-chose-cosine-similarity-gpt-evaluation-over-bleu-rouge-and-meteor-for-llm-response-67fdd67eab88?gi=a4eab63f58e7

Peeperkorn, M., Kouwenhoven, T., Brown, D., & Jordanous, A. (2024, May). Is Temperature the Creativity Parameter of Large Language Models? [arXiv:2405.00492 [cs]]. https://doi.org/10.48550/arXiv.2405.00492

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research [Publisher: Taylor & Francis]. *Journal of Management Information Systems*, *24*(3), 45–77. https://doi.org/10.2753/mis0742-1222240302

Peng, J.-L., Cheng, S., Diau, E., Shih, Y.-Y., Chen, P.-H., Lin, Y.-T., & Chen, Y.-N. (2024). A survey of useful LLM evaluation [Publisher: Cornell University]. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2406.00936

Proctor, R. A. (1991). The importance of creativity in the management field [Publisher: Wiley]. *British Journal of Management*, *2*(4), 223–230. https://doi.org/10.1111/j.1467-8551.1991.tb00028.x

Rabeyah, A. A., Góes, L. F. W., Volpe, M., & Medeiros, T. (2024). Do llms agree on the creativity evaluation of alternative uses? [Publisher: Cornell University]. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2411.15560

Radharapu, B., Revel, M., Ung, M., Ruder, S., & Williams, A. (2025). Arbiters of ambivalence: Challenges of using llms in noconsensus tasks [Publisher: arXiv]. https://doi.org/10.48550/ARXIV.2505.23820

Raiaan, M. A. K., Mukta, M. S. H., Fatema, K., Fahad, N. M., Sakib, S., Mim, M. M. J., Ahmad, J., Ali, M. E., & Azam, S. (2024, January). A review on large language models: Architectures, applications, taxonomies, open issues and challenges [ISSN: 2169-3536 Journal abbreviation: IEEE Access Pages: 26839-26874 Volume: 12]. https://doi.org/10.1109/access.2024.3365742

Rashid, A. (2024, March). *Untitled* (tech. rep.). https://doi.org/10.55277/researchhub.vq5dnd6h

Raza, M., Jahangir, Z., Riaz, M. B., Saeed, M. J., & Sattar, M. A. (2025). Industrial applications of large language models [Publisher: Nature Portfolio]. *Scientific Reports*, *15*(1). https://doi.org/10.1038/s41598-025-98483-1

Runco, M. A. (1993). Divergent thinking, creativity, and giftedness [Publisher: SAGE Publishing]. *Gifted Child Quarterly*, *37*(1), 16–22. https://doi.org/10.1177/001698629303700103

Schneider, J., & Basalla, M. (2020). Creativity of deep learning: Conceptualization and assessment [Publisher: Cornell University]. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2012.02282

Schubert, E. (2021). Creativity is optimal novelty and maximal positive affect: A new definition based on the spreading activation model [Publisher: Frontiers Media]. *Frontiers in Neuroscience*, *15*. https://doi.org/10.3389/fnins.2021.612379

Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., Li, Y., Gupta, A., Han, H., Schulhoff, S., Dulepet, P. S., Vidyadhara, S., Ki, D., Agrawal, S., Pham, C., Kroiz, G., Li, F., Tao, H., Srivastava, A., ... Resnik, P. (2024). The prompt report: A systematic survey of prompting techniques [Publisher: Cornell University]. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2406.06608

Shamay-Tsoory, S. G., Adler, N., Aharon-Peretz, J., Perry, D., & Mayseless, N. (2010). The origins of originality: The neural bases of creative thinking and originality [Publisher: Elsevier BV]. *Neuropsychologia*, *49*(2), 178–185. https://doi.org/10.1016/j.neuropsychologia.2010.11.020

Shao, Y., Zhang, C., Zhou, J., Gu, T., & Yuan, Y. (2019, May). How does culture shape creativity? A mini-review [ISSN: 1664-1078 Volume: 10]. https://doi.org/10.3389/fpsyg.2019.01219

Smajic, E., Avdic, D., Pasic, A., Prcic, A., & Stancic, M. (2022). Mixed Methodology of Scientific Research in Healthcare. *Acta Informatica Medica*, *30*(1), 57. https://doi.org/10.5455/aim.2022.30.57-60

Soares, A. M., Farhangmehr, M., & Shoham, A. (2006). Hofstede's dimensions of culture in international marketing studies [Publisher: Elsevier BV]. *Journal of Business Research*, *60*(3), 277–284. https://doi.org/10.1016/j.jbusres.2006.10.018

Sternberg, R. J., & Grigorenko, E. L. (2001). Guilford's structure of intellect model and model of creativity: Contributions and limitations [Publisher: Taylor & Francis]. *Creativity Research Journal*, *13*, 309–316. https://doi.org/10.1207/s15326934crj1334_08

Stureborg, R., Alikaniotis, D., & Suhara, Y. (2024). Large language models are inconsistent and biased evaluators [Publisher: Cornell University]. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2405.01724

Su, J., Jiang, C., Jin, X., Qiao, Y., Xiao, T., Ma, H., Wei, R. Q., Jing, Z., Xu, J., & Lin, J. (2024). Large language models for forecasting and anomaly detection: A systematic literature review [Publisher: Cornell University]. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2402.10350

Sukiennik, N., Gao, C., Xu, F., & Li, Y. (2025). An evaluation of cultural value alignment in LLM [Publisher: arXiv]. https://doi.org/10.48550/ARXIV.2504.08863

Tanti, T., Maison, M., Syefrinando, B., Daryanto, M., & Salma, H. (2020). Students' self-regulation and motivation in learning science [Number: 4]. *International Journal of Evaluation and Research in Education (IJERE)*, *9*(4), 865–873. https://doi.org/10.11591/ijere.v9i4.20657

Tarasov, A. (2022, February). Why artificial intelligence lacks creativity and what can de done to help it. https://www.unite.ai/why-artificial-intelligence-lacks-creativity-and-what-can-de-done-to-help-it/

Types of divergent thinking. (2024, August). https://www.canr.msu.edu/resources/divergent-thinking-types

Vedel, I., Kaur, N., Hong, Q. N., El Sherif, R., Khanassov, V., Godard-Sebillotte, C., Sourial, N., Yang, X. Q., & Pluye, P. (2019). Why and how to use mixed methods in primary health care research. *Family Practice*, *36*(3), 365–368. https://doi.org/10.1093/fampra/cmy127

Wafa, A. W. A., & Hussain, M. H. M. (2021, June). A literature review of artificial intelligence [ISSN: 2791-1268, 2791-1276 Issue: 1 Pages: 1-1 Volume: 1]. https://doi.org/10.32350/air.11.01

Wang, H., Zou, J., Mozer, M. C., Zhang, L., Goyal, A., Lamb, A., Deng, Z., Xie, M. Q., Brown, H. A., & Kawaguchi, K. (2024). Can AI be as creative as humans? [Publisher: Cornell University]. *arXiv (Cornell University)*. https://doi.org/10.48550/arXiv.2401.01623

Wang, Y.-H., & Ajovalasit, M. (2020). Involving cultural sensitivity in the design process: A design toolkit for chinese cultural products [Publisher: Wiley]. *International Journal of Art & Design Education*, *39*(3), 565–584. https://doi.org/10.1111/jade.12301

Wei, J., Wang, X., Schuurmans, D., Bosma, M., ichter brian, b., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in neural information processing systems* (pp. 24824–24837, Vol. 35). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with ChatGPT [Publisher: Cornell University]. *arXiv (Cornell University)*. https://doi.org/10.48550/arXiv.2302.11382

Wu, T., Terry, M., & Cai, C. J. (2021). AI chains: Transparent and controllable human-AI interaction by chaining large language model prompts [Publisher: Cornell University]. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2110.01691

Wu, Z., Ji, D., Yu, K., Zeng, X., Wu, D., & Shidujaman, M. (2021, January). AI creativity and the human-AI co-creation model [ISSN: 0302-9743, 1611-3349]. In *Lecture notes in computer science* (pp. 171–190). Springer Science+Business Media. https://doi.org/10.1007/978-3-030-78462-1_13

Yang, X., Chen, A., PourNejatian, N., Shin, H. C., Smith, K. E., Parisien, C., Compas, C. B., Martin, C., Costa, A., Flores, M. G., Zhang, Y., Magoč, T., Harle, C. A., Lipori, G., Mitchell, D. A., Hogan, W. R., Shenkman, E., Bian, J., & Wu, Y. (2022). A large language model for electronic health records [Publisher: Nature Portfolio]. *npj*

Digital Medicine, *5*(1). https://doi.org/10.1038/s41746-022-00742-2

Yong, K., Mannucci, P. V., & Lander, M. (2020). Fostering creativity across countries: The moderating effect of cultural bundles on creativity [Publisher: Elsevier BV]. *Organizational Behavior and Human Decision Processes*, *157*, 1–45. https://doi.org/10.1016/j.obhdp.2019.12.004

Zhao, Y., Zhang, R., Li, W., Huang, D., Guo, J., Peng, S., Hao, Y., Wen, Y., Hu, X. J., Du, Z., Guo, Q., Li, L., & Chen, Y. (2024). Assessing and understanding creativity in large language models [Publisher: Cornell University]. *arXiv (Cornell University)*. https://doi.org/10.48550/arXiv.2401.12491