Do Persona-Infused LLMs Affect Performance in a Strategic Reasoning Game?

John Licato and Stephen Steinle
Bellini College of Artificial Intelligence,
Cybersecurity and Computing
University of South Florida

Brayden Hollis
Information Directorate
Air Force Research Library

Abstract

Although persona prompting in large language models appears to trigger different styles of generated text, it is unclear whether these translate into measurable behavioral differences, much less whether they affect decision-making in an adversarial strategic environment that we provide as open-source. We investigate the impact of persona prompting on strategic performance in PERIL, a world-domination board game. Specifically, we compare the effectiveness of persona-derived heuristic strategies to those chosen manually. Our findings reveal that certain personas associated with strategic thinking improve game performance, but only when a mediator is used to translate personas into heuristic values. We introduce this mediator as a structured translation process, inspired by exploratory factor analysis, that maps LLM-generated inventory responses into heuristics. Results indicate our method enhances heuristic reliability and face validity compared to directly inferred heuristics, allowing us to better study the effect of persona types on decision-making. These insights advance our understanding of how persona prompting influences LLM-based decision-making and propose a heuristic generation method that applies psychometric principles to LLMs.

1 Introduction

"If you would read a [person's] Disposition, see him Game, you will then learn more of him in one hour, than in seven Years Conversation," according to a letter of advice written over 300 years ago (Lingard and Erb, 1907). If this advice is correct, perhaps nowhere is one's personality more apparent than in strategic adversarial games, where individual behavioral tendencies such as aggression, patience, caution, and others dictate the heuristics that guide players' decision-making. Such settings present a unique opportunity to study the relationship between how modern large language models (LLMs) relate personality descriptions (often called *personas*) and decision-making in strategic environments.

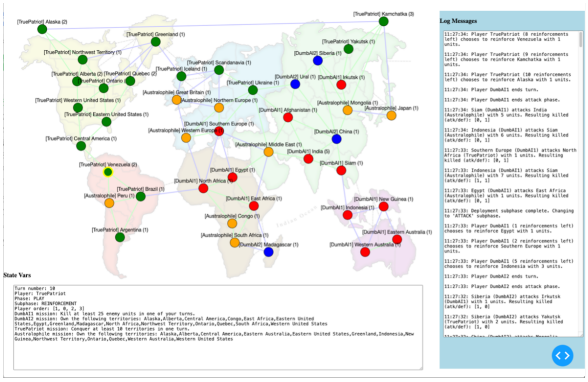

Figure 1: PERIL, our implementation inspired by a popular strategic world conquest board game, was used to study the effects of persona-based prompting in LLMs.

In this paper, we investigate whether personality traits inferred from prompts reliably translate into actionable heuristics in a strategy board game. Strategic reasoning is a critical capability for advancing AI in decision-making and human-machine collaboration. Beyond gaming, the findings have broader implications for simulation, training, and the development of automated systems requiring strategic adaptability. Such work contributes to advancing AI's role in team-based environments, military simulations, and other domains where human-like variability and strategic decision-making are essential.

Isolating strategic reasoning ability is difficult, especially when that strategy must apply to an environment that has a large search space, nondeterministic outcomes, and high-pressure conditions that require rapid decision-making under time and computation constraints. In such environments, strategic actors may benefit from *precommitment*, or the fixing of rules and heuristics that will constrain one's behavior ahead of time, so that cog-

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 3497–3528

nitive overhead, behavioral inconsistencies, and the effect of time pressure on decisions will be minimized later. In this paper, we will focus on this aspect of strategic thinking by studying the performance of LLMs in a variant of a popular strategy board game. More specifically, we will study the effect of *persona prompting*, a prompting strategy in which a pre-trained LLM is prompted with a description of a personality and asked to act in accordance with it. We do this through a fixed set of heuristics tailored to the PERIL environment, which should be understood as design choices rather than a comprehensive taxonomy of player attributes. Despite its promise, the effect of persona prompting on tasks requiring strategic reasoning, particularly in dynamic and uncertain environments, remains underexplored.

We primarily address two research questions: **(1)** Does persona prompting using personalities with traits associated with strategic reasoning improve performance on strategic games? **(2)** Does using a personality inventory to translate persona descriptions into heuristics lead to decision-making heuristics with more face validity?

Novel Contributions and Summary of Findings

- This is the first work to specifically study the effect of persona prompting on decision-making in a strategic reasoning game. We found a positive relationship between personality traits that intuitively would lead to better performance in the game and actual game performance, thus contributing to the ongoing research on how and when to use persona prompting.

- We introduce PERIL, a new platform for evaluating strategic decision-making capabilities of AI players. In this paper, we compare the effects of persona prompting on players with the same mission, but the platform we implemented allows for multiple missions (which can in the future be used for studying strategic deception). We will make our full source code and platform available.

- We introduce the use of personality inventory questionnaires to translate personas into heuristic choices in an end-to-end fashion. We observe that this method results in heuristics that align with features of those personalities (much more so than when the questionnaire is not used). Without it the variation in how each persona translates into heuristics is small, suggesting that persona prompting alone does not lead to significant behavior differences.

2 Related Work

In recent years, pre-trained LLMs have become increasingly difficult to fine-tune, due to a combination of model size, computation requirements, and reduced access to pre-trained models' weights. As a result, many researchers have turned to strategies exploring the extent to which prompts can be adjusted to improve performance. An approach rapidly gaining popularity is based on the concept of the *persona*, where a personality description is provided to the LLM, and it is asked to act in accordance with that personality (Tseng et al., 2024; Zhang et al., 2024; Bhandari et al., 2025a). New frameworks are rapidly emerging to compare different persona prompts on a variety of tasks (Pan et al., 2024; Lin et al., 2024; Samuel et al., 2024; Liu et al., 2024; Potertì et al., 2025), and datasets of high-quality LLM-generated personas are now available (Schuller et al., 2024; Chan et al., 2024; Wang et al., 2025).

However, it is unclear how well persona prompting can affect performance in action spaces. There are early results showing how LLMs can play a role in the command and conquer domain space (Goecks and Waytowich, 2024) and in strategic board games (Meta Fundamental AI Research Diplomacy Team (FAIR) et al., 2022; Lorè and Heydari, 2024; Hu et al., 2024; Belle et al., 2025), but very little studying the effects of persona prompting. There are mixed results on its general effectiveness: (Liu et al., 2024) used multiple personas to improve scientific ideation. (Kamruzzaman and Kim, 2024) found that nationality-based persona prompting introduced preference biases towards the assigned country of origin. And other authors have found that multi-agent systems consisting of multiple personas working together can produce differences in reasoning, hallucination rate, and more (Sreedhar and Chilton, 2024; Wang et al., 2024; Olea et al., 2024; Jiang et al., 2025).

Strangely, not all recent work agrees that persona prompting has a significant effect, or even an effect that would be expected given the persona. For example, (Hu and Collier, 2024) find that persona variables accounted for less than 10% of annotation variance in subjective NLP datasets. (Kim et al.,

2024) found that role-playing prompts *decreased* reasoning abilities in some datasets. And (Zheng et al., 2024) find that on a subset of the MMLU benchmark tasks (Hendrycks et al., 2021), the use of personas does not seem to improve performance beyond random chance. Conflicting results can appear even in the same study, suggesting that persona prompting results are very sensitive to prompt structure (Phelps and Russell, 2025). Further clouding the effective application of personas is their use in so-called psychological inventories (Li et al., 2025; Wu et al., 2025). While there is plenty of quality literature on the reliability of using personas in inventories (Bhandari et al., 2025b; Frisch and Giulianelli, 2024), there is significantly less work exploring the validity of such inventory results (Maharjan et al., 2025). Another relevant line of work examines how persona prompting interacts with broader tendencies to anthropomorphize LLMs. Researchers have proposed multi-level frameworks describing how human-like qualities are attributed across perceptual, linguistic, and cognitive dimensions (Xiao et al., 2025), as well as taxonomies of linguistic expressions that make language technologies appear more human-like (DeVrio et al., 2025). These works address that part of the difficulty in interpreting persona effects may stem from how human expectations shape prompts. This is without even addressing the central concern of what theoretical foundation such results rest upon (even assuming reliability and validity) since the responses generated by LLMs do not conform to human distributions (Pratelli and Petrocchi, 2025).

Given this gap in the literature, it is clear more work is needed to observe the effects of persona prompting in new scenarios, so that a deeper theory can be developed. Drawing on work showing that the use of personality questionnaires can improve LLM performance (Bodroža et al., 2024; Jiang et al., 2024), we set out to study the extent to which persona-based prompting, augmented with personality questionnaires, can affect performance in a game of strategic reasoning, as well as gain deeper insight into how LLMs translate personas into decisions. As discussed in the previous paragraph, we do not make claims about the theory behind *why* the LLM makes generations as it does, but rather to show that personas *do* result in meaningful changes and that an exploratory factor analysis method bears fruit when applied to LLM reasoning.

3 Experimental Setup

Our experimental design has four components: game environment, persona selection, heuristic generation, and tournament. The game environment introduces our PERIL software and the set of heuristics used to guide play. Persona selection explains how we sampled personas and identified the subset used in experiments. Heuristic generation describes how we constructed a questionnaire with GPT-4 and used it to map persona responses into heuristic values. Finally, tournament games outlines how matches were run and how performance metrics were collected.

3.1 Game Environment

For our strategic reasoning test environment, we implemented a game loosely inspired by the board game Risk®, which we call PERIL.[1] In our game, up to six players control *units* on a map of *regions* divided into *zones*. The objective of the game is to achieve an assigned mission, but for the present work all players have the same mission, in order to minimize confounding factors: to achieve world domination by occupying all regions. The regions are connected to each other and either connect over land or water. Players control a set of units representing armies. Every region must have at least one unit on it (except in the beginning of the game, when no regions have any). Regions can contain an unlimited number of units, but all units on a single region must belong to the same player.

The game starts in an initialization phase, where all players are given a number of units. They alternate, placing one unit on an unoccupied region at a time, until all regions are occupied, and then all units initially given to each player are placed. Each player then takes their regular turns. Each turn is broken up into three subphases: reinforcement (players receive additional units they can place), attack (players decide which regions to attack, where they are only able to attack regions adjacent to those they control), and redeployment (players can move units to connected regions).

Using Python and the Dash Cytoscape library (Inc., 2018), we implemented a framework for creating custom maps and allowing AI-controlled players to compete (Figure 1).[2] We also implemented the ability to assign different missions to players,

[1] PERIL stands for **P**lease **E**veryone, we're **R**epelling **I**nfringement **L**awsuits.

[2] https://github.com/Advancing-Machine-Human-Reasoning-Lab/PERIL

Table 1: Decision-Making Heuristics Available in PERIL

PHASE	
Initialization / Deployment Phase	
PTM/PTL	Place a unit in a region T_1 that is adjacent to region T_2 if T_2 is owned by the player with the most (PTM) or least (PTL) regions.
PUM/PUL	Place a unit in a region T_1 that is adjacent to region T_2 if T_2 is owned by the player with the most (PUM) or least (PUL) units.
PCM/PCL	Place a unit in a region T_1 that is adjacent to region T_2 if T_2 is owned by the player with the most (PCM) or least (PCL) zones owned (measured by total zone bonuses).
ETE/ETN	Place a unit in a region if it is adjacent to an enemy region (ETE) / not adjacent to any enemy regions (ETN).
EAC	Place a unit in a region if it is on a zone boundary.
EACM/EACL	Place a unit in a region if it is adjacent to the largest (EACM) / smallest (EACL) zone.
EACO	Place a unit in a region if it is adjacent to a zone that is completely owned by an enemy player.
Attack Phase	
PTM/PTL	Attack from region T_1 to T_2 if T_2 is owned by the player with the most (PTM) or least (PTL) regions.
PUM/PUL	Attack from region T_1 to T_2 if T_2 is owned by the player with the most (PUM) or least (PUL) units.
PCM/PCL	Attack from region T_1 to T_2 if T_2 is owned by the highest (PCM) or lowest (PCL) number of zones owned (measured by total zone bonuses).
ONM/ONL	Attack if the units in T_1 are greater (ONM) / fewer (ONL) than the units in T_2.
ON2	Attack if the units in T_1 are at least twice the number of units in T_2.
ICD/ICS	Attack from T_1 to T_2 if they are (ICD) / are not (ICS) in different zones.
L	Attack from T_1 to T_2 if T_2 connects T_1 to a region you own that it isn't currently connected to.
PASS	Likelihood of passing, ending your turn without more attacks. If set to 100, you will never attack; if set to 0, you will always attack.
Redeployment Phase	
OBTM/OBTL	Move from T_1 to T_2 if T_2 is adjacent to more regions occupied by the player with the most (OBTM) or least (OBTL) regions.
OBUM/OBUL	Move from T_1 to T_2 if T_2 is adjacent to more regions occupied by the player with the most (OBUM) or least (OBUL) units.
OBCM/OBCL	Move from T_1 to T_2 if T_2 is adjacent to more regions occupied by the player with the most (OBCM) or least (OBCL) zones owned (measured by total zone bonuses).
CNM/CNL	Move from T_1 to T_2 if T_2 is connected to more (CNM) / fewer (CNL) regions.
CB	Move from T_1 to T_2 if T_2 is on a zone boundary and T_1 is not.
CA	Move from T_1 to T_2 if T_2 is adjacent to at least one region occupied by an enemy player and T_1 is not.
CAC	Move from T_1 to T_2 if T_2 is on a zone boundary and T_1 is not.
M/L	Move from T_1 to T_2 if T_2 has more (M) / fewer (L) units on it.
SI	Move from T_1 to T_2 if T_2 is adjacent to a region that has a higher chance of successful invasion than those connected to T_1, calculated using the ratio of available troops from attacking region to troops on target region.
PASS	Likelihood of passing, ending your turn without more redeployments. If set to 100, you will never redeploy; if set to 0, you will always redeploy.

although that feature was not utilized in this paper. AI-controlled players can access the entire game state at any point of the game directly, but for this paper we focused on the setting of strategic priorities in the form of *heuristics*, implemented as follows: At the beginning of each game, the AI-controlled player is given its assigned persona and a list of allowed heuristics in the game. It is then asked to assign priorities for each of these heuristics (more details on how these priorities are selected in §3.2). During the game, at each game phase the set of allowed moves are enumerated and assigned point values based on the heuristics selected by the player. The move is then selected using a random selection with the point value as the weight. By default, all heuristics have the point value of 5. The set of heuristics is in Table 1.

3.2 Persona Selection

To select our personas, we started with Persona Hub (Chan et al., 2024), a collection of one billion diverse, synthetically-generated persona descriptions. Each persona consists of a short personality description, e.g., "An elderly woman who loves watching makeup tutorials online and enjoys discussing beauty products." We randomly-selected 175,000 personas and annotated them using GPT-4o-mini, by asking it to independently rate each persona on the following features that we hypothesized would be predictive of game performance:

- **strategicThinker** a strategic thinker.
- **domainExpert** someone who has experience in combat, military warfare, or similar areas of expertise.

- **perilSpecific** someone who is likely to perform well on the game of PERIL specifically.
- **riskTaker** someone who is likely to take risks.
- **doOrBe** an instruction to act a certain way by describing actions (do) or if it is an instruction to play a role by describing a personality or character background (be).

Instructions were provided to rate all features on a scale from -1 to +1, in increments of 0.5. A qualitative description for each rating was provided to increase reliability. These ratings serve as a reference for evaluating how well heuristic generation methods correlate with persona characteristics, but they are not intended as ground-truth measures of personality. We then used a greedy algorithm to find the 50 personas that maximize the product of the variances across all features F:

$$\arg\max_{P} \prod_{f \in F} \sum_{p \in P} (f_p - \overline{f_P})^2 \quad (1)$$

Where f_p is the rating of feature f given to persona p in subset P, and $\overline{f_P}$ is the average f rating of all items in P. To find a greedy approximation, for each persona p in the original set of 175K personas, we started with the set $\{p\}$ and iteratively added the persona that most increased Equation 1, until we reached the maximum set size of 50. This was repeated for all possible personas i as starting points. The best set P is the final set of personas we will refer to for the remainder of this paper.

3.3 Heuristic Selection

Four LLMs were selected to generate heuristics: gpt-4o-2024-11-20, Meta-Llama-3-8B-Instruct, Llama-4-Maverick-17B-128E-Instruct-FP8, and Mistral-Small-Instruct-2409. These models were selected due to their relevance in current literature (GPT4) and to examine how model size impacts heuristic generation (Llama). Mistral small was chosen to provide an additional small sized model. Non-GPT models were downloaded from HuggingFace[3] and run with stock configurations on 6 H100 GPUs. Two strategies were used to convert personas into heuristic values. In the first (which we will call the "**direct heuristic**" (DH) players), the LLM was prompted with the instructions of the game, a list of the heuristics available for a given game phase, and an example of how to provide values for each heuristic (ranging from 0 to 100, with a default value of 5). However, this generation method can lead to inconsistencies. For example, two similar personas may lead to proportional but different values assigned to each heuristic. Furthermore, the same persona may manifest in different ways in the three different phases, motivating a need to have personality features translate more evenly into heuristic values.

For this reason, our second strategy (which we will call the "**personality inventory**" (PI) players) utilizes a personality inventory, inspired by the American Psychiatric Association's Personality Inventory for DSM-5 (PID-5)—Adult (Krueger et al., 2012). A set of question items are provided, each with a first-person statement (e.g., "I deserve special treatment.") and the option to select how true this statement is: Very false or often false (-2 points), Sometimes or somewhat false (-1 points), Sometimes or somewhat true (+1 points), Very true or often true (+2 points). Each item maps either positively or negatively to some heuristics. For example, one item is "I prefer to spread my influence to less contested or peripheral regions." This maps positively to heuristics **PTL, PUL, PCL, ETN**, and **EACL**. It maps negatively to **PTM, PUM, PCM, ETE**, and **EACM**. If a LLM responds to this item with "very true or often true", then 2 points will be added to heuristics in the positive set, and 2 points will be subtracted from the negative set.

To generate the PI heuristics, we followed three steps. First, we used an LLM (GPT-4) to draft a questionnaire inventory. Second, we manually curated this inventory to ensure clarity and consistency. Third, we prompted an LLM to answer each item in the questionnaire using a persona (note that heuristic information and mappings were not provided at this stage). These point values were then converted into heuristic weights in a range matching that of the direct heuristic players. For some heuristic value H, let $r(H)$ be the number of points assigned to that heuristic divided by the maximum amount of possible points that could have been assigned to it. Then the weight $w(H)$ is:

$$\begin{array}{ll} \max(0, \lambda * (r(H)/5) + 5) & r(H) \leq 0 \\ \min(100, \lambda * (95 * r(H)) + 5) & r(H) > 0 \end{array}$$

This transformation ensures that if $r(H) = 0$, then $w(h) = 5$, and $0 \leq r(H) \leq 100$, since 5 was the default value in the prompt. For the experiments in this paper, we use $\lambda = 0.5$, which made the average heuristic value across all personas of personality

[3] https://huggingface.co/models

inventory players approximately equal to that of the direct heuristic players.

3.4 Tournament Games

To compare players, we set up a series of matches between the fifty personas. In each round, the 50 players were paired up randomly. When all 25 games are played, the players are paired up again for another round. This process is repeated for 49 rounds per run (1225 games). We carried out two runs with the 50 personality inventory players, and two additional runs with the 50 direct heuristic players. To measure player skill level, we use the TrueSkill algorithm (Herbrich et al., 2006), a generalization of the more well-known ELO score, allowing for games involving more than two players at a time (although we do not use that feature in this paper). If a game extended past 250 turns, it is declared a draw and all players are counted as having lost for the purpose of the TrueSkill calculation (this occurred in $< 0.05\%$ of games). We use two-player games to reduce the confounding effect of larger numbers of players, but note that many of the heuristics we defined (Table 1) only produce observably different behaviors when there are more than two players total. Nevertheless, we retain them here to study how persona-prompted LLMs select values for those heuristics.

4 Results

Our results can be summarized in four main findings, presented in the order they appear below. First, we compare player performance, showing that strategically themed personas (e.g., strategists, officers, athletes) consistently achieve higher TrueSkill scores than other personas (e.g., students or children; Table 2). Second, we analyze correlations between inferred persona features and generated heuristics, finding that the PI method produces significant correlations with features such as *strategicThinker*, *domainExpert*, and *perilSpecific*, whereas the DH method shows weak or inconsistent correlations (Table 3). Third, we examine cross-run reliability of final rankings, finding that DH players were more stable across trials, but this stability was not tied to persona features. Finally, we evaluate opposite-value consistency in heuristic assignments, where GPT-4 generated coherent mappings under the PI method, while smaller models produced noisier and less reliable patterns (Table 4). Overall, these results demonstrate that the PI method yields more distinct and interpretable persona-driven behaviors than the DH method.

4.1 Comparison of Personality Inventory Players

The relationship between actual performance and inferred personality features are apparent when comparing the top 5 and bottom 5 performers of the first run of personality inventory players (Table 2). To determine whether a relationship existed between inferred personality features and final TrueSkill score, we calculated the Spearman correlation (Table 3). For personality inventory runs, performance was significantly correlated with the features *strategicThinker*, *domainExpert*, and *perilSpecific*, but not with *riskTaker* and *doOrBe*. For the direct heuristic runs, the majority of features showed no significant correlation with performance. Additionally, the DH feature correlations are prone to dramatic changes across trials providing additional concern for the reliability of directly prompting for inventory responses. Conversely, the PI generations repeatedly provided highly significant correlations across multiple trials. LLaMA 3 was an exception to this trend, but as Mistral was able to generate highly correlated scores the difference in results is not accounted for by model size but by training or architecture choices.

These results support our suspicion that the direct heuristic method does not produce actual behaviors that differentiate strongly between personality prompts, whereas the personality inventory method does (at least with respect to the personality features we studied here). Furthermore, the correlations of personality inventory player performance with the LLM-annotated **perilSpecific** feature (row 3 of Table 3) is a promising sign that, given a shallow description of a player's personality, and assuming their personality translates predictably into their play style, LLMs may have some ability to predict player performance. Note that this does *not* necessarily demonstrate that the reason for the performance difference is that the top performers actually are exhibiting the personality traits we have identified as winning. At a minimum, this data shows that some qualitative difference exists which: (1) affects performance in the game of Peril, (2) was elicited by the persona prompting method augmented with personality inventories, and (3) was successfully identified by an LLM annotator which looked only at persona descriptions. Additionally, these early results show that the personality inven-

Rating	Persona Description
28.08	A geopolitical strategist who often appears on different networks presenting an alternative viewpoint on policies and events
27.89	A government agency using GIS analysis to plan efficient land use and infrastructure development
27.81	A competitive collegiate football player always seeking for custom-designed team merchandise
27.35	A retired intelligence officer who had previously worked for the CIA.
27.11	A person who struggles with Discardia – a fear of throwing things away.
23.03	A genealogist researching family histories connected to Biddeford, Maine.
22.04	A genealogist helping clients trace their family roots, particularly those with connections to the Somme department in France.
21.83	A healthcare blogger who spreads misinformation about vaccines and challenges the nurse's beliefs
21.70	A struggling high school student who has no interest in biology.
18.64	A young child who laughs uncontrollably at the street performer's antics

Table 2: TrueSkill for GPT4 - PI1 - Run 1. Higher TrueSkill ratings align with human expectations: tactically minded personas (e.g., strategists, officers, athletes) outperform less focused ones (e.g., underperforming students, children).

tory method of translating persona descriptions into heuristic choices leads to more observably distinct behaviors than the direct heuristic method. We suspect that this is due to the personality inventory's ability to enforce that interpretations of personality traits apply more evenly across all heuristics.

However, the TrueSkill algorithm has a bit of a locking-in effect, where after a large number of games, the amount that subsequent games change one's TrueSkill rating is decreasingly small. This is the reason that we had two separate runs for the personality inventory players, and likewise for the direct method players. Given this fact, was the final ranking of players consistent between these runs? The final player ranking of all players in the first and second runs had a close-to-significant Spearman correlation ($\rho = 0.278, p = 0.051$). However, interestingly (and counter-intuitively), the correlation between runs three and four was much stronger ($\rho = 0.524, p < 0.001$). Likewise, the final rankings of personality inventory player runs did not correlate significantly with final rankings of the direct heuristic runs (all were $p > 0.2$). Thus, although the personality inventory method produced behaviors whose performance spreads more closely aligned with personality features (Table 3), the direct heuristic method produced behaviors that performed more consistently relative to other players. Alternatively, this shows that although previous performance of direct heuristic players is more predictive of future performance than previous performance of personality inventory players, inferred personality features are more predictive of personality inventory player performance than it is of direct heuristic player performance.

4.2 Comparison of Personality Inventory and Direct Heuristic Players

One of the motivators for our introduction of the personality inventory method was the observation that the direct heuristic method led to inconsistent heuristic choices. It is expected that, given a single persona *P*, and a set of heuristics, the value assigned to each heuristic should have what we call *opposite-value consistency*: heuristics specifying opposite properties should have opposite values (Table 4). To measure opposite-value consistency, for each heuristic that has a direct opposite, we measure the ratio between them (using the larger value as the numerator), and average across all players $\mathbf{P}$. Because the range of possible heuristic values is 0 to 100, we cap this ratio at 100. For some opposite heuristic pair (h_1, h_2), if player p has heuristic values h_1^p, h_2^p, the opposite-value consistency is:

$$\left(\sum_{p \in \mathbf{P}} \max\left(\frac{h_1^p}{h_2^p}, \frac{h_2^p}{h_1^p}, 100\right) \right) / |\mathbf{P}|$$

The opposite-value consistency for personality inventory heuristics are strongly impacted by the models used to generate them, and the smaller models did not behave in a manner consistent with expectations based on the prompts. The ability for GPT-4 to correctly identify mutually exclusive heuristics may need to manually be accounted for when transferring the prompts to smaller models.

5 Conclusion

In this work, we explored the differences between two methods of translating persona descriptions into actual behaviors. The first, direct heuristics, is by far the most common method seen in current literature. However, we observed that this method leads to heuristic selection that was inconsistent for a given persona. As shown in Table 3, the DH method resulted in correlations that were both weak and non-significant. We therefore proposed a personality inventory technique, which involves the

Feature	Model	PI1	PI2	DH1	DH2
strategicThinker	GPT4	**0.49*****	**0.40*****	0.08	0.05
	Mistral R1	0.2447	**0.3826****	0.0046	0.0645
	Mistral R2	**0.4391****	**0.4360****	0.2218	0.2644
	LLaMA 3 R1	0.2192	0.2400	0.0211	0.1124
	LLaMA 3 R2	0.2076	**0.3938****	0.0370	-0.0572
	LLaMA 4 R1	**0.5707*****	0.2480	0.2218	**0.3531***
	LLaMA 4 R2	**0.4730*****	**0.5040*****	0.1999	0.0892
domainExpert	GPT4	**0.41*****	**0.44*****	0.12	0.03
	Mistral R1	**0.2848***	0.2688	-0.0269	0.1239
	Mistral R2	**0.3538***	**0.4900*****	0.1985	0.2696
	LLaMA 3 R1	0.1994	0.1583	0.1027	0.1561
	LLaMA 3 R2	0.2713	**0.3687****	-0.0275	-0.0548
	LLaMA 4 R1	**0.6166*****	**0.2796***	0.2438	**0.3817****
	LLaMA 4 R2	**0.4709*****	**0.5021*****	**0.3036***	0.0813
perilSpecific	GPT4	**0.41*****	**0.39*****	0.11	0.02
	Mistral R1	**0.2961***	**0.3460***	-0.0008	0.0743
	Mistral R2	**0.4572*****	**0.4859*****	0.2003	0.2448
	LLaMA 3 R1	0.2033	0.2355	0.0221	0.0979
	LLaMA 3 R2	**0.3161***	**0.4151****	0.0017	-0.0514
	LLaMA 4 R1	**0.6948*****	**0.3967****	0.2272	**0.3838****
	LLaMA 4 R2	**0.4826*****	**0.5814*****	0.1877	0.0314
riskTaker	GPT4	0.14	0.07	0.09	0.02
	Mistral R1	0.1778	0.1473	0.0019	-0.0192
	Mistral R2	**0.4516*****	**0.3366***	-0.0015	0.0879
	LLaMA 3 R1	0.0053	-0.0901	0.0142	0.2775
	LLaMA 3 R2	-0.0776	-0.1119	-0.0350	0.0275
	LLaMA 4 R1	0.1159	0.1205	-0.0401	0.2409
	LLaMA 4 R2	0.0562	0.1416	0.1959	0.0950
doOrBe	GPT4	-0.04	0.10	0.06	0.15
	Mistral R1	0.2119	0.0182	-0.0814	-0.0293
	Mistral R2	0.2516	0.2240	0.0337	0.0829
	LLaMA 3 R1	0.0422	0.1111	0.2060	0.0007
	LLaMA 3 R2	-0.1231	-0.0647	0.0082	-0.0422
	LLaMA 4 R1	-0.1930	-0.2326	0.0940	0.1138
	LLaMA 4 R2	-0.2227	-0.2128	-0.0643	0.2227

Table 3: The figures show the correlation between each personality feature and heuristic weights, as chosen by players using the direct and inventory heuristic methods. Each figure caption indicates the heuristic (DH/PI) used and its generation batch (1/2). All batches other than GPT4 were generated twice. Additionally, the statistical significance of each entry is indicated by asterisks "*" as follows: $* = (p \leq 0.05), ** = (p \leq 0.01), *** = (p \leq 0.005)$. The direct heuristic methods are consistently less statistically significant than the inventory heuristics for all models and runs. This is to be expected as direct generation of personality traits is known to have lower reliability than generation via inventories.

Phase	Heuristics	Mistral	LLaMA 3	LLaMA 4	GPT4
0	EACM–EACL	18.16	38.25	6.515	**-2.61**
	ETE–ETN	3.93	8.82	39.13	**-1.63**
	PCM–PCL	26.18	17.46	29.45	**-13.03**
	PTM–PTL	23.02	22.46	18.55	**-4.00**
	PUM–PUL	23.82	22.38	33.82	**-13.00**
1	ICD–ICS	8.75	25.30	**-2.05**	-5.20
	ONM–ONL	26.22	7.73	71.20	**-4.20**
	PCM–PCL	12.85	16.99	**10.40**	-47.87
	PTM–PTL	21.49	11.19	12.76	**-0.61**
	PUM–PUL	18.09	14.74	**14.13**	-14.62
2	CNM–CNL	-4.98	**-4.92**	38.49	-15.33
	M–L	**-3.49**	11.61	16.93	-11.28
	OBCM–OBCL	14.21	33.15	21.18	**1.36**
	OBTM–OBTL	**4.33**	26.49	20.74	-37.74
	OBUM–OBUL	**2.60**	39.03	19.21	-18.84

Table 4: Average difference in scores between opposite heuristics for models across phases. Positive values indicate higher DH scores and negative values indicate higher PI scores. Bold values indicate more similar opposite value consistency across DH and PI methods. Non-GPT models produced more conflicting from PI than DH, shown in larger average DH values.

creation of an inventory questionnaire that translates personas into heuristic values. We showed that the inventory method leads to heuristic values that are more consistent also shown in Table 3. The overwhelming majority of significant high correlation heuristics were generated by the PI method. It is also important to note that while smaller models (Mistral and LLaMA 3) did not provide correlations as strong, they still clearly behave similarly to the larger models in regards to heuristic correlation. Additional results can be found in the Appendix.

This work contributes to research showing that although persona prompting alone may lead to LLM outputs with different styles of text, it does not necessarily lead to substantially different decision-making behaviors. This further highlights the difficulty in translating personality descriptions into anything beyond surface-level expressions, since predicting behavioral differences between personalities requires a much deeper knowledge of the causality underlying human behavior than simply mimicking speech patterns. To address this, we showed that a personality inventory questionnaire-based approach can be more effective at eliciting behavioral heuristics that seem to better align with expectations of various personality descriptions, when compared to an approach that directly infers heuristics from personality descriptions. *In short: a main takeaway from this work is that simply asking an LLM to act in accordance with a persona, without using a moderator like the personality inventory method, may not suffice to produce realistic and diverse behaviors and decision-making.*

6 Future Work.

As part of this work, we did implement a PERIL *mission mode*, in which all players are assigned missions to achieve victory other than world domination. Although we did not utilize this functionality in the present paper (as it introduces another confounding variable), we will release the source code both of PERIL and the code used to replicate this paper's results, in order to encourage other researchers to explore the interesting problems in this space. For example, recent work shows that multi-agent systems may perform better at simulations than single-agent systems (Sreedhar and Chilton, 2024; Bui et al., 2025), but it is unclear how current state-of-the-art agents perform when playing adversarially against other potentially deceptive agents.

7 Limitations.

It should be noted that any observations we make here about how well a persona-prompted LLM matches its given persona can only at most have *face validity*—i.e., they only appear intuitively to match personality archetypes. We cannot fully establish whether these patterns have deeper alignment with actual human personalities without a proper psychometrically-designed empirical study on a population of people. Instead, the value in our present work is in the introduction of the personality inventory method, and the finding that without it, the variance in behaviors between persona prompts and the adherence to expected patterns such as opposite-value consistency are very small. Indeed, if it can be shown that the translation of personality features into heuristic behaviors has more than face validity, it can lead to powerful simulation technologies, as well as tools for studying the effect of personality on decision-making. For example, it might allow us to predict that a human being who matches a given persona would behave a certain way in a new scenario.

In this exploratory work, a single strategic environment (the game of PERIL). Within this environment, we used a limited number of heuristics, which are not fully representative of the range of possible decision-making heuristics in this game. Furthermore, our heuristic-guided agents made their heuristic choices in the beginning of the game, which reduced their adaptivity, since they could not adjust those heuristics in response to game conditions. Finally, we used TrueSkill as a way of estimating player performance, but it should be noted that not enough games were played to allow every player to face every other one, and due to the way TrueSkill is calculated, ordering affects final scores. This effect seemed to affect individual players more in the personality inventory runs than in the direct heuristic runs (as measured by correlation of final rankings of players), but it did not significantly change the effect of personality features on final player ordering (Table 3). The low correlation between final player rankings in the two PI runs is counter-intuitive to us, and future work will need to explore why this was the case.

Ethical Statement

We do not anticipate significant ethical issues, as this work did not involve human subjects or the collection of personal data. This research explores artificial intelligence strategies in the strategy-based board game PERIL and draws inspiration from the mechanics of Risk®, a trademarked board game owned by Hasbro, Inc. The game developed and described in this paper is an independent project and is not affiliated with, endorsed by, or associated with Hasbro, Inc. The use of Risk® as a reference point is solely for comparative analysis and academic purposes under the principles of fair use. No proprietary elements, such as trademarked names, graphics, or copyrighted texts, have been reproduced in this work. In the interest of research transparency and replicability, upon acceptance for publication we will release the full source code, along with all prompts used in this work, on a publicly accessible GitHub repository.

Acknowledgments

This research was supported in part by the Air Force Research Laboratory, Information Directorate, through the Air Force Office of Scientific Research Summer Faculty Fellowship Program®, Contract Numbers FA8750-15-3-6003, FA9550-15-0001 and FA9550-20-F-0005.

References

Nikolas Belle, Dakota Barnes, Alfonso Amayuelas, Ivan Bercovich, Xin Eric Wang, and William Wang. 2025. Agents of change: Self-evolving llm agents for strategic planning. *arXiv preprint arXiv:2506.04651*.

Pranav Bhandari, Nicolas Fay, Michael Wise, Amitava Datta, Stephanie Meek, Usman Naseem, and Mehwish Nasim. 2025a. Can llm agents maintain a persona in discourse? *arXiv preprint arXiv:2502.11843*.

Pranav Bhandari, Usman Naseem, Amitava Datta, Nicolas Fay, and Mehwish Nasim. 2025b. Evaluating personality traits in large language models: Insights from psychological questionnaires. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 868–872.

Bojana Bodroža, Bojana M. Dinić, and Ljubiša Bojić. 2024. Personality testing of large language models: limited temporal stability, but highlighted prosociality. *Royal Society Open Science*, 11(10).

Ngoc Bui, Hieu Trung Nguyen, Shantanu Kumar, Julian Theodore, Weikang Qiu, Viet Anh Nguyen, and Rex Ying. 2025. Mixture-of-personas language models for population simulation. *arXiv preprint arXiv:2504.05019*.

Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data cre-

ation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*.

Alicia DeVrio, Myra Cheng, Lisa Egede, Alexandra Olteanu, and Su Lin Blodgett. 2025. A taxonomy of linguistic expressions that contribute to anthropomorphism of language technologies. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–18.

Ivar Frisch and Mario Giulianelli. 2024. Llm agents in interaction: Measuring personality consistency and linguistic alignment in interacting populations of large language models. *arXiv preprint arXiv:2402.02896*.

Vinicius G. Goecks and Nicholas Waytowich. 2024. Coa-gpt: Generative pre-trained transformers for accelerated course of action development in military operations. In *Proceedings of the 2024 IEEE Conference on Artificial Intelligence (IEEE AI 2024)*, pages 1–8.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Ralf Herbrich, Tom Minka, and Thore Graepel. 2006. Trueskill™: a bayesian skill rating system. In *Proceedings of the 19th International Conference on Neural Information Processing Systems*, NIPS'06, page 569–576, Cambridge, MA, USA. MIT Press.

Sihao Hu, Tiansheng Huang, Fatih Ilhan, Selim Tekin, Gaowen Liu, Ramana Kompella, and Ling Liu. 2024. A survey on large language model-based game agents. *Preprint*, arXiv:2404.02039.

Tiancheng Hu and Nigel Collier. 2024. Quantifying the persona effect in llm simulations. *arXiv preprint arXiv:2402.10811*.

Plotly Technologies Inc. 2018. Dash cytoscape: Interactive network visualization in python.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. PersonaLLM: Investigating the ability of large language models to express personality traits. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3605–3627, Mexico City, Mexico. Association for Computational Linguistics.

Kenan Jiang, Li Xiong, and Fei Liu. 2025. Harbor: exploring persona dynamics in multi-agent competition. *arXiv preprint arXiv:2502.12149*.

Mahammed Kamruzzaman and Gene Louis Kim. 2024. Exploring changes in nation perception with nationality-assigned personas in llms. *arXiv preprint arXiv:2406.13993*.

Junseok Kim, Nakyeong Yang, and Kyomin Jung. 2024. Persona is a double-edged sword: Mitigating the negative impact of role-playing prompts in zero-shot reasoning tasks. *arXiv preprint arXiv:2408.08631*.

R F Krueger, J Derringer, K E Markon, D Watson, and A E Skodol. 2012. Initial construction of a maladaptive personality trait model and inventory for dsm-5. *Psychol Med*, 42(9):1879–1890.

Ang Li, Haozhe Chen, Hongseok Namkoong, and Tianyi Peng. 2025. Llm generated persona is a promise with a catch. *arXiv preprint arXiv:2503.16527*.

Shuhang Lin, Wenyue Hua, Lingyao Li, Che-Jui Chang, Lizhou Fan, Jianchao Ji, Hang Hua, Mingyu Jin, Jiebo Luo, and Yongfeng Zhang. 2024. Battleagent: Multi-modal dynamic emulation on historical battles to complement historical analysis. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 172–181.

R. Lingard and Frank C. Erb. 1907. *A Letter of Advice to a Young Gentleman Leaving the University Concerning His Behaviour and Conversation in the World*. McAufliffe & Booth, New York. Retrieved from the Library of Congress, www.loc.gov/item/07007481/.

Yiren Liu, Pranav Sharma, Mehul Jitendra Oswal, Haijun Xia, and Yun Huang. 2024. Personaflow: Boosting research ideation with llm-simulated expert personas. *arXiv preprint arXiv:2409.12538*.

Nunzio Lorè and Babak Heydari. 2024. Strategic behavior of large language models and the role of game structure versus contextual framing. *Scientific Reports*, 14(1):18490.

Julina Maharjan, Ruoming Jin, Jianfeng Zhu, and Deric Kenne. 2025. Do large language models really understand personality? *Journal of Medical Internet Research*, 21(05).

Meta Fundamental AI Research Diplomacy Team (FAIR), Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, and others. 2022. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074.

Carlos Olea, Holly Tucker, Jessica Phelan, Cameron Pattison, Shen Zhang, Maxwell Lieb, Doug Schmidt, and Jules White. 2024. Evaluating persona prompting for question answering tasks. In *Proceedings of the 10th International Conference on Artificial Intelligence and Soft Computing*.

Xuchen Pan, Dawei Gao, Yuexiang Xie, Yushuo Chen, Zhewei Wei, Yaliang Li, Bolin Ding, Ji-Rong Wen, and Jingren Zhou. 2024. Very large-scale multi-agent simulation in agentscope. *arXiv preprint arXiv:2407.17789*.

Steve Phelps and Yvan I Russell. 2025. The machine psychology of cooperation: can gpt models operationalize prompts for altruism, cooperation, competitiveness, and selfishness in economic games? *Journal of Physics: Complexity*, 6(1):015018.

Daniele Potertì, Andrea Seveso, and Fabio Mercorio. 2025. Designing role vectors to improve llm inference behaviour. *arXiv preprint arXiv:2502.12055*.

Manuel Pratelli and Marinella Petrocchi. 2025. Evaluating the simulation of human personality-driven susceptibility to misinformation with llms. *arXiv preprint arXiv:2506.23610*.

Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik Narasimhan, and Vishvak Murahari. 2024. Personagym: Evaluating persona agents and llms. *arXiv preprint arXiv:2407.18416*.

Andreas Schuller, Doris Janssen, Julian Blumenröther, Theresa Maria Probst, Michael Schmidt, and Chandan Kumar. 2024. Generating personas using llms and assessing their viability. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24)*, pages 179:1–179:7.

Karthik Sreedhar and Lydia Chilton. 2024. Simulating human strategic behavior: Comparing single and multi-agent llms. *arXiv preprint arXiv:2402.08189*.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two tales of persona in llms: A survey of role-playing and personalization. *arXiv preprint arXiv:2406.01171*.

Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2024. Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 257–279.

Ziyi Wang, Yuxuan Lu, Wenbo Li, Amirali Amini, Bo Sun, Yakov Bart, Weimin Lyu, Jiri Gesi, Tian Wang, Jing Huang, et al. 2025. Opera: A dataset of observation, persona, rationale, and action for evaluating llms on human online shopping behavior simulation. *arXiv preprint arXiv:2506.05606*.

Shenghan Wu, Yang Deng, Yimo Zhu, Wynne Hsu, and Mong Li Lee. 2025. From personas to talks: Revisiting the impact of personas on llm-synthesized emotional support conversations. *arXiv preprint arXiv:2502.11451*.

Yunze Xiao, Lynnette Hui Xian Ng, Jiarui Liu, and Mona T Diab. 2025. Humanizing machines: Rethinking llm anthropomorphism through a multi-level framework of design. *arXiv preprint arXiv:2508.17573*.

Yadong Zhang, Shaoguang Mao, Tao Ge, Xun Wang, Adrian de Wynter, Yan Xia, Wenshan Wu, Ting Song, Man Lan, and Furu Wei. 2024. Llm as a mastermind: A survey of strategic reasoning with large language models. *arXiv preprint arXiv:2404.01230*.

Mingqian Zheng, Jiaxin Pei, Lajanugen Logeswaran, Moontae Lee, and David Jurgens. 2024. When "a helpful assistant" is not really helpful: Personas in system prompts do not improve performances of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15126–15154, Miami, Florida, USA. Association for Computational Linguistics.

A Appendix

This section contains the PERIL gameplay cycle, prompts, heuristic correlation tables, top 5 and bottom 5 TrueSkill persona rankings, and opposite value consistency tables. This material is not required to understand the main body of work, but readers may find it useful to gain a broader understanding of model impact on the results.

A.1 Gameplay Cycle

- **Initialize Game:** Up to six players are placed on a map of regions divided into zones. The mission is world domination by occupying all regions. Each region must eventually contain at least one unit.

- **Initialization Phase:** Each player is given a pool of units. Players alternate placing one unit on an unoccupied region until all regions are taken. Remaining units are then placed on the regions they control.

- **Gameplay Loop (per turn):**

  - *Reinforcement:* Player receives new units and places them on controlled regions.

  - *Attack:* Player may attack adjacent enemy regions using armies from their own regions.

  - *Redeployment:* Player may move units between connected regions they control.

- **End Condition:** If a player occupies all regions, they are declared the winner.

A.2 Graphic

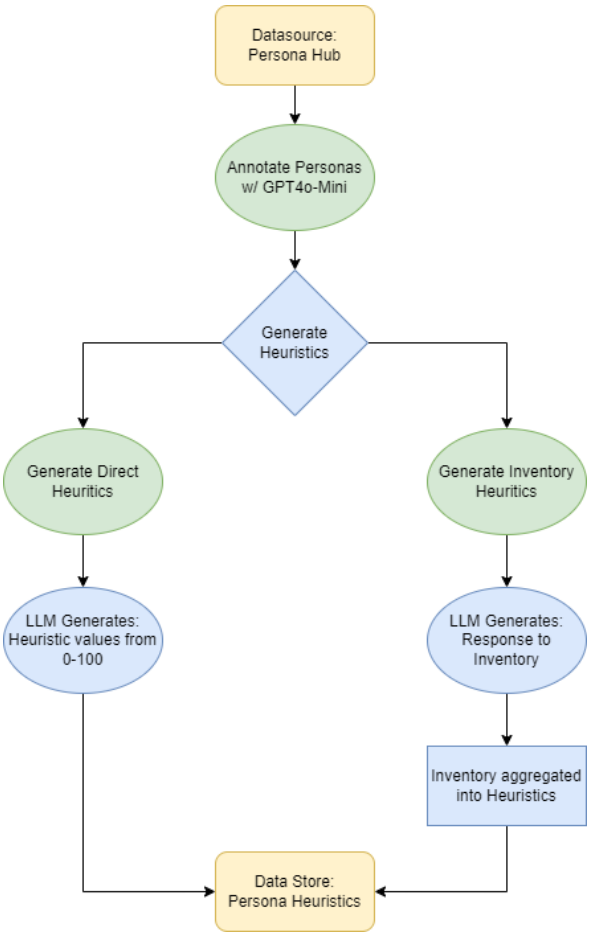

Figure 2: Pipeline from persona descriptions to heuristic agents. Note that LLMs generate heuristic proxies (DH or PI), but do not play PERIL directly.

A.3 Prompts

We used three prompt sets to generate the heuristics used in this work. The first set was used to generate the initial assessments used for identifying fifty personas that are maximally distinct from each other. Usage of the assessment results is discussed in detail in section 3.2. The second set was to generate the direct heuristics used in playing the PERIL game. The third set was also used in gameplay, but it was used to generate the inventory heuristics. When generating any assessment or heuristic, the persona in question is always included in the prompt.

A.3.1 Assessment Prompt

You will be given a personality of an individual. You must estimate how well that person would do on the strategic board game Peril. In the game of Peril, you are a player who must achieve world

conquest, very similar to the popular board game Risk. For the personality you are given, return a JSON object with the following values: - 'index': An integer describing the index of the prompt in the dataset. We will provide this for you. - 'personality': The personality of the individual, which we also provided to you. - 'strategicThinker': A rating of whether this personality describes a strategic thinker. This should be one of the numbers in -1, -0.5, 0, 0.5, 1. Use the following scale: - +1 : This is certain to be a very capable strategic thinker, who combines systematic thinking with appropriate use of intuition and can consistently perform at a high level. - +0.5 : This is more likely than not to be someone with good strategic thinking capabilities, but they may not be consistent or fully developed. - 0 : There is no evidence to suggest that this person is or is not a capable strategic thinker. - -0.5 : This is more likely than not to be someone with poor or no strategic thinking capabilities. - -1: This is certain to be someone with poor or no strategic thinking capabilities, who is unable to perform at an acceptable level at any task requiring strategic thinking. - 'domainExpert': A rating of whether this personality describes someone who has experience in combat, military warfare, or other similar areas of expertise. This should be one of the numbers in -1, -0.5, 0, 0.5, 1. Use the following scale: - +1 : This is certain to be someone who has high level experience in combat, military warfare, or other similar areas of expertise. - +0.5 : This is more likely than not to be someone who has experience in combat, military warfare, or other similar areas of expertise. - 0 : There is no evidence to suggest that this person does or does not have experience in combat, military warfare, or other similar areas of expertise. - -0.5 : This is more likely than not to be someone who has poor or no experience in combat, military warfare, or other similar areas of expertise. - -1: This is certain to be someone with poor or no experience in combat, military warfare, or other similar areas of expertise. - 'perilSpecific': A rating of whether this personality describes someone who is likely to perform well on the game Peril specifically. This should be one of the numbers in -1, -0.5, 0, 0.5, 1. Use the following scale: - +1 : This is certain to be someone who is likely to perform well on the game Risk. - +0.5 : This is more likely than not to be someone who is likely to perform well on the game Risk. - 0 : There is no evidence to suggest that this person is or is not likely to perform well on the game Risk. - -0.5 : This is more likely to be someone who will perform poorly on the game Risk. - -1: This is certain to be someone who will perform poorly on the game Risk. - 'riskTaker': A rating of whether this personality describes someone who is likely to be a high risk taker. This should be one of the numbers in -1, -0.5, 0, 0.5, 1. Use the following scale: - +1 : This is certain to be someone who is likely to take dangerous risks, always making moves that are likely to have a high payoff but also a high risk of failure. - +0.5 : This is more likely than not to be someone who is likely to take dangerous risks, often making moves that are likely to have a high payoff but also a high risk of failure. - 0 : There is no evidence to suggest that this person is or is not likely to take dangerous risks. - -0.5 : This is more likely to be someone who will not take dangerous risks, often making moves that are likely to have a low payoff but also a low risk of failure. - -1: This is certain to be someone who will not take dangerous risks, always making moves that are likely to have a low payoff but also a low risk of failure. - 'doOrBe': A rating of whether this personality description is an instruction to act a certain way by describing actions (do) or if it is an instruction to play a role by describing a personality or character background (be). This should be one of the numbers in -1, 0, 1. Use the following scale: - +1 : This personality description describes how to act a certain way primarily by describing specific actions. - 0 : This personality description has a balance between describing specific actions and describing a personality or character background. - -1: This personality description describes how to play a role primarily by describing a personality or character background.

A.3.2 Peril Introduction Prompt

You are playing the board game "Peril" with other players, which is inspired by the popular board game "Risk". In this game, you are battling to conquer the world. The first player to achieve this wins the game.

BOARD The board consists of seven zones (North America, South America, Europe, Asia, Africa, and Australia), which are divided into regions. The regions are connected to each other and either connect over land, or over water. Players control units, which can be thought of as armies. Every region must have at least one unit on it (except in the beginning of the game, when no regions have any). Regions can contain an unlimited number of

units, but all units on a single region must belong to the same player.

GAMEPLAY - First, all players are given a number of units. They alternate, placing one unit on an unoccupied region at a time. When all regions are occupied, players can place units on regions they already occupy. At no time are players allowed to place units in regions occupied by other players. They continue in this manner, placing one unit at a time, until all players have placed their units. This ends the initial placement phase. - Each player now takes their regular turns. Each turn is broken up into three subphases: reinforcement, attack, and deployment. - In the reinforcement phase, the player is given a number of units depending on which regions and zones they own. If they own all regions in a zone at the beginning of their turn, they get a bonus number of units depending on the size of the zone. They are allowed to place units on regions they already occupy. - In the attack phase, players may attack from a region they own into any adjacent enemy-occupied region. They must declare how many units they wish to send in the attack, and this number must be less than or equal to the number of units on the attacking region. A number of attackers and defenders will be killed in this attack, and the higher the number of attackers, the greater the chance of success. If all defenders are killed, then the number of units remaining in the attack must all move to the defending region, leaving at least one unit behind in the attacking region. The player may then perform additional attacks until they are done or cannot attack any more. - In the redeployment phase, the player may relocate any of their units to any other region they control, so long as: (1) they leave at least one unit on every region they own, and (2) if a unit is moved from one region T1 to another region T2, T1 and T2 must be either directly connected, or connected via a sequence of connected regions that are all owned by the player.

A.3.3 Direct Heuristic Prompt

We need to determine what values to assign these heuristics, based on your assigned personality. For each of the heuristics above, we need to determine a value between 0 and 100 that represents how much this heuristic should be weighted in the decision making process. This value should be based on your assigned personality. For example, a player who is very aggressive might have a high value for the heuristic that says to attack from region T1 to region T2 if T2 is owned by the player with the fewest units. By default, every heuristic has a value of 5. So if you want to deprioritize a heuristic, you can assign it a value less than 5. If you want to prioritize a heuristic, you can assign it a value greater than 5. If you want to ignore a heuristic, you can assign it a value of 0. If you want to use a heuristic as much as possible, you can assign it a value of 100.

A.3.4 Deployment Heuristics (Phase 0) Prompt

PTM - Place a unit in a region T1 that is adjacent to region T2 if T2 is owned by the player with the most regions. PTL - Place a unit in a region T1 that is adjacent to region T2 if T2 is owned by the player with the fewest regions. PUM - Place a unit in a region T1 that is adjacent to region T2 if T2 is owned by the player with the most units. PUL - Place a unit in a region T1 that is adjacent to region T2 if T2 is owned by the player with the fewest units. PCM - Place a unit in a region T1 that is adjacent to region T2 if T2 is owned by the player with the most zones owned (measured by total zone bonuses). PCL - Place a unit in a region T1 that is adjacent to region T2 if T2 is owned by the player with the fewest zones owned (measured by total zone bonuses). ETE - Place a unit in region T if T is adjacent to an enemy region. ETN - Place a unit in region T if T is not adjacent to any enemy regions. EAC - Place a unit in region T if T is on a zone boundary. EACM - Place a unit in region T if T is adjacent to the largest zone (by number of regions). EACL - Place a unit in region T if T is adjacent to the smallest zone (by number of regions). EACO - Place a unit in region T if T is adjacent to a zone that is completely owned by another player.

A.3.5 Attack Heuristics (Phase 1) Prompt

PTM - Attack from region T1 to region T2 if T2 is owned by the player with the most regions. PTL - Attack from region T1 to region T2 if T2 is owned by the player with the fewest regions. PUM - Attack from region T1 to region T2 if T2 is owned by the player with the most units. PUL - Attack from region T1 to region T2 if T2 is owned by the player with the fewest units. PCM - Attack from region T1 to region T2 if T2 is owned by the player with the most zones owned (measured by total zone bonuses). PCL - Attack from region T1 to region T2 if T2 is owned by the player with the fewest zones owned (measured by total zone bonuses).

ONM - Attack if the units on T1 are greater than the units on T2. ONL - Attack if the units on T1 are fewer than the units on T2. ON2 - Attack if the units on T1 are at least 2x the number of units on T2. ICD - Attack if T1 and T2 are in different zones. ICS - Attack if T1 and T2 are in the same zone. ICOE - Attack if T2 is in a zone completely owned by a single player. L - Attack if T2 connects to a region you own that T1 isn't currently linked to. PASS - Likelihood of passing (ending your turn and not doing any more attacks). If set to 100, you will never attack; if 0, you will always attack.

A.3.6 Redeployment Heuristics (Phase 2) Prompt

OBTM - Move from region T1 to T2 if T2 is adjacent to more regions occupied by the player with the most regions. OBTL - Move from region T1 to T2 if T2 is adjacent to more regions occupied by the player with the fewest regions. OBUM - Move from region T1 to T2 if T2 is adjacent to more regions occupied by the player with the most units. OBUL - Move from region T1 to T2 if T2 is adjacent to more regions occupied by the player with the fewest units. OBCM - Move from region T1 to T2 if T2 is adjacent to more regions occupied by the player with the most zones owned (measured by total zone bonuses). OBCL - Move from region T1 to T2 if T2 is adjacent to more regions occupied by the player with the fewest zones owned (measured by total zone bonuses). CNM - Move from region T1 to T2 if T2 is connected to more regions than T1 is. CNL - Move from region T1 to T2 if T2 is connected to fewer regions than T1 is. CB - Move from region T1 to T2 if T2 is on a zone boundary and T1 is not. CA - Move from region T1 to T2 if T2 is adjacent to at least one enemy-owned region and T1 is not. CAC - Move from region T1 to T2 if T2 is adjacent to a region in a zone completely owned by a single enemy player and T1 is not. M - Move from region T1 to T2 if T2 has more units than T1. L - Move from region T1 to T2 if T2 has fewer units than T1. SI - Move from region T1 to T2 if T2 is adjacent to a region with a higher chance of successful invasion than any of those connected to T1, calculated using the ratio of available troops from attacking region to troops on target region. PASS - Likelihood of passing (ending your turn and not doing any more redeployments). If set to 100, you will never redeploy; if 0, you will always redeploy.

A.3.7 Redeployment Heuristics (Phase 2) Prompt

We are interested in how you would describe yourself. Given the statement "item_question", you must choose a number from 0 to 3: 0 - Very false or often false 1 - Sometimes or somewhat false 2 - Sometimes or somewhat true 3 - Very true or often true

A.4 Correlation Between Heuristics and Personality

This section expands upon the relationship between the annotated assessments in section 3.2 and the generated heuristics (both direct and inventory) as described in section 3.3. The figures show the correlation between each personality feature and heuristic weights, as chosen by players using the direct and inventory heuristic methods. Each figure caption indicates the heuristic (DH/PI) used and its generation batch (1/2). All batches other than GPT4 were generated twice. The figures are ordered by phase from top to bottom: Phase 1 heuristics (initialization / deployment), Phase 2 (attack), Phase 3 (redeployment). Additionally, the statistical significance of each entry is indicated by asterisks "*" as follows: $* = (p \leq 0.05), ** = (p \leq 0.01), *** = (p \leq 0.005)$. Heuristic appearing as 'nan' are due to not having been selected by the LLM. Relevant Figures: 3 - 16.

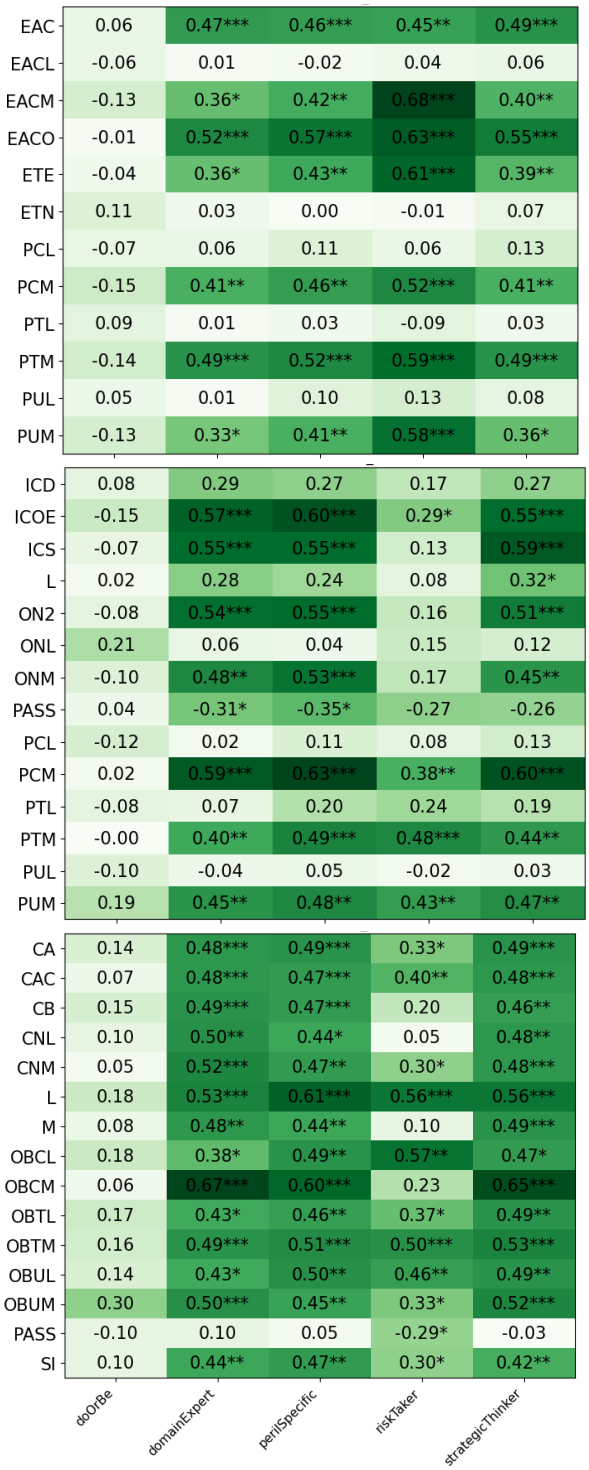

Figure 3: Heuristic Correlations - DH1 - GPT4

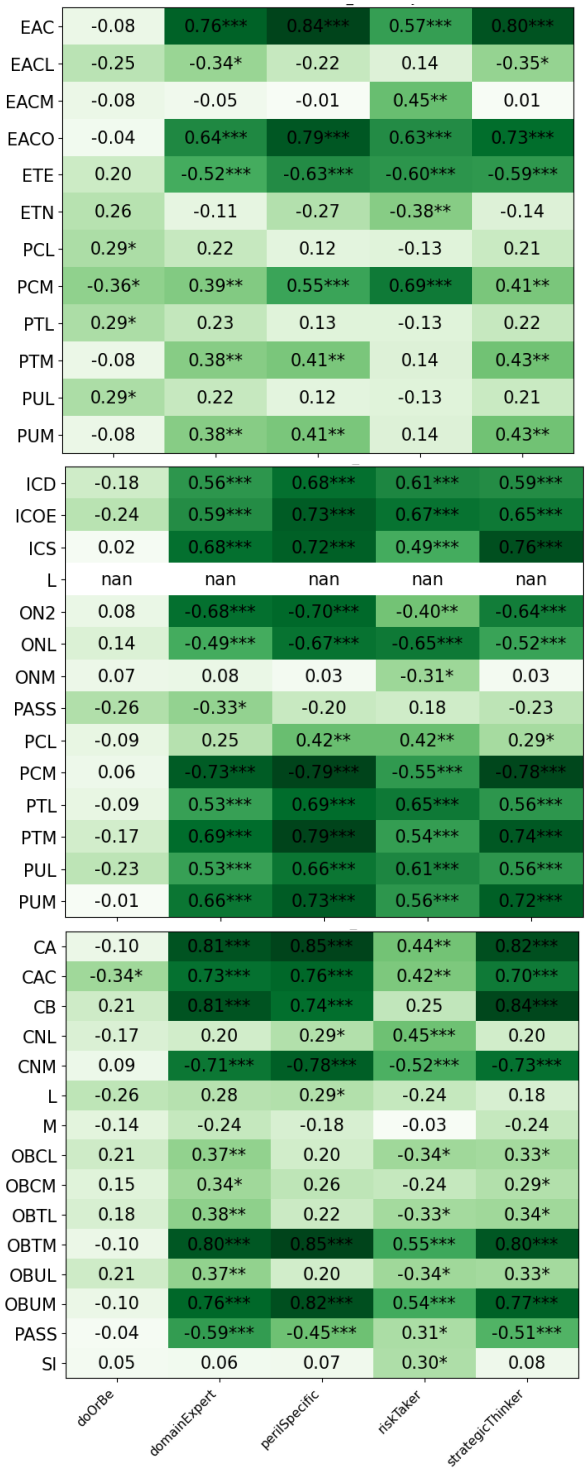

Figure 4: Heuristic Correlations - PI1 - GPT4

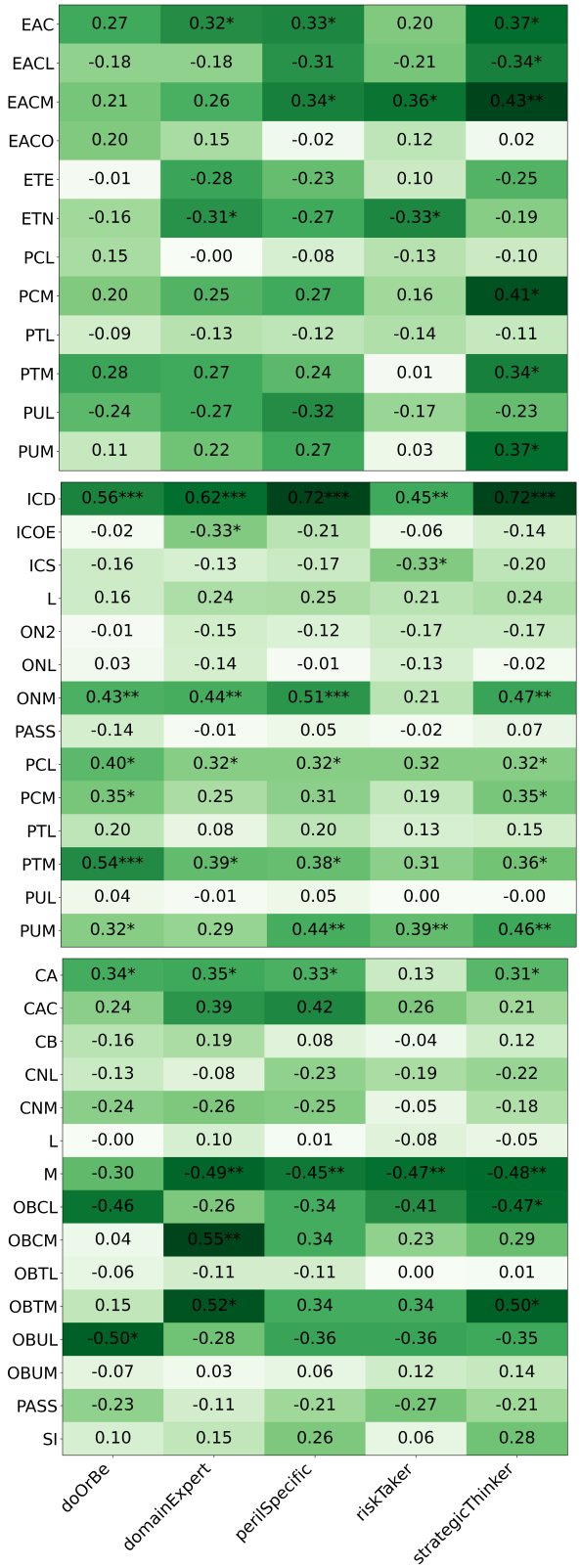

Figure 5: Heuristic Correlations - DH1 - Mistral Small

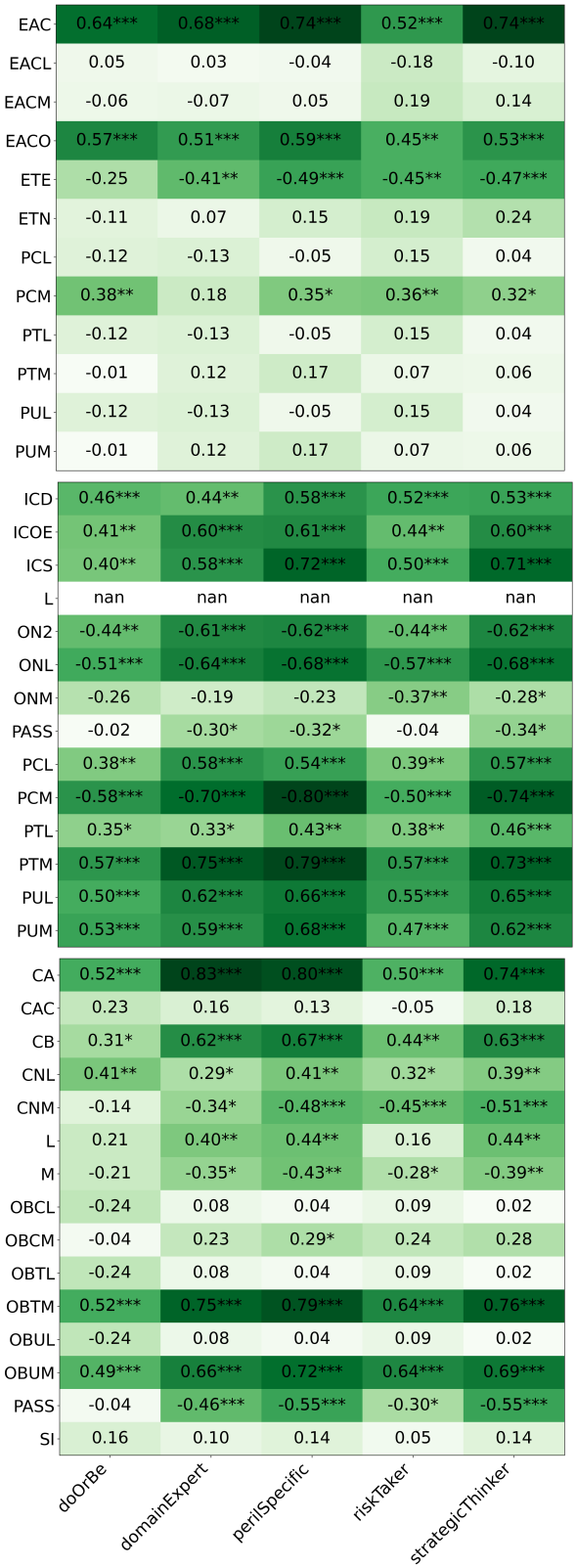

Figure 6: Heuristic Correlations - PI1 - Mistral Small

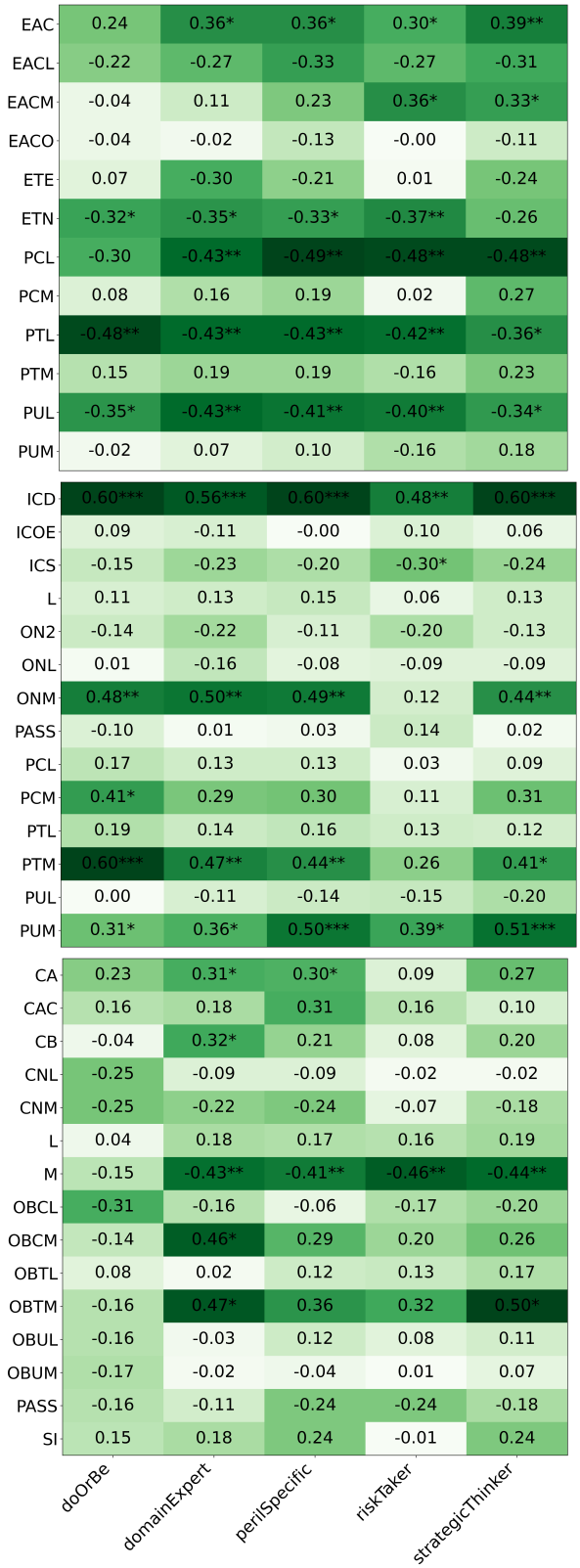

Figure 7: Heuristic Correlations - DH2 - Mistral Small

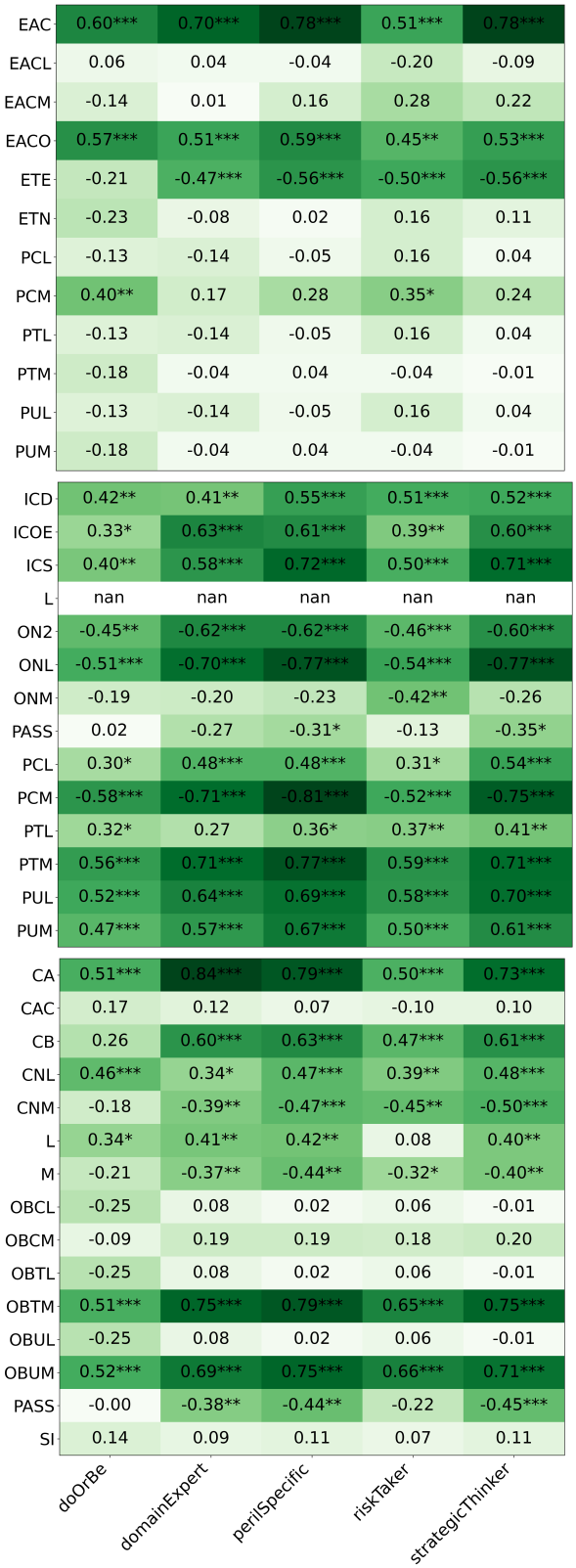

Figure 8: Heuristic Correlations - PI2 - Mistral Small

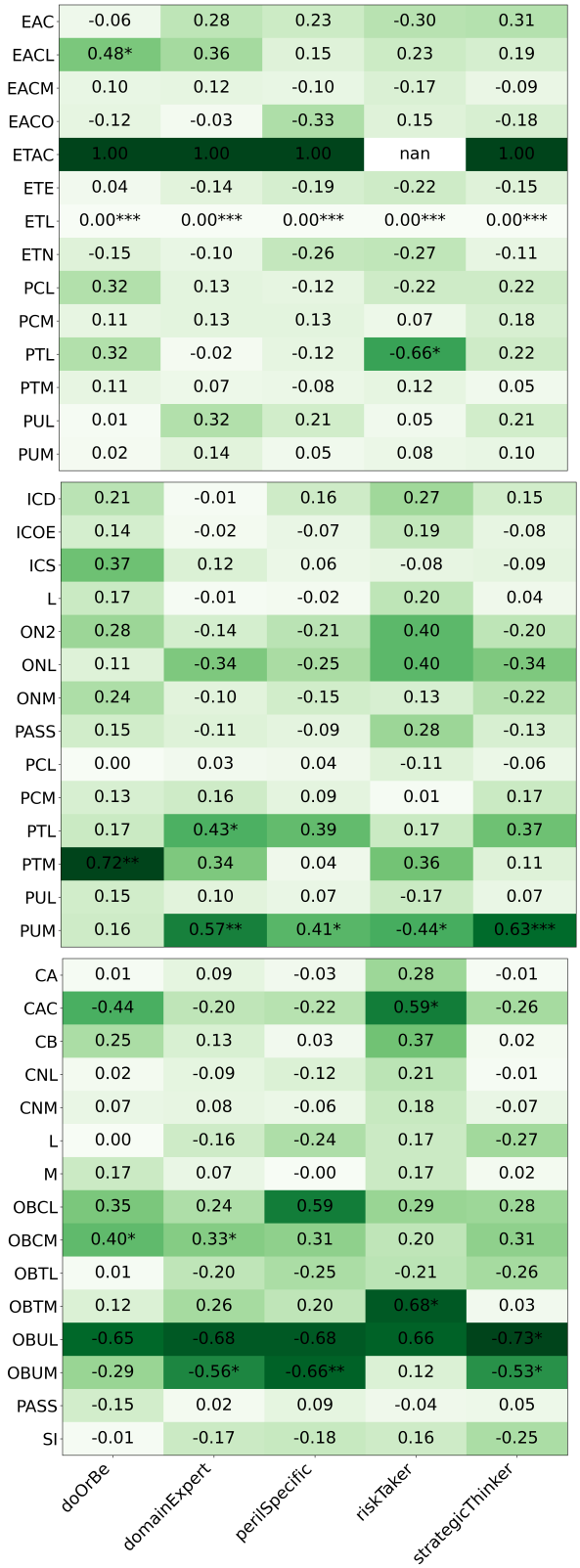

Figure 9: Heuristic Correlations - DH1 - LLaMA 3

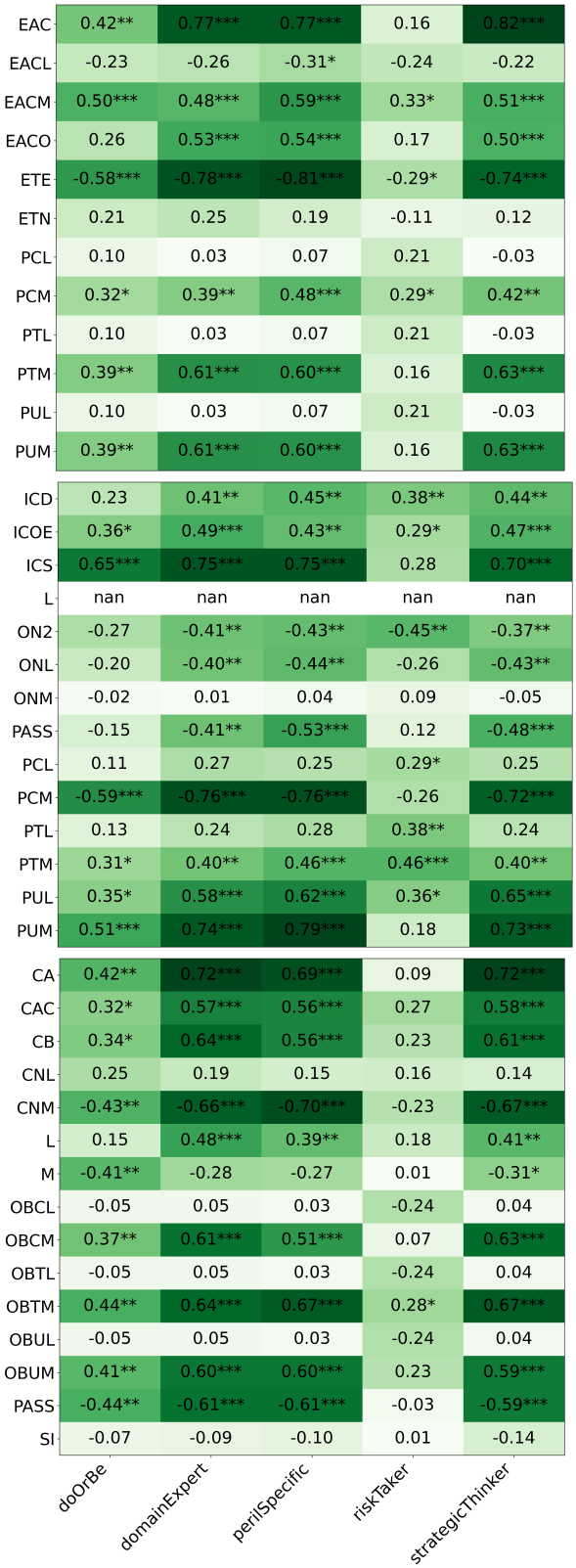

Figure 10: Heuristic Correlations - PI1 - LLaMA 3

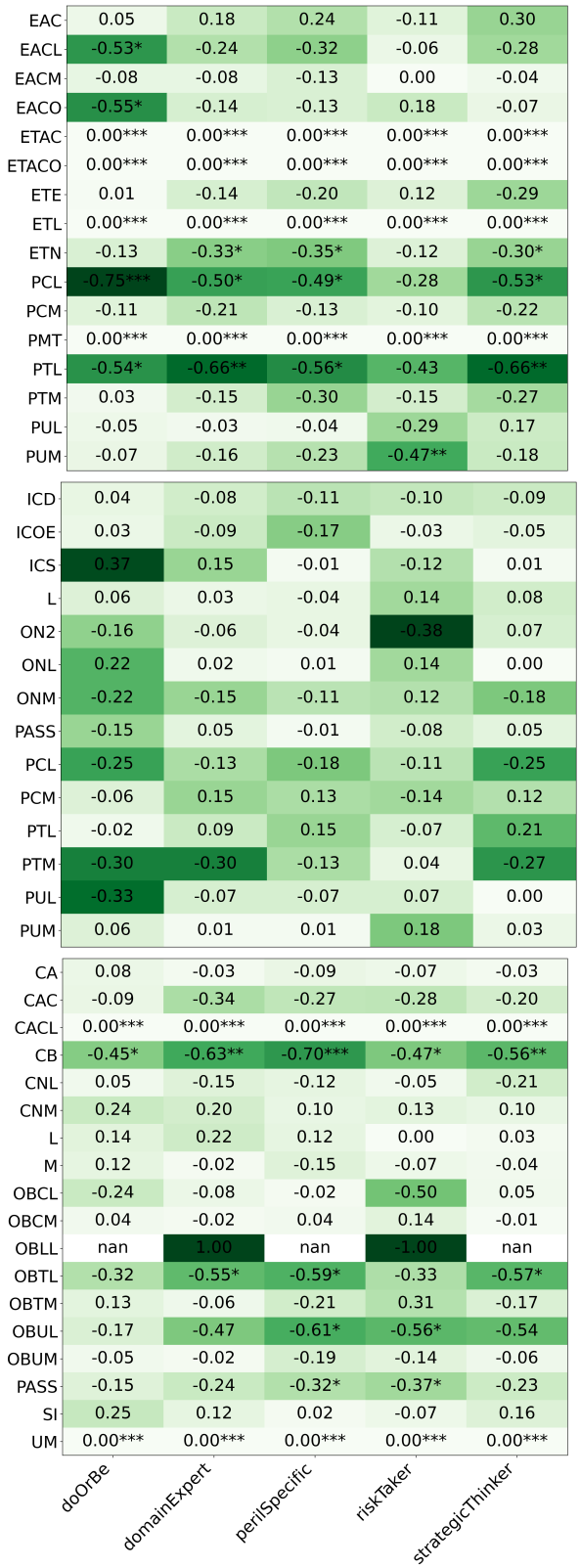

Figure 11: Heuristic Correlations - DH2 - LLaMA 3

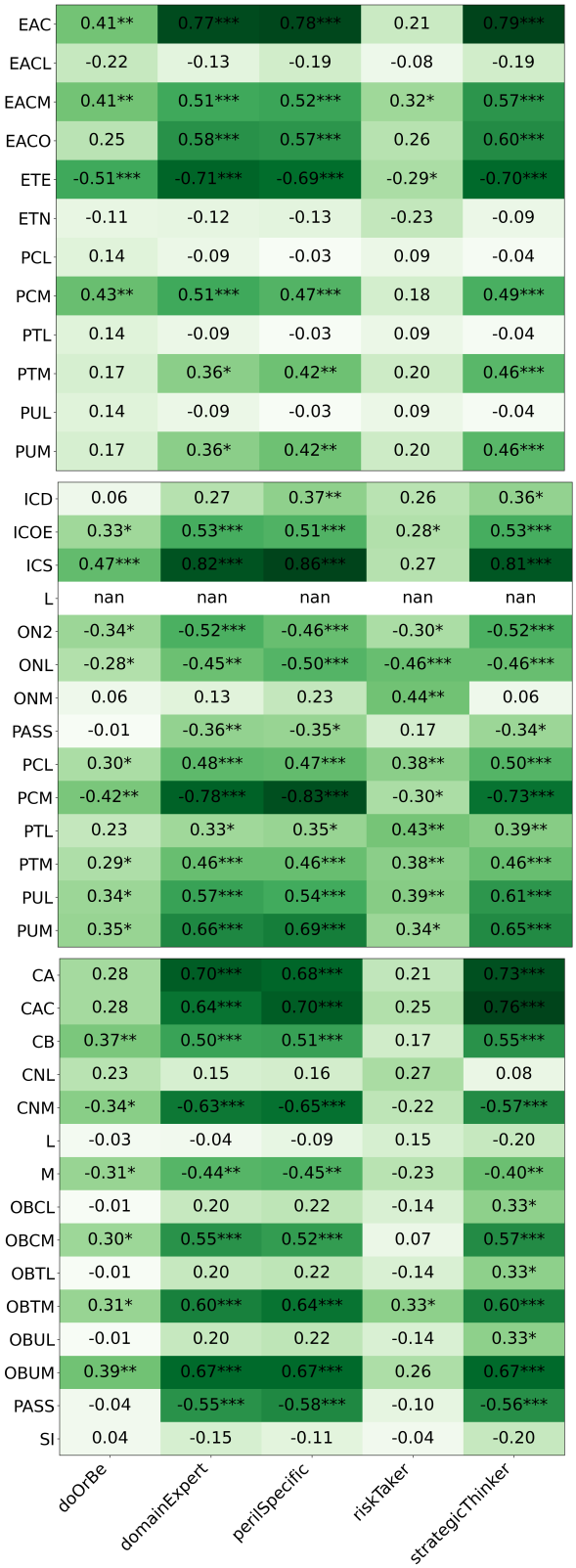

Figure 12: Heuristic Correlations - PI2 - LLaMA 3

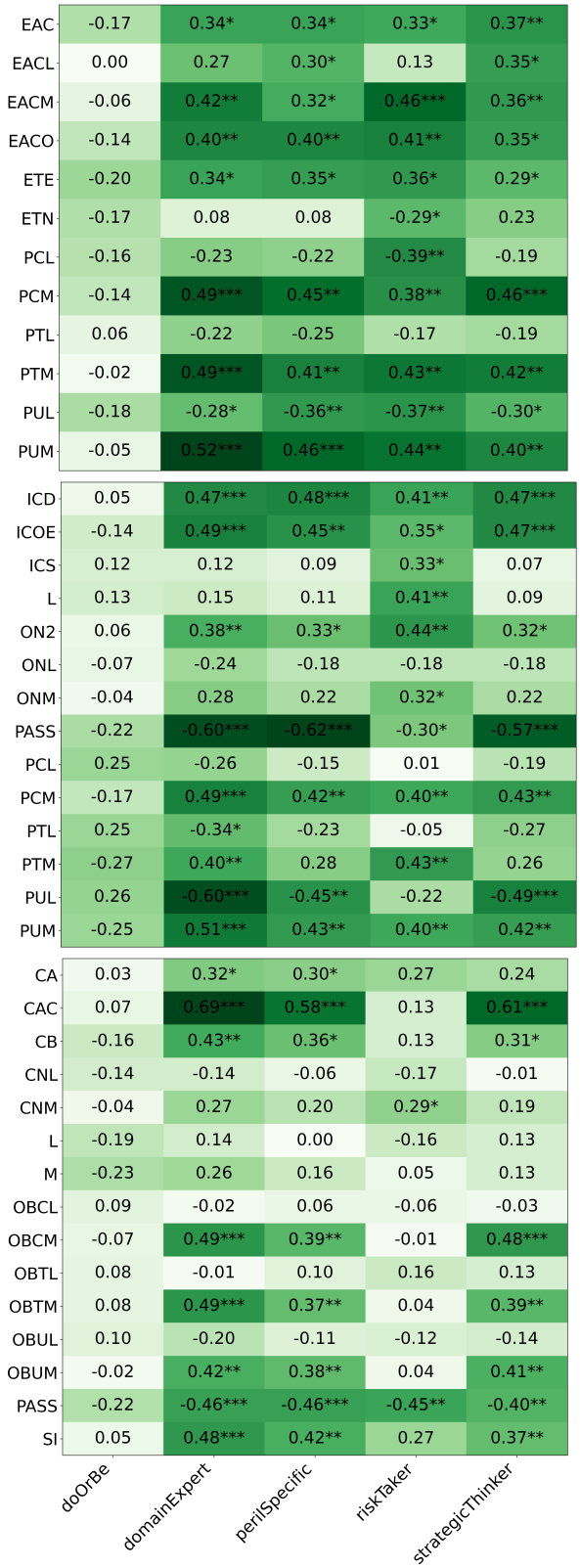

Figure 13: Heuristic Correlations - DH1 - LLaMA 4

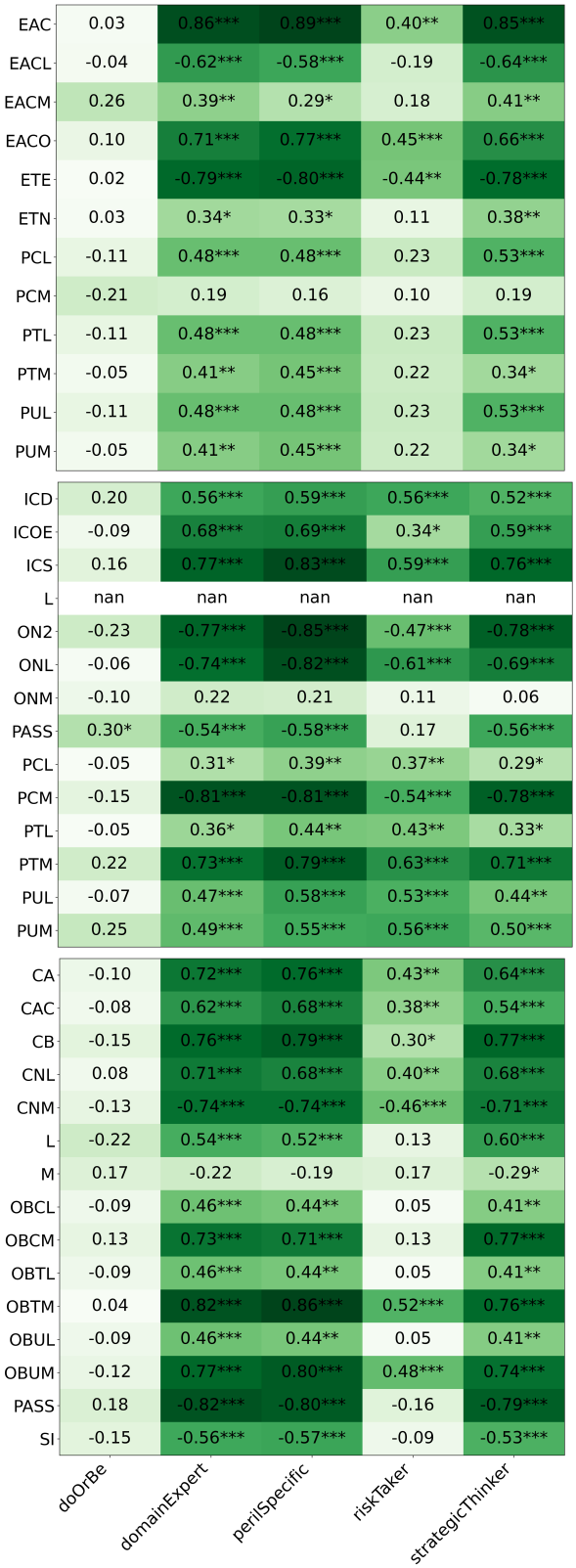

Figure 14: Heuristic Correlations - PI1 - LLaMA 4

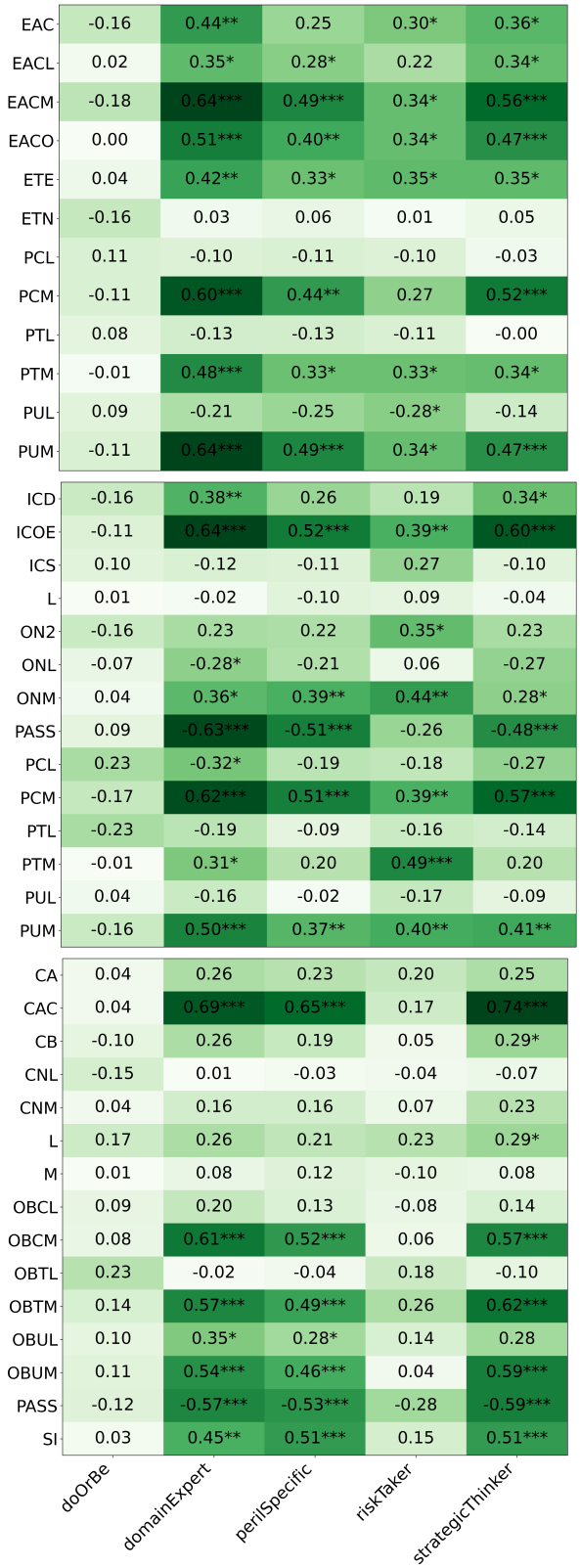

Figure 15: Heuristic Correlations - DH2 - LLaMA 4

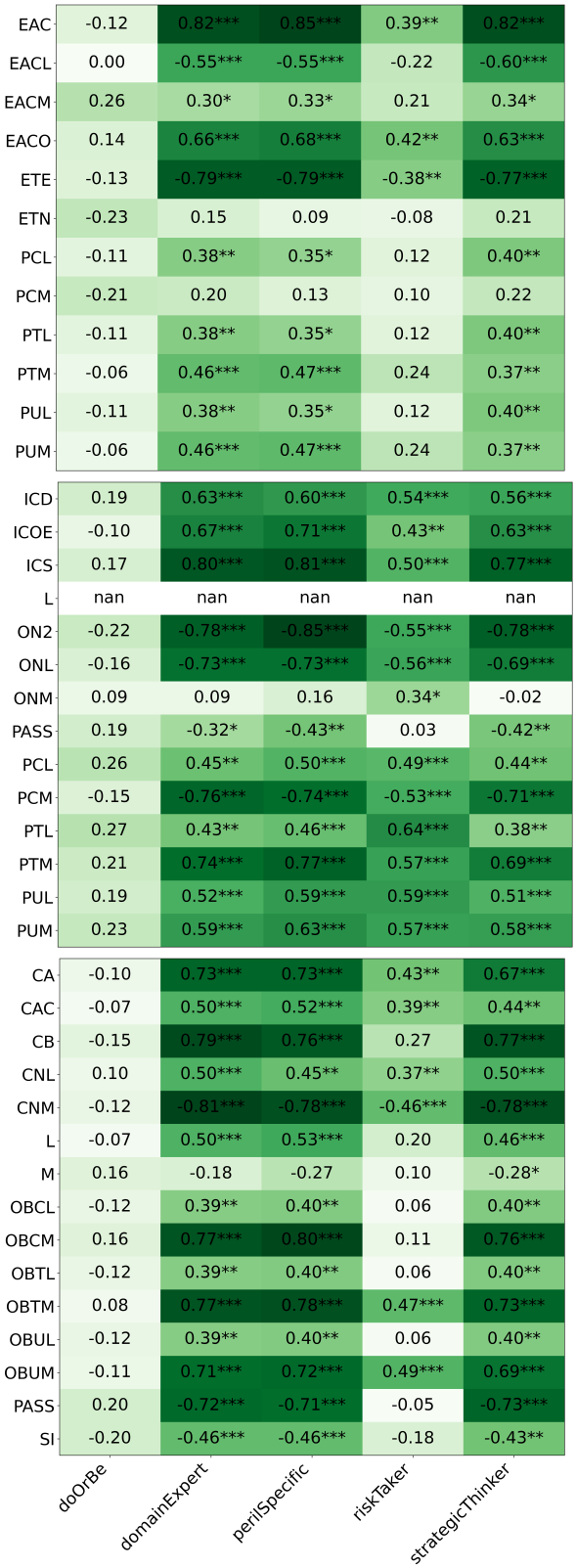

Figure 16: Heuristic Correlations - PI2 - LLaMA 4

A.5 Persona Descriptions

This section contains tables similar to Table 2. These tables show the top and bottom five performing personas based on their final TrueSkill ratings after 1200 tournament games. Relevant Tables: 5 - 28.

Rating	Persona Description
29.5914	A nanny who works with the widower to provide additional support and care for the children
29.1766	A literary critic who dismisses the Madam Tulip series as shallow and predictable
28.7519	A military historian writing a book on the impact of military leaders on shaping discipline and leadership
28.5374	A startup CTO who emphasizes rapid deployment and market responsiveness over meticulous code craftsmanship
28.2439	A recently promoted Major General in the Nigerian Army
18.3503	A U.S. Army veteran who served alongside the retired interpreter and shares a deep bond of mutual trust and respect
20.1616	A military historian specializing in the Napoleonic Wars, with a focus on the Peninsular Campaign and the strategies employed by both the Allied and French forces
20.6356	An amateur genealogist tracing family histories with possible connections to noble lineages in Scotland and England
21.2993	A struggling high school student who has no interest in biology
21.3212	A supply chain and logistics consultant for aviation and defense industries, aiming to analyze the effectiveness of Axis supply lines and Allied interdiction efforts

Table 5: TrueSkill for LLaMA 3 - DH1 - Run 1

Rating	Persona Description
29.1663	A retired naval officer with experience in aircraft carrier operations and a deep knowledge of US Navy history and traditions.
28.9431	A military historian writing a book on the impact of military leaders on shaping discipline and leadership.
28.7413	A retired military general with extensive experience in national defense, providing insights on border security strategies.
27.5964	A competitive collegiate football player always seeking for custom-designed team merchandise
27.3898	An ancient deity known for their wisdom in cosmic affairs and mastery of temporal phenomena
18.7251	A U.S. Army veteran who served alongside the retired interpreter and shares a deep bond of mutual trust and respect
19.4858	A military historian specializing in the Napoleonic Wars, with a focus on the Peninsular Campaign and the strategies employed by both the Allied and French forces.
21.9361	A photographer skilled at capturing detailed images of the artifacts.
21.9867	A geopolitical strategist who often appears on different networks presenting an alternative viewpoint on policies and events
22.0864	A 7-year-old child diagnosed with autism spectrum disorder, seeking support in managing their social behaviors

Table 6: TrueSkill for LLaMA 3 - DH1 - Run 2

Rating	Persona Description
30.2914	An elderly relative who relies on the journalist's explanations to stay up-to-date on the latest technology trends.
27.9751	A teenager struggling with anxiety and looking for coping mechanisms
27.7713	A government agency using GIS analysis to plan efficient land use and infrastructure development
27.7561	A military strategist in the Department of Defense, interested in the use of real-time satellite data to support decision-making and improve mission effectiveness.
27.6661	A struggling high school student who has no interest in biology.
19.4921	A young child who comes to the shop every day with their parents to buy Kinder Eggs
20.6314	A young child who laughs uncontrollably at the street performer's antics
20.8708	A curious toddler who eagerly explores and enjoys playing with the DIY toys
21.3603	A startup CTO who emphasizes rapid deployment and market responsiveness over meticulous code craftsmanship
21.3628	A grassroots activist advocating for defunding the police and investing in alternative community-based solutions

Table 7: TrueSkill for LLaMA 3 - PI1 - Run 1

Rating	Persona Description
28.9910	A retired intelligence officer who had previously worked for the CIA.
28.0686	A 7-year-old child diagnosed with autism spectrum disorder, seeking support in managing their social behaviors
27.8405	A retired military general with extensive experience in national defense, providing insights on border security strategies.
27.6526	A struggling high school student who has no interest in biology.
27.5939	A wargaming enthusiast who designs and simulates historical military scenarios to explore the decision-making processes, strategies, and outcomes of various engagements, including the Capture of Guam.
20.3536	A startup CTO who emphasizes rapid deployment and market responsiveness over meticulous code craftsmanship
21.4205	A young child who comes to the shop every day with their parents to buy Kinder Eggs
21.9011	A person who struggles with Discardia – a fear of throwing things away.
22.3848	A young child who laughs uncontrollably at the street performer's antics
22.5022	A nanny who works with the widower to provide additional support and care for the children

Table 8: TrueSkill for LLaMA 3 - PI1 - Run 2

Rating	Persona Description
29.5645	An elderly woman who relies on the apothecary for her herbal remedies and trusts their expertise.
29.0417	A genealogist studying the origins and variations of Arabic-language surnames, including Al-Marri and its related surnames such as Marri and Al Murrah.
28.7365	A young child fascinated by Disney movies and loves to hear stories from their collection
28.3787	A genealogist researching family histories connected to Biddeford, Maine.
28.0665	A military historian writing a book on the impact of military leaders on shaping discipline and leadership.
16.0860	A recently promoted Major General in the Nigerian Army
17.3984	A rare bird species critically affected by habitat loss
19.1394	An amateur genealogist tracing family histories with possible connections to noble lineages in Scotland and England.
20.2137	A traditional comedian who believes in adhering to mainstream comedy and disapproves of pushing boundaries
20.5467	A U.S. Army veteran who served alongside the retired interpreter and shares a deep bond of mutual trust and respect

Table 9: TrueSkill for LLaMA 3 - DH2 - Run 1

Rating	Persona Description
30.0065	A young child who comes to the shop every day with their parents to buy Kinder Eggs
29.1161	A 7-year-old child diagnosed with autism spectrum disorder, seeking support in managing their social behaviors
28.8581	A nanny who works with the widower to provide additional support and care for the children
28.8233	A genealogist researching family histories connected to Biddeford, Maine.
28.6546	A highly respected admiral known for their strategic thinking and ability to inspire and motivate sailors
14.7580	A rare bird species critically affected by habitat loss
15.9706	A traditional comedian who believes in adhering to mainstream comedy and disapproves of pushing boundaries
16.1548	A recently promoted Major General in the Nigerian Army
18.7521	An amateur genealogist tracing family histories with possible connections to noble lineages in Scotland and England.
20.2914	A seasoned military strategist sharing stories from their covert operations and offering guidance in a fast-paced world of intelligence

Table 10: TrueSkill for LLaMA 3 - DH2 - Run 2

Rating	Persona Description
30.4708	An Indian military personnel who served his country for 30 years and received the 50th Independence Anniversary Medal.
28.5833	A confused person who is not familiar with iGEM and Synthetic Biology
27.7430	A retired Navy veteran who now works as a military consultant and shares practical knowledge
27.4178	An elderly woman who relies on the apothecary for her herbal remedies and trusts their expertise.
27.3379	A retired intelligence officer who had previously worked for the CIA.
19.7349	A young child fascinated by Disney movies and loves to hear stories from their collection
20.3243	A supply chain and logistics consultant for aviation and defense industries, aiming to analyze the effectiveness of Axis supply lines and Allied interdiction efforts.
21.2134	A young child who laughs uncontrollably at the street performer's antics
21.3314	A curious toddler who eagerly explores and enjoys playing with the DIY toys
21.5132	A startup CTO who emphasizes rapid deployment and market responsiveness over meticulous code craftsmanship

Table 11: TrueSkill for LLaMA 3 - PI2 - Run 1

Rating	Persona Description
29.4089	A military tactician specializing in special operations, particularly in counter-terrorism and hostage situations, with a focus on strategic planning and equipment selection for elite units.
29.3058	An elderly woman who relies on the apothecary for her herbal remedies and trusts their expertise.
28.9895	A military historian writing a book on the impact of military leaders on shaping discipline and leadership.
28.0034	A traditional comedian who believes in adhering to mainstream comedy and disapproves of pushing boundaries
27.9695	A retired military general with extensive experience in national defense, providing insights on border security strategies.
18.5998	A young child who laughs uncontrollably at the street performer's antics
19.2635	A grassroots activist advocating for defunding the police and investing in alternative community-based solutions
19.3120	A startup CTO who emphasizes rapid deployment and market responsiveness over meticulous code craftsmanship
21.0551	A person who struggles with Discardia – a fear of throwing things away.
21.3481	A young child fascinated by Disney movies and loves to hear stories from their collection

Table 12: TrueSkill for LLaMA 3 - PI2 - Run 2

Rating	Persona Description
29.4089	A military tactician specializing in special operations, particularly in counter-terrorism and hostage situations, with a focus on strategic planning and equipment selection for elite units.
29.3058	An elderly woman who relies on the apothecary for her herbal remedies and trusts their expertise.
28.9895	A military historian writing a book on the impact of military leaders on shaping discipline and leadership.
28.0034	A traditional comedian who believes in adhering to mainstream comedy and disapproves of pushing boundaries
27.9695	A retired military general with extensive experience in national defense, providing insights on border security strategies.
18.5998	A young child who laughs uncontrollably at the street performer's antics
19.2635	A grassroots activist advocating for defunding the police and investing in alternative community-based solutions
19.3120	A startup CTO who emphasizes rapid deployment and market responsiveness over meticulous code craftsmanship
21.0551	A person who struggles with Discardia – a fear of throwing things away.
21.3481	A young child fascinated by Disney movies and loves to hear stories from their collection

Table 13: TrueSkill for LLaMA 4 - DH1 - Run 1

Rating	Persona Description
29.0400	A retired intelligence officer who had previously worked for the CIA.
28.5427	A young child who comes to the shop every day with their parents to buy Kinder Eggs
28.4108	A genealogist helping clients trace their family roots, particularly those with connections to the Somme department in France.
27.7573	A military historian writing a book on the impact of military leaders on shaping discipline and leadership.
27.6677	A teenager struggling with anxiety and looking for coping mechanisms
16.8731	A curious toddler who eagerly explores and enjoys playing with the DIY toys
20.3059	A nanny who works with the widower to provide additional support and care for the children
20.6517	A struggling high school student who has no interest in biology.
21.2158	A photographer skilled at capturing detailed images of the artifacts.
21.2810	A wargaming enthusiast who enjoys designing and playing strategic simulations of historical military conflicts, particularly those involving Chinese forces.

Table 14: TrueSkill for LLaMA 4 - DH1 - Run 2

Rating	Persona Description
28.4912	A competitive collegiate football player always seeking for custom-designed team merchandise
28.3990	A retired naval officer with experience in aircraft carrier operations and a deep knowledge of US Navy history and traditions.
27.7828	An Indian military personnel who served his country for 30 years and received the 50th Independence Anniversary Medal.
27.6983	A geopolitical strategist who often appears on different networks presenting an alternative viewpoint on policies and events
27.5775	A retired intelligence officer who had previously worked for the CIA.
20.3577	A graphic designer seeking advice on how to prevent repetitive strain injuries
21.1707	A curious toddler who eagerly explores and enjoys playing with the DIY toys
21.6260	A 7-year-old child diagnosed with autism spectrum disorder, seeking support in managing their social behaviors
21.6414	A young child who laughs uncontrollably at the street performer's antics
21.7782	A photographer skilled at capturing detailed images of the artifacts.

Table 15: TrueSkill for LLaMA 4 - PI1 - Run 1

Rating	Persona Description
29.4422	A retired military general with extensive experience in national defense, providing insights on border security strategies.
29.1101	A retired Navy veteran who now works as a military consultant and shares practical knowledge
28.5151	A struggling high school student who has no interest in biology.
27.9740	A genealogist helping clients trace their family roots, particularly those with connections to the Somme department in France.
27.8130	A geopolitical strategist who often appears on different networks presenting an alternative viewpoint on policies and events
19.0300	A young child who laughs uncontrollably at the street performer's antics
19.6486	A startup CTO who emphasizes rapid deployment and market responsiveness over meticulous code craftsmanship
19.9231	A graphic designer seeking advice on how to prevent repetitive strain injuries
20.9841	A 7-year-old child diagnosed with autism spectrum disorder, seeking support in managing their social behaviors
21.1041	A curious toddler who eagerly explores and enjoys playing with the DIY toys

Table 16: TrueSkill for LLaMA 4 - PI1 - Run 2

Rating	Persona Description
30.0335	A grassroots activist advocating for defunding the police and investing in alternative community-based solutions
28.3202	A healthcare blogger who spreads misinformation about vaccines and challenges the nurse's beliefs
28.1103	A retired naval officer with experience in aircraft carrier operations and a deep knowledge of US Navy history and traditions.
27.9493	A wargaming enthusiast who enjoys designing and playing strategic simulations of historical military conflicts, particularly those involving Chinese forces.
27.7971	An elderly woman who relies on the apothecary for her herbal remedies and trusts their expertise.
16.9432	A teenager struggling with anxiety and looking for coping mechanisms
19.2549	A traditional comedian who believes in adhering to mainstream comedy and disapproves of pushing boundaries
20.7565	A confused person who is not familiar with iGEM and Synthetic Biology
21.7609	A struggling high school student who has no interest in biology.
22.7316	A retired Navy veteran who now works as a military consultant and shares practical knowledge

Table 17: TrueSkill for LLaMA 4 - DH2 - Run 1

Rating	Persona Description
28.1846	A supply chain and logistics consultant for aviation and defense industries, aiming to analyze the effectiveness of Axis supply lines and Allied interdiction efforts.
27.8466	A geopolitical strategist who often appears on different networks presenting an alternative viewpoint on policies and events
27.8174	A startup CTO who emphasizes rapid deployment and market responsiveness over meticulous code craftsmanship
27.5886	A grassroots activist advocating for defunding the police and investing in alternative community-based solutions
27.5369	An elderly woman who relies on the apothecary for her herbal remedies and trusts their expertise.
21.3166	A confused person who is not familiar with iGEM and Synthetic Biology
21.5359	A struggling high school student who has no interest in biology.
21.7931	A disaster relief coordinator working on preparedness and emergency response plans for towns in Tornado Alley
21.9628	A traditional comedian who believes in adhering to mainstream comedy and disapproves of pushing boundaries
22.9009	A genealogist researching family histories connected to Biddeford, Maine.

Table 18: TrueSkill for LLaMA 4 - DH2 - Run 2

Rating	Persona Description
28.4403	A nanny who works with the widower to provide additional support and care for the children
28.0372	A geopolitical strategist who often appears on different networks presenting an alternative viewpoint on policies and events
27.4786	A highly respected admiral known for their strategic thinking and ability to inspire and motivate sailors
27.3847	A rare bird species critically affected by habitat loss
27.2912	A retired naval officer with experience in aircraft carrier operations and a deep knowledge of US Navy history and traditions.
20.3037	A healthcare blogger who spreads misinformation about vaccines and challenges the nurse's beliefs
20.7256	A curious toddler who eagerly explores and enjoys playing with the DIY toys
21.2932	A startup CTO who emphasizes rapid deployment and market responsiveness over meticulous code craftsmanship
21.6368	An elderly French woman who shares stories of her youth in Paris and offers traditional baking tips.
21.6622	A young child who laughs uncontrollably at the street performer's antics

Table 19: TrueSkill for LLaMA 4 - PI2 - Run 1

Rating	Persona Description
30.2238	A competitive collegiate football player always seeking for custom-designed team merchandise
29.2156	A nanny who works with the widower to provide additional support and care for the children
28.2542	A teenager struggling with anxiety and looking for coping mechanisms
27.9331	A wargaming enthusiast who designs and simulates historical military scenarios to explore the decision-making processes, strategies, and outcomes of various engagements, including the Capture of Guam.
27.9087	A retired naval officer with experience in aircraft carrier operations and a deep knowledge of US Navy history and traditions.
19.4226	A curious toddler who eagerly explores and enjoys playing with the DIY toys
20.1626	A 7-year-old child diagnosed with autism spectrum disorder, seeking support in managing their social behaviors
21.2523	A government agency using GIS analysis to plan efficient land use and infrastructure development
21.5645	A startup CTO who emphasizes rapid deployment and market responsiveness over meticulous code craftsmanship
21.6047	A healthcare blogger who spreads misinformation about vaccines and challenges the nurse's beliefs

Table 20: TrueSkill for LLaMA 4 - PI2 - Run 2

Rating	Persona Description
29.8093	An elderly relative who relies on the journalist's explanations to stay up-to-date on the latest technology trends.
27.7350	A teenager struggling with anxiety and looking for coping mechanisms
27.6069	A grassroots activist advocating for defunding the police and investing in alternative community-based solutions
27.5729	A retired intelligence officer who had previously worked for the CIA.
27.3498	A 7-year-old child diagnosed with autism spectrum disorder, seeking support in managing their social behaviors
21.0631	A risk management consultant advising organizations on safety measures in high-risk regions, including Afghanistan.
21.3249	A confused person who is not familiar with iGEM and Synthetic Biology
21.4664	An elderly woman who relies on the apothecary for her herbal remedies and trusts their expertise.
21.8820	A genealogist studying the origins and variations of Arabic-language surnames, including Al-Marri and its related surnames such as Marri and Al Murrah.
22.1924	A graphic designer seeking advice on how to prevent repetitive strain injuries

Table 21: TrueSkill for Mistral Small - DH1 - Run 1

Rating	Persona Description
29.3532	A teenager struggling with anxiety and looking for coping mechanisms
28.5395	A military strategist in the Department of Defense, interested in the use of real-time satellite data to support decision-making and improve mission effectiveness.
28.5144	An elderly relative who relies on the journalist's explanations to stay up-to-date on the latest technology trends.
28.4752	A 7-year-old child diagnosed with autism spectrum disorder, seeking support in managing their social behaviors
28.0034	A young child fascinated by Disney movies and loves to hear stories from their collection
20.0316	A risk management consultant advising organizations on safety measures in high-risk regions, including Afghanistan.
20.0726	A person who struggles with Discardia – a fear of throwing things away.
22.0895	A young child who laughs uncontrollably at the street performer's antics
22.1544	A startup CTO who emphasizes rapid deployment and market responsiveness over meticulous code craftsmanship
22.2729	A healthcare blogger who spreads misinformation about vaccines and challenges the nurse's beliefs

Table 22: TrueSkill for Mistral Small - DH1 - Run 2

Rating	Persona Description
30.7565	A seasoned military strategist sharing stories from their covert operations and offering guidance in a fast-paced world of intelligence
29.9239	A military historian specializing in the Napoleonic Wars, with a focus on the Peninsular Campaign and the strategies employed by both the Allied and French forces.
28.7850	A nanny who works with the widower to provide additional support and care for the children
28.3912	A government agency using GIS analysis to plan efficient land use and infrastructure development
28.1505	A retired Navy veteran who now works as a military consultant and shares practical knowledge
19.6633	A young child who laughs uncontrollably at the street performer's antics
20.8942	A genealogist researching family histories connected to Biddeford, Maine.
21.5948	A curious toddler who eagerly explores and enjoys playing with the DIY toys
21.6085	A young child who comes to the shop every day with their parents to buy Kinder Eggs
21.8016	A struggling high school student who has no interest in biology.

Table 23: TrueSkill for Mistral Small - PI1 - Run 1

Rating	Persona Description
28.9724	Another military strategist from a rival nation, constantly attempting to outwit and outmaneuver
28.1851	A government agency using GIS analysis to plan efficient land use and infrastructure development
28.1572	A wargaming enthusiast who enjoys designing and playing strategic simulations of historical military conflicts, particularly those involving Chinese forces.
28.0173	A recently promoted Major General in the Nigerian Army
27.4219	A wargaming enthusiast who designs and simulates historical military scenarios to explore the decision-making processes, strategies, and outcomes of various engagements, including the Capture of Guam.
18.5387	A young child who laughs uncontrollably at the street performer's antics
19.4294	A struggling high school student who has no interest in biology.
19.7970	A curious toddler who eagerly explores and enjoys playing with the DIY toys
22.2818	A genealogist researching family histories connected to Biddeford, Maine.
22.3399	A young child who comes to the shop every day with their parents to buy Kinder Eggs

Table 24: TrueSkill for Mistral Small - PI1 - Run 2

Rating	Persona Description
31.4374	A nanny who works with the widower to provide additional support and care for the children
29.8648	A struggling high school student who has no interest in biology.
28.6752	A teenager struggling with anxiety and looking for coping mechanisms
28.5229	A retired military general with extensive experience in national defense, providing insights on border security strategies.
28.3828	A wargaming enthusiast who enjoys designing and playing strategic simulations of historical military conflicts, particularly those involving Chinese forces.
17.7759	a confused person who is not familiar with iGEM and Synthetic Biology
17.8619	A young child who laughs uncontrollably at the street performer's antics
18.0121	A person who struggles with Discardia – a fear of throwing things away.
18.8108	A literary critic who dismisses the Madam Tulip series as shallow and predictable.
20.9075	A genealogist researching family histories connected to Biddeford, Maine.

Table 25: TrueSkill for Mistral Small - DH2 - Run 1

Rating	Persona Description
29.0947	An elderly relative who relies on the journalist's explanations to stay up-to-date on the latest technology trends.
28.8704	A 7-year-old child diagnosed with autism spectrum disorder, seeking support in managing their social behaviors
28.5627	A young child who comes to the shop every day with their parents to buy Kinder Eggs
28.1359	A retired military general with extensive experience in national defense, providing insights on border security strategies.
27.6784	A nanny who works with the widower to provide additional support and care for the children
15.6622	A person who struggles with Discardia – a fear of throwing things away.
17.5323	A young child who laughs uncontrollably at the street performer's antics
17.8052	a confused person who is not familiar with iGEM and Synthetic Biology
18.3785	A literary critic who dismisses the Madam Tulip series as shallow and predictable.
20.2392	A risk management consultant advising organizations on safety measures in high-risk regions, including Afghanistan.

Table 26: TrueSkill for Mistral Small - DH2 - Run 2

Rating	Persona Description
30.0935	A seasoned military strategist sharing stories from their covert operations and offering guidance in a fast-paced world of intelligence
29.1228	A wargaming enthusiast who enjoys designing and playing strategic simulations of historical military conflicts, particularly those involving Chinese forces.
29.0169	A geopolitical strategist who often appears on different networks presenting an alternative viewpoint on policies and events
28.1472	A wargaming enthusiast who designs and simulates historical military scenarios to explore the decision-making processes, strategies, and outcomes of various engagements, including the Capture of Guam.
27.3938	A government agency using GIS analysis to plan efficient land use and infrastructure development
20.5554	A young child who laughs uncontrollably at the street performer's antics
20.5604	A genealogist researching family histories connected to Biddeford, Maine.
21.3149	A curious toddler who eagerly explores and enjoys playing with the DIY toys
21.3640	A young child who comes to the shop every day with their parents to buy Kinder Eggs
22.0657	A literary critic who dismisses the Madam Tulip series as shallow and predictable.

Table 27: TrueSkill for Mistral Small - PI2 - Run 1

Rating	Persona Description
28.0526	A 7-year-old child diagnosed with autism spectrum disorder, seeking support in managing their social behaviors
28.0013	A competitive collegiate football player always seeking for custom-designed team merchandise
27.7001	A military historian specializing in the Napoleonic Wars, with a focus on the Peninsular Campaign and the strategies employed by both the Allied and French forces.
27.6578	An Indian military personnel who served his country for 30 years and received the 50th Independence Anniversary Medal.
27.4933	A wargaming enthusiast who designs and simulates historical military scenarios to explore the decision-making processes, strategies, and outcomes of various engagements, including the Capture of Guam.
20.2685	A young child who laughs uncontrollably at the street performer's antics
21.0522	A young child who comes to the shop every day with their parents to buy Kinder Eggs
21.1988	A healthcare blogger who spreads misinformation about vaccines and challenges the nurse's beliefs
21.8348	a confused person who is not familiar with iGEM and Synthetic Biology
22.1529	A struggling high school student who has no interest in biology.

Table 28: TrueSkill for Mistral Small - PI2 - Run 2

A.6 Opposite Value Consistency

As discussed in Table 4, the direct heuristic method almost always results in values that are significantly higher than those generated by the inventory prompts. Relevant Tables: 29 - 35.

Phase	Categories	DH avg.	PI avg.
1	PTM, PTL	5.02	**7.63**
	PUM, PUL	4.45	**6.08**
	PCM, PCL	6.32	**19.35**
	ETE, ETN	6.84	**10.84**
	EACM, EACL	4.58	**17.58**
2	PTM, PTL	3.26	**8.46**
	PUM, PUL	4.09	**8.29**
	PCM, PCL	4.04	**51.91**
	ONM, ONL	9.99	**10.60**
	ICD, ICS	2.01	**16.63**
3	OBTM, OBTL	3.10	**18.43**
	OBUM, OBUL	3.49	**14.77**
	OBCM, OBCL	**4.64**	3.28
	CNM, CNL	3.66	**41.40**
	M, L	2.73	**21.57**

Table 29: Direct heuristics (DH avg.) and inventory heuristics (PI avg.) by heuristic pair and phase for the first heuristic set generated with GPT4.

Phase	Heuristic Pair	DH avg.	PI avg.
0	EACM-EACL	31.51	**11.60**
0	ETE-ETN	10.43	**7.03**
0	PCM-PCL	32.27	**4.20**
0	PTM-PTL	26.52	**4.44**
0	PUM-PUL	28.87	**3.83**
1	ICD-ICS	16.81	**9.80**
1	ONM-ONL	31.07	**13.61**
1	PCM-PCL	22.06	**11.25**
1	PTM-PTL	24.36	**3.66**
1	PUM-PUL	22.32	**3.53**
2	CNM-CNL	5.74	**12.35**
2	M-L	**5.82**	10.69
2	OBCM-OBCL	16.99	**3.50**
2	OBTM-OBTL	15.35	**7.15**
2	OBUM-OBUL	9.26	**6.08**

Table 30: Direct heuristics (DH avg.) and inventory heuristics (PI avg.) by heuristic pair and phase for the first heuristic set generated with Mistral Small.

Phase	Heuristic Pair	DH avg.	PI avg.
0	EACM-EACL	27.89	**11.48**
0	ETE-ETN	11.82	**7.37**
0	PCM-PCL	28.82	**4.54**
0	PTM-PTL	28.45	**4.50**
0	PUM-PUL	26.47	**3.87**
1	ICD-ICS	20.75	**10.27**
1	ONM-ONL	47.20	**12.23**
1	PCM-PCL	25.98	**11.10**
1	PTM-PTL	25.82	**3.54**
1	PUM-PUL	21.25	**3.87**
2	CNM-CNL	**10.32**	13.66
2	M-L	**8.10**	10.21
2	OBCM-OBCL	18.57	**3.65**
2	OBTM-OBTL	7.94	**7.49**
2	OBUM-OBUL	8.20	**6.19**

Table 31: Direct heuristics (DH avg.) and inventory heuristics (PI avg.) by heuristic pair and phase for the second heuristic set generated with Mistral Small.

Phase	Heuristic Pair	DH avg.	PI avg.
0	EACM-EACL	47.19	**10.94**
0	ETE-ETN	15.04	**8.65**
0	PCM-PCL	22.27	**4.57**
0	PTM-PTL	41.73	**5.14**
0	PUM-PUL	24.31	**4.38**
1	ICD-ICS	19.54	**9.73**
1	ONM-ONL	20.37	**11.78**
1	PCM-PCL	34.46	**12.93**
1	PTM-PTL	13.68	**1.97**
1	PUM-PUL	21.50	**2.72**
2	CNM-CNL	**7.46**	9.28
2	M-L	25.24	**10.02**
2	OBCM-OBCL	45.90	**4.90**
2	OBTM-OBTL	26.08	**4.06**
2	OBUM-OBUL	43.79	**3.11**

Table 32: Direct heuristics (DH avg.) and inventory heuristics (PI avg.) by heuristic pair and phase for the first heuristic set generated with LLaMA 3.

Phase	Heuristic Pair	DH avg.	PI avg.
0	EACM-EACL	50.77	**10.52**
0	ETE-ETN	18.50	**7.26**
0	PCM-PCL	21.96	**4.74**
0	PTM-PTL	13.96	**5.64**
0	PUM-PUL	29.62	**4.79**
1	ICD-ICS	48.57	**7.79**
1	ONM-ONL	18.75	**11.88**
1	PCM-PCL	26.71	**14.26**
1	PTM-PTL	12.53	**1.86**
1	PUM-PUL	12.91	**2.21**
2	CNM-CNL	**3.35**	11.37
2	M-L	16.16	8.17
2	OBCM-OBCL	29.11	**3.82**
2	OBTM-OBTL	35.31	**4.35**
2	OBUM-OBUL	40.81	**3.43**

Table 33: Direct heuristics (DH avg.) and inventory heuristics (PI avg.) by heuristic pair and phase for the second heuristic set generated with LLaMA 3.

Phase	Heuristic Pair	DH avg.	PI avg.
0	EACM-EACL	18.17	**11.53**
0	ETE-ETN	44.36	**7.08**
0	PCM-PCL	43.85	**4.85**
0	PTM-PTL	23.07	**2.96**
0	PUM-PUL	42.09	**2.63**
1	ICD-ICS	**2.05**	5.51
1	ONM-ONL	83.72	**14.55**
1	PCM-PCL	27.81	**14.07**
1	PTM-PTL	14.19	**4.62**
1	PUM-PUL	15.94	**2.41**
2	CNM-CNL	52.89	**15.15**
2	M-L	30.02	**12.77**
2	OBCM-OBCL	29.95	**4.92**
2	OBTM-OBTL	34.30	**4.72**
2	OBUM-OBUL	24.43	**3.68**

Table 34: Direct heuristics (DH avg.) and inventory heuristics (PI avg.) by heuristic pair and phase for the first heuristic set generated with LLaMA 4.

Phase	Heuristic Pair	DH avg.	PI avg.
0	EACM-EACL	16.34	**9.95**
0	ETE-ETN	47.17	**6.19**
0	PCM-PCL	25.00	**5.11**
0	PTM-PTL	19.76	**2.78**
0	PUM-PUL	30.67	**2.49**
1	ICD-ICS	**3.58**	4.21
1	ONM-ONL	86.18	**12.95**
1	PCM-PCL	20.96	**13.90**
1	PTM-PTL	19.83	**3.88**
1	PUM-PUL	17.00	**2.27**
2	CNM-CNL	52.39	**13.15**
2	M-L	28.34	**11.74**
2	OBCM-OBCL	21.32	**3.99**
2	OBTM-OBTL	15.83	**3.94**
2	OBUM-OBUL	21.14	**3.47**

Table 35: Direct heuristics (DH avg.) and inventory heuristics (PI avg.) by heuristic pair and phase for the second heuristic set generated with LLaMA 4.