# Persona Non Grata: LLM Persona-Driven Generations in MCQA are Unstable in Distinct Dimensions

**César Guerra-Solano**[1]*, **Xiang Lorraine Li**[2]
[1]Department of Computer Sciences, University of Wisconsin-Madison
[2]Department of Computer Science, University of Pittsburgh
**Correspondence:** guerrasolano@wisc.edu, xianglli@pitt.edu

## Abstract

Persona-driven generations (PDGs) have seen prolific use in research and industry applications, where a large language model (LLM) takes on a "persona" while completing some task. While persona expressed through free-form text (like dialogue) has substantial work investigating stability or consistency, relatively, persona expressed in non-text-heavy outputs (like in multiple-choice question answering, or MCQA) is often overlooked. We work to address this gap, seeking to understand the instability of LLM PDGs in MCQA tasks. We develop three metrics investigating the performance, outcome, and question correctness stability, evaluating three distinct dimensions. Using these metrics, we find that instability varies consistently between model families and model size, and across question domains, with math/commonsense questions leading to greater instability. We also find task prompt format introduces more prediction instability than other hyperparameters, like temperature. Finally, we find that instability is related to task accuracy, and using our instability metrics, find different experimental settings that result in different best and worst personas for tasks, despite their similarity. This reveals the importance of checking hyperparameter instability in PDGs.

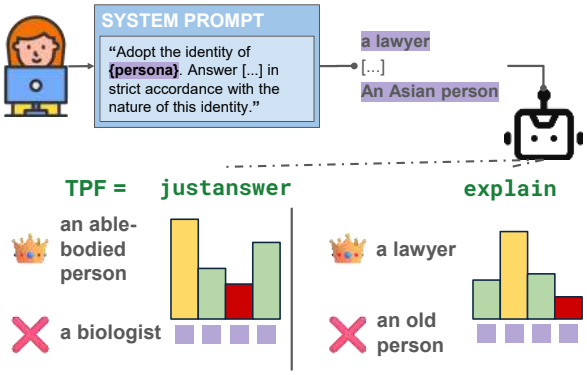



Figure 1: Depicting instability in non-text-heavy persona-driven generations (PDGs). Given different, although still reasonable, experiment configurations, represented by different settings for the task prompt format (TPF) hyperparameter, an LLM's persona-driven performance on a multiple-choice evaluation differs greatly.

sectionIntroduction

Large language models (LLMs) have seen prolific use in a variety of domains, with their adaptability and broad capabilities lending themselves to high performance across a range of tasks and opportunities for personalization for user- and task-specific needs (Kojima et al., 2022; Brown et al., 2020; Grattafiori et al., 2024; Yang et al., 2025; Zhang et al., 2024). With this, **persona-driven generations (PDGs)** have become prevalent – here, by leveraging the power of prompting, LLMs can role-play as a "persona" while carrying out some task, such as general user assistance (e.g. "You are a teacher. Explain [...]") or more specific task completion (e.g. "You are a hiring manager. Rate these resumes [...]"). As PDGs see use in high-stakes domains, from medicine to education, evaluating their stability is critical (Sun et al., 2025; Yuan et al., 2025; Kyung et al., 2025; Li et al., 2024).

With PDGs, two main uses can be observed: in free-form text generation settings, such as in dialogue systems like Character.AI (Character.AI), and non-text-heavy outputs, such as in persona-based multiple-choice question answering (MCQA). In the latter case, LLM persona is expressed through task completion: rather than producing text that reflects the *style* of a persona, an LLM answers questions in accordance with the *capabilities* associated with that persona, when relevant (e.g., a biologist answering biology questions), or to remain unaffected by persona choice when the persona is irrelevant (e.g., race should not influence performance on mathematical questions).

However, relative to text generation, little work has characterized or improved PDGs in MCQA.

*Work completed while at the University of Pittsburgh.



*Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 2308–2332
November 6-10, 2026 ©2026 Association for Computational Linguistics

Prior studies indicate that PDGs in this setting are implicitly biased and unpredictable (Gupta et al., 2024; Zheng et al., 2024), motivating further work to characterize them and their potential flaws. Additionally, these studies bear little standardization, with large differences in model choice and hyperparameters across their experiment settings – features known to affect model performance, thus being targets of optimization (Wang et al., 2023; Sclar et al., 2023). With the diverse experiment settings that PDGs are used in, differing in persona, task, and prompt format, **additional investigations are necessary to find the effects of these variations** to ensure current and future findings are robust.

We work to address this gap, seeking to further understand PDGs in MCQA, quantifying their stability across diverse environments. We do so relative to **three persona-specific metrics**, targeting unique and distinct features of PDG paradigms – we look at not only stability of **accuracy** as in related past work (Gupta et al., 2024; Zheng et al., 2024; de Araujo et al., 2025), but also stability of **experimental outcomes** and **persona-specific question correctness**.

We evaluate 5 open-source models across 48 experiment settings and 41 personas, finding:

- PDGs in MCQA are **heavily unstable**, with great variation across similar experiment settings used in past work, and consistent patterns with model size and family in the models we evaluate.
- Instability is **specific**, with distinct relationships relative to 3 dimensions of instability, types of content in the input prompt (such as the question formatting vs. persona formatting), and the subject domain of the task.
- Our instability metrics are **actionable**, with demonstrated usage in identifying stable and unstable experiment settings that greatly differ in overall accuracy and persona-related conclusions, such as the best persona for a task, or inter-persona distribution shifts.

## 1 Related Work

**Persona-Driven Generations** Persona-based generation is widely used across LLM downstream tasks, including dialogue systems (Huang et al., 2023), narrative generation (Zhang et al., 2024), and LLM-as-a-judge settings where models adopt expert roles (Dong et al., 2024; Yu et al., 2024; Li et al., 2025). These applications generally fall into two categories: free-form text generation, where persona is embedded into generation style and content (e.g., Character.AI (Character.AI)), and non-text-heavy tasks, like persona-based task-taking, where persona is reflected in task performance.

Evaluation of PDGs has primarily focused on text-heavy settings, examining consistency, stability, and failures such as sensitivity or bias in long-form generation (Zhang et al., 2025; Liu et al., 2024; Maharana et al., 2024; Tseng et al., 2024; de Araujo et al., 2025; Cheng et al., 2023). By contrast, non-text-heavy PDGs—though common in tasks like LLM-as-a-judge or MCQA (de Araujo et al., 2025; Kyung et al., 2025; Salewski et al., 2023; Tang et al., 2024)[1]—remain underexplored, with only limited attention to bias, faithfulness, or unpredictability (Zheng et al., 2024; Gupta et al., 2024; de Araujo et al., 2025), leaving a key gap in comprehensively assessing their robustness.

Our work is related to de Araujo et al. (2025) which evaluates how well personas *match human expectations or needs* with personas, like if expert personas confer a performance advantage. However, we focus on how *stable* PDGs are, i.e, the consistency of MCQA performance given personas across noted variations in experiment setting such as prompt format and generation temperature.

**Generation Sensitivity** Prior work on model generation sensitivity has primarily focused on model temperature and input prompt styles/formatting, with evaluation metrics emphasizing macro-level aspects such as task or response structure and performance (Sclar et al., 2023; Chatterjee et al., 2024; Li et al., 2025; Brown et al., 2020; He et al., 2024). However, these studies primarily center on persona-agnostic contexts, leaving the need to investigate how experiment setting variations influence LLMs' role-playing capabilities unaddressed. Moreover, existing sensitivity metrics (Zhuo et al., 2024) largely focus on accuracy, overlooking outcome- and question-level sensitivity, a gap our work addresses.

## 2 Instability Metrics

In this paper, we define instability in persona-driven generations as robustness to experimental settings. We examine how evaluation trends shift across hyperparameter settings such as prompt

[1]Please look at de Araujo et al. (2025)'s Table 3 for a compilation of PDG usage in prior literature, including in tasks with objective correctness.

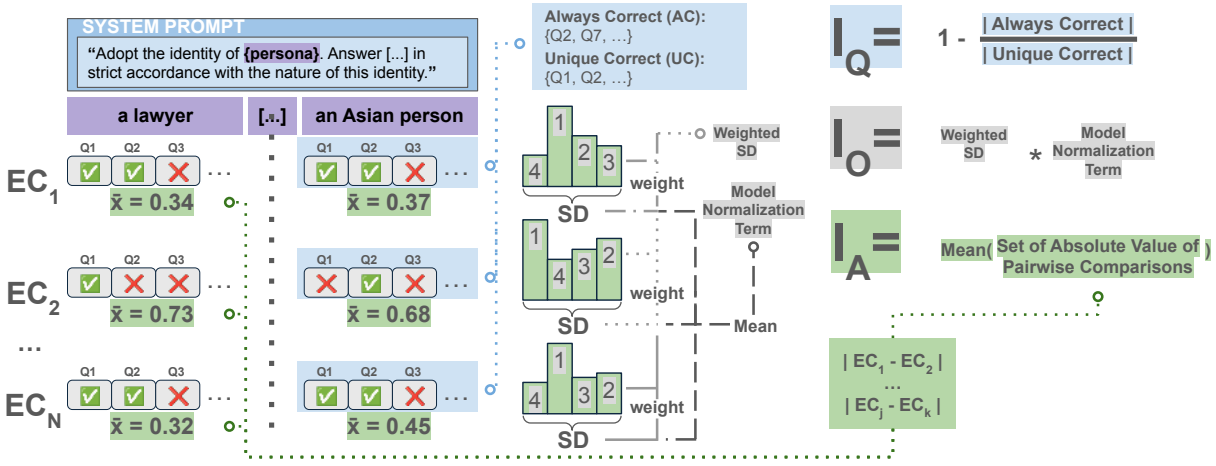



Figure 2: A figure depicting how instability metrics are calculated based on PDGs for our dataset. Across runs in several unique experimental configurations (ECs), we consider mean performance and question-wise correctness. Mean performance is used for standard deviation calculations for computing $I_A$ and used to get inter-persona rankings within one EC for $I_O$. The set of correct questions for a persona across all ECs is used for $I_Q$ calculation. Calculations shown are for one persona, although all personas are integrated as shown in Section 2. We explore how varying the question set and experiment setting set sizes affects metric calculations in Appendices E.6 and E.7.

format and model temperature, with each setting $s \in S$ representing a hyperparameter combination.

To quantify performance variation, we propose three metrics that measure differences in **accuracy**, **outcome**, and **correct questions**. Prior work (Zhuo et al., 2024; Sclar et al., 2023) has primarily focused only on accuracy differences. However, instability exists agnostic of accuracy – you can have large accuracy differences but no difference in inter-persona ranking, or identical accuracy but great inconsistency in *what* questions are answered correctly. Thus, further dimensions of instability should be investigated, especially in PDGs. We define each metric using the following terms.[2]

**Experimental Settings:** Each metric is calculated relative to a set of considered experimental settings, $S$. This allows us to analyze instability on broader (e.g., considering all experimental settings) or smaller (e.g., varying one hyperparameter within an otherwise fixed setting) scales.

**Personas:** Each metric considers a set of personas, $P$, that are provided to the model. Each persona is evaluated under multiple experimental settings. Expectations and desiderata with respect to how performance should change (if at all) when varying personas can vary based on the task and specific personas considered – please look to our discussion in Appendix D.2.1 for more information.

**Persona Accuracy:** The accuracy of a persona $p$ is calculated for a set of questions $Q$ under an experimental setting $s$ denoted as, $Acc(p; s, Q)$. Note that the question set $Q$ could be the entire dataset, or specific subsets, such as domain-specific questions.

**Persona Rank:** Given persona accuracy $Acc(p; s, Q)$, the rank of a persona $p$ relative to other personas is defined as the index of $p$ in a list ranked by accuracy, denoted as $R(p; s, Q)$.[3]

## 2.1 Performance Instability: $I_A$

The first metric aims to understand how sensitive model accuracy is across all experimental settings (averaged over personas). Intuitively, we seek to understand that, if we randomly selected two experimental settings from a set of potential settings, what is the expected difference in accuracy?

$$I_A(p) = \frac{1}{\binom{|S|}{2}} \sum_{s_i, s_j \in S} \left| \text{Acc}(p; s_i) - \text{Acc}(p; s_j) \right|$$

$$I_A = \frac{1}{|P|} \sum_{p \in P} I_A(p)$$

The formulation is similar to Zhuo et al. (2024), but fitted to a persona-specific context. For a persona, we calculate the pair-wise accuracy differences between $s_i$ and $s_j$ for all experiment settings, then average them. $\binom{|S|}{2}$ represents the count of experiment settings pairs for a single persona. We

[2]For validation, we report the correlation of our metric results with metrics from existing literature or reasonable correlates in Appendix E.4.

[3]In the rest of section 2, we omit $Q$ in accuracy $Acc(p; s)$ and rank $R(p; s)$ as $Q$ is implied in computing these metrics.

aggregate the persona instability to represent the accuracy instability $I_A$ for a given question set, and a model, where $|P|$ is the total number of personas.

### 2.2 Outcome Instability $I_O$

This metric measures how variable the conclusion of a persona-driven experiment is, such as identifying the best or worst persona for a task. To do this, we compare persona ranking $R(p; s)$ across experimental settings. Importantly, we note that for each experimental setting, both minor and substantial differences in persona average accuracies can change their rankings, leading to outcome instability. However, intuitively, we argue that the degree of this instability is different: larger accuracy differences should have a larger outcome instability score. To account for this, we define the outcome instability of each persona as the weighted standard deviation of its rankings, where the weight assigned to each experimental setting corresponds to the standard deviation across all personas' accuracies in that setting. At the model level, overall outcome instability is computed by averaging the persona-level instabilities and scaling by the average of $w_s$ across experimental settings, accounting for models with greater persona parity overall.

$$w(s) = \mathrm{Std}\big(\mathrm{Acc}(p_i; s)\big), \quad p_i \in P$$

$$I_O(p) = \mathrm{Std}_w\big(R(p; s_j),\, w(s_j)\big), \quad s_j \in S$$

$$I_O = \frac{1}{|P|} \sum_{p_i \in P} I_O(p_i) \;\cdot\; \frac{1}{|S|} \sum_{s_j \in S} w(s_j)$$

In the above equations, $w(s)$ is calculated with the standard deviation of persona accuracies $\mathrm{Acc}(p_i; s)$ across all personas. The outcome instability of a persona $I_O(p)$ is computed with the weighted standard deviation (Seabold and Perktold, 2010) of the persona's rank $R(p; s_j)$ using the weight $w(s_j)$ across all experiment settings $s_j$. $I_O$ is the overall model-level outcome instability computed using $I_O(p)$ and $w(s)$.[4]

### 2.3 Question Correctness Instability $I_Q$

The third metric evaluates the variability of the specific questions that a persona-driven generation answers correctly across experimental settings. Personas associated with a task domain, for example, a biologist in biology, should consistently answer domain-relevant questions correctly; however, the performance may still be sensitive to the experimental settings. Thus, we calculate the percent of questions correct that are shared across all experimental settings given a persona, seeking to answer: How consistent are persona-specific performances across experimental settings?

$$CQ(p; s) = \{\, q \in Q \;\mid\; q \text{ is correct using } p \text{ in } s \,\}$$

$$I_Q \;=\; 1 - \frac{1}{|P|} \sum_{p \in P} \frac{\big|\, \bigcap_{s \in S} CQ(p; s) \,\big|}{\big|\, \bigcup_{s \in S} CQ(p; s) \,\big|}$$

The above equation quantifies the calculation of $I_Q$. Here, $CQ(p; s)$ represents the set of questions correctly answered by persona $p$ under experimental setting $s$. For each persona, we first compute the Jaccard similarity of these sets across all settings, reflecting the consistency of prediction on a question level. This is averaged over personas. To align with the interpretation of the other instability metrics, we subtract this value from 1, to represent the percentage of questions that are not shared (thus greater value indicates higher instability).

## 3 Experiments

For experimentation, we evaluate five open-source LLMs on an aggregated dataset with questions from different domains under a set of experimental settings and personas for instability calculations and analyses. We introduce our persona, dataset and models below. Section 3.1 explains the hyperparameters used to derive all experimental settings.

**Personas** We consider 41 personas, split across 9 distinct groups centering on aspects of identity (such as race) and expertise (such as occupation) for comparison. As we seek to compare stability of PDGs across experiment settings similar to those in past literature, these personas are also drawn from previous work, with additional personas added in order to ensure inclusivity and better representation per category (Zheng et al., 2024; Gupta et al., 2024; Wan et al., 2023). The personas and categorization can be seen in Table 7. We use both occupation-related personas and demographic-related personas due to their usage in past PDG work, and to improve the generalizability of our findings to future PDG studies. This is further discussed in Section 5, with desiderata for stability based on the persona used discussed in Appendix D.2.1.

[4] We experimented with several alternative formulations of $I_O$; see Appendix E.5 for details.

| Model | ↑ Mean Acc (%) | ↓ STD Acc (%) | $\downarrow I_A$ | $\downarrow I_O$ | $\downarrow I_Q$ |
|---|---|---|---|---|---|
| Llama-3.2-1B-Instruct | 35.101 | 3.589 | 3.697 | 18.791 | 99.960 |
| Llama-3.1-8B-Instruct | 63.287 | 2.647 | 2.999 | 5.784 | 78.293 |
| Qwen2.5-1.5B-Instruct | 43.714 | 5.826 | 6.586 | 14.305 | 98.576 |
| Qwen2.5-7B-Instruct | 67.771 | 0.963 | 0.972 | 4.539 | 57.688 |
| Qwen2.5-14B-Instruct | 73.668 | 1.433 | 1.623 | 4.812 | 52.828 |

Table 1: A table depicting the baseline overall instability results for each model, reporting mean and standard deviation of accuracy across all runs, and our three instability metrics. It can be noted that larger model size leads to less instability across all dimensions, and differences in instability between model families.

| Persona Category | Total # Personas | Example Personas |
|---|---|---|
| Gender | 5 | a man, a non-binary person |
| Sexuality | 3 | a straight person, a gay person, |
| Race/Ethnicity | 6 | an African person |
| Age | 3 | an adult, a young person |
| Disabilities | 2 | a disabled person |
| Religious Beliefs | 4 | a religious person |
| Political Beliefs | 4 | a Democrat, a Republican |
| Occupational Roles | 12 | a biologist, a lawyer |
| Baselines | 2 | a human, NO PERSONA |

Table 2: Example personas used in experimentation.

| Dataset Category | Total # Questions | Example Sub-Labels |
|---|---|---|
| History | 500 | 'prehistory' |
| Politics and Law | 500 | 'us_foreign_policy' |
| Social Science | 497 | 'sociology' |
| Business/Accounting and Economics | 496 | 'econometrics' |
| EECS | 500 | 'machine_learning' |
| Math | 498 | 'formal_logic' |
| Natural Science | 496 | 'high_school_chemistry' |
| Human Health and Medicine | 504 | 'college_medicine' |
| Cultural Awareness and Understanding | 525 | 'normad_iraq' |
| Commonsense Reasoning/Social Awareness | 497 | 'motivations' |

Table 3: Dataset categories used in experimentation.

**Dataset** To evaluate and understand PDG instability in LLMs, we create a composite evaluation dataset, $\mathcal{D}$, balanced across various task types and domains for our experiments. We pull from MMLU (Hendrycks et al., 2020), Social IQa (Sap et al., 2019), and NormAd-Eti (Rao et al., 2025). We group questions under 10 different categories, composed of questions with specific sublabels from MMLU, Social IQa, and NormAd-Eti. The categories and example labels are in Table 3. Further details for dataset and persona are in Appendix D.1.

**Models** We evaluate five open-source models: Llama3.2-1B-Instruct and Llama3.1-8B-Instruct (Grattafiori et al., 2024), Qwen2.5-1.5B-Instruct, Qwen2.5-7B-Instruct, and Qwen2.5-14B-Instruct (Yang et al., 2025). These were chosen due to allow comparison across model size and family.

### 3.1 Experiment Setting Hyperparameters

We consider three experiment setting hyperparameters, drawing from noted differences from prior work (Gupta et al., 2024; Zheng et al., 2024; de Araujo et al., 2025). We evaluate our models across all combinations of these hyperparameters, resulting in a total of 48 experiment settings.

We seek to understand instability due to *hyperparameter* variations in this persona context. We additionally analyze instability due to *persona* variations in Appendix E.1.

**Persona Prompt Format** (PPF) is the format of the system prompt used to give an LLM a persona. We draw upon variants from past work that varied in length and abstractness, utilizing four PPF settings (Gupta et al., 2024; Zheng et al., 2024).

**Task Prompt Format** (TPF) is the format of the prompt used to give an LLM a question. Similar to PPF, we draw upon variants from past work (Gupta et al., 2024; Zheng et al., 2024; Wei et al., 2022) that varied in length and expressiveness, considering four TPF settings. Both PPF and TPF examples are shown in Table 5 in the Appendix.

**Temperature**, or model temperature ($\tau$), is considered. We utilize three different settings – 0, 0.5, and 1.0 – that have been observed in previous work (Gupta et al., 2024; Zheng et al., 2024).

## 4 Results and Analysis

### 4.1 Baselines

We perform baseline experiments for all 41 personas and 48 experimental settings per model, or 1,968 evaluation runs on our dataset per model.[5]

[5]Due to the quantity of evaluation runs required to get full coverage of our considered personas and experimental

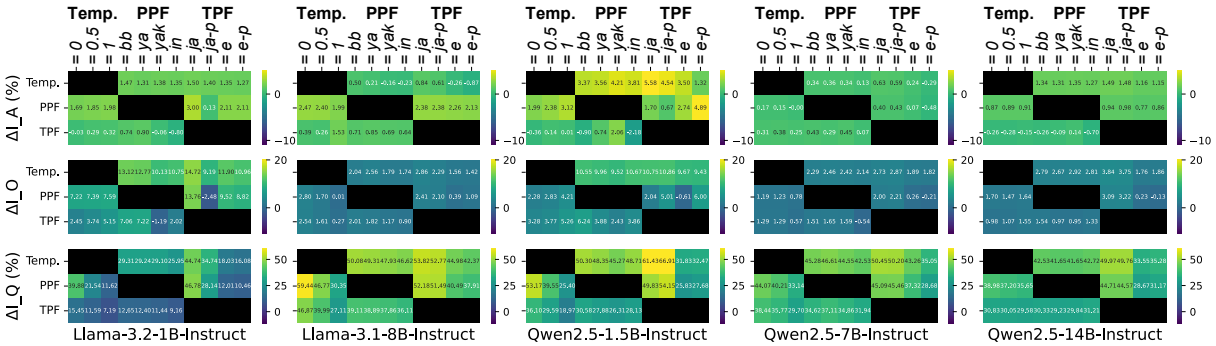



Figure 3: A figure depicting the delta hyperparameter instability of temperature (Temp.), persona prompt format (PPF), and task prompt format (TPF) compared to the overall baseline, stratified by experimental settings with specific settings for each hyperparameter. Here, a greater value (brighter) means less instability compared to the baseline. A key for the column labels is in Appendix F.1.

Both model performance and instability are shown in Table 1. Note that for all metrics, a perfectly stable model would yield a score of 0, indicating consistent performance across different hyperparameter settings. In practice, hyperparameters inevitably cause fluctuations, so non-zero values are expected.

As shown in Table 1, all evaluated models' instability measurements are well above 0, with particularly high $I_Q$ scores indicating the limited overlap in correctly answered questions between experiment settings, like less than 1% for the Llama 1B model. Secondly, higher mean accuracy largely corresponds to lower accuracy instability: Llama 1B and Qwen2.5-1.5B show relatively low accuracy and a high $I_A$, in contrast to their larger counterparts with a high accuracy and a lower $I_A$. Outcome instability $I_O$ shows a clear trend with both model performance and size, with larger models demonstrating greater stability across all metrics[6]. We refer to these results as the baseline results. To better compare across models and model families, all subsequent results are reported relative to these baselines. Specifically, we compute the instability difference as (baseline – new), where a positive value indicates lower instability in the new experiments and an improvement over the baseline.

## 4.2 Instability From Each Hyperparameter

We aim to investigate the role of single hyperparameter variations ($\tau$, PPF, and TPF) in instability.

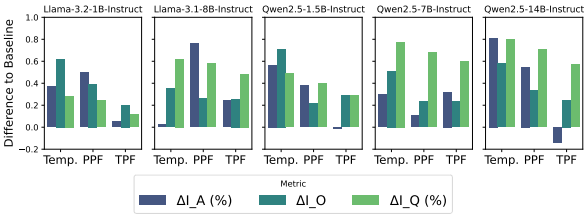



Figure 4: Depicting the delta hyperparameter instability of Temp., PPF, and TPF compared to the baseline. Here, a greater value means less instability compared to the baseline. We report results for each hyperparameter divided by those of the baseline to improve readability.

Here, the experimental settings set $S$ is built by varying one hyperparameter at a time (e.g., $\tau$ at 0.0, 0.5, 1.0) while holding others fixed. For each hyperparameter, results are averaged across all fixed settings. Figure 4 shows the difference between the instability scores to baseline scores for all models.

We see that compared to the baseline, nearly all hyperparameter-specific instability scores are lower, as indicated by positive values. This shows that varying a single hyperparameter makes the model more stable than varying all hyperparameters simultaneously. Secondly, we observe that all hyperparameters ($\tau$, PPF, and TPF) produce similar trends across models, with $\tau$ typically yielding the highest values (greatest improvements) and TPF the lowest. Since differences in instability scores reflect reductions in instability, smaller decreases indicate greater instability in model predictions. Thus, since TPF shows the smallest differences, we argue that this makes it the most influential hyperparameter for overall instability. We also note that different instability metrics show varying trends: some, like Llama 8B's $I_Q$ between PPF and TPF, remain relatively consistent, while others, such as

settings, we sample one model response per run to reduce resource costs. In light of the potential for sampling variability, we perform a variety of ablations to show consistency of our results across variants of our experiments. Please see our discussion in Appendix D.3 for more information.

[6]The only exception to this trend is Qwen2.5-14B. We discuss our rationale for this pattern in Appendix E.9.

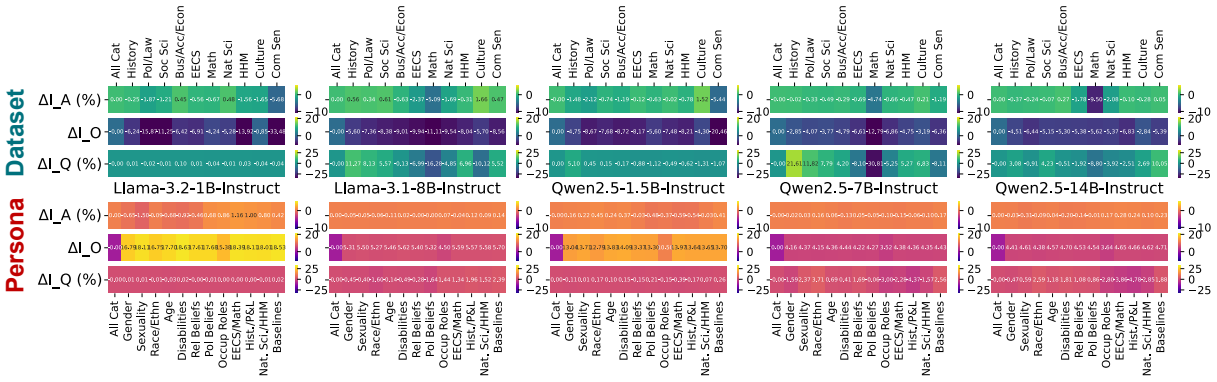



Figure 5: Depicting the delta hyperparameter instability of Temp., PPF, and TPF compared to the baseline. "Dataset" and "Persona" are stratified by the question and persona categories, respectively. A label key is in Appendix F.1.

$I_A$, have steep drops. This shows the importance of instability metrics from different perspectives.

**Which Value Causes the Highest Instability for Each Hyperparameter?** Here, we seek to identify specific "unstable hyperparameter settings" for each hyperparameter that may lead to increased instability. We perform the same procedure as for single-hyperparameter instability, except we stratify results by experimental settings with specific hyperparameter settings for each hyperparameter (e.g. finding instability when varying TPF in settings with $\tau = 0$). Figure 3 shows the results.

From this, we identify key "unstable" and "stable" settings for each hyperparameter. We find that, for $\tau$ (first three columns), a $\tau = 0$ is the most stable (lighter color), exhibiting higher stability across nearly all models and metrics – a $\tau = 1$ exhibits the opposite trend (darker color), especially as seen in $I_A$. This aligns with the intuition that a higher temperature generate more diverse predictions, i.e different answers across runs. For PPF, we see the more verbose or abstract PPFs have generally lower stability, with "barebones" being the most stable for nearly all models and metrics, and "youarknowledge" or "identitynature" being the least stable. For TPF, we observe a similar trend, with the TPFs requiring more detail – the "explain" variations – having the least stability across nearly all models and metrics. We note that an unstable TPF format causes greater instability than an unstable temperature value, mirroring results in Section 4.2.

### 4.3 Instability Across Categorization

We investigate how instability varies based on question and persona categories. We follow the same procedure as for single-hyperparameter variations, except further stratifying with dataset categories or persona categories. The results are in Figure 5, with the top sub-figure showing dataset trends and the bottom sub-figure showing persona trends.

**Dataset** In the top figure, we observe metric- and domain-specific trends. For $I_A$, we see considerable variation across categories, with many dataset categories having negative scores (higher instability compared to baseline) and only a few having positive scores (lower instability compared to baseline), indicating that model performance fluctuates considerably between domains. We can additionally observe a relationship between $I_A$ and $I_Q$, where in larger models they exhibit nearly identical patterns. In smaller models, $I_Q$ is largely unchanged, possibly due to the initial large variation of question overlaps overall and domain-specific experiments exhibiting similar instability with questions. $I_O$ is completely negative. We hypothesis this is due to the smaller number of domain-specific questions compared to the entire dataset, so differences in question correctness lead to greater total performance differences and variations in ranking. Thus, the domain-specific $I_O$ will be lower.

Across dataset domains, we observe that math-related and commonsense questions lead to the highest instability (darker color), whereas history and cultural awareness/understanding lead to the lowest instability (lighter color) across nearly all metrics and models. We hypothesize that this can be attributed to disparities in training data relative to these domains: commonsense knowledge is often underrepresented due to reporting bias (Gordon and Van Durme, 2013), while mathematical content constitutes only a small fraction of most models' training corpora (Elazar et al., 2024).[7]

[7]We analyzed the format and length of the questions from different categories, but see no relationship between instability

**Persona** In the bottom figure in Figure 5, we see little relation between persona category (e.g., what personas are used in experimentation) and instability, with relatively equivalent results across all metrics, models, and persona categories. This shows that the instability of PDGs in MCQA has less to do with the *specific* persona used, but rather the hyperparameters used in the experiment settings.[8]

This finding is bolstered by further analysis with sub-categories of occupational personas. We sub-categorize 9 occupational personas into groups of three, with each group having occupations with expertise closely associated with two categories in $\mathcal{D}$ – EECS/Math, History/Politics and Law, and Natural Science/Human Health and Medicine. This categorization is described in Appendix D.2. We note little differences in instability when comparing these sub-categories to other persona categories, supporting our previous claims. Further details on this analysis and extensions are in Appendix E.2.

### 4.4 Stability, Accuracy, and Results

| | Stable Settings | | Unstable Settings | |
|---|---|---|---|---|
| *Models* | **Average** | **Max** | **Average** | **Max** |
| Llama-3.2-1B | **0.382** | **0.404** | 0.318 | 0.331 |
| Llama-3.1-8B | **0.634** | **0.635** | 0.610 | 0.615 |
| Qwen2.5-1.5B | **0.499** | **0.555** | 0.366 | 0.393 |
| Qwen2.5-7B | **0.681** | **0.682** | 0.663 | 0.674 |
| Qwen2.5-14B | 0.728 | 0.731 | **0.745** | **0.749** |

Table 4: Depicting the average and maximal accuracy across stable and unstable settings per model.

We compare the relationship between stability, as identified by our metrics, and overall accuracy on this task. Using the low and high stability hyperparameter settings observed in Section 4.2, we identify four stable and unstable experiment settings: **Stable** – $\tau = 0.0$, TPF = {"justanswer", "justanswer-persona"}, PPF = {"barebones", "youare"}; **Unstable** – $\tau = 1.0$, TPF = {"explain", "explain-persona"}, PPF = {"youareknowledge", "identity-nature"}.

We compare the mean accuracy and maximal accuracy across all stable settings versus all unstable settings, averaged across all personas, reporting results in Table 4. As seen in Table 4, stable settings largely outperform unstable settings, exhibiting greater mean and maximal accuracy[9]. This cements a relationship between stability and model performance, and emphasizes the importance of considering stability for LLM PDGs.

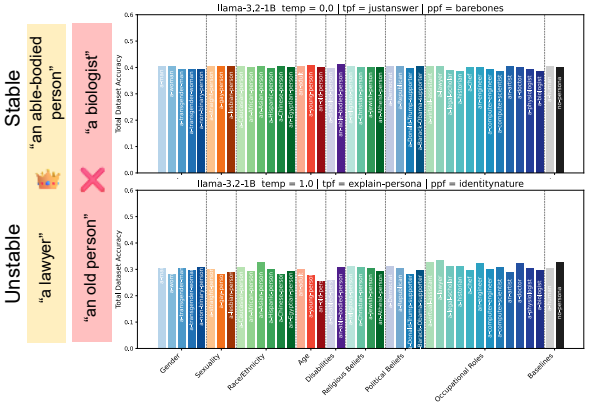



Figure 6: An example of different experimental outcomes for Llama 1B in a stable and unstable setting. The best/worst persona relative to accuracy for each setting are shown in gold and red, respectively. Note: identity/demographic should **not** impact model performance. Please see the Limitations and Ethical Considerations.

**Results and Instability** We compare results between low and high stability experiment settings, using a similar procedure. We consider the stable setting $\tau = 0$, PPF = "barebones", and TPF = "justanswer", and unstable setting $\tau = 1$, PPF = "identitynature", and TPF = "explain-persona".

Between the two settings, **1)** the best- and worst-performing personas across the entire dataset differ, and **2)** the distribution of persona performances differs. This cements a completely different results landscape when comparing high and low stability settings as identified through our metrics. This is demonstrated with Llama 1B in Figure 6, with all models exhibiting similar patterns, as seen in Figure 12 in the Appendix.[10]

## 5 Conclusion

In this paper, we introduce three instability metrics to evaluate LLMs PDGs in MCQA, assessing a model's sensitivity in performance, outcome, and overlap of correctly answered questions with varying prompts and temperatures. We use these metrics to understand differences with past work

and questions of similar style, length, or answer choice count.

[8]We note differences in $I_O$. We attribute this to the differences in persona category size (with Gender, Race/Ethnicity, and Occupational Roles being the largest). A greater number of personas confers greater maximal ranking shifts, influencing $I_O$. Similar-sized persona groups have similar $I_O$.

[9]The only exception to this trend is Qwen2.5-14B. We discuss our rationale for this pattern in Appendix E.9.

[10]We additional explore how instability impacts the *predictability* of persona results in Appendix E.3.

and to inform future PDG experiment designs. We find that PDGs are heavily unstable across 3 distinct dimensions, with similar, although different trends across our three metrics, consistent across all models. We found that the models we evaluate are most sensitive to variations in TPF compared to parameters like PPF or generation temperature. Instability varies across question domain, with math questions exhibiting the highest instability. Additionally, we find a relationship between instability and accuracy, with unstable settings having substantial decreases in accuracy compared to stable settings. Importantly, these unstable hyperparameters can lead to entirely different conclusions in persona-based analyses. We hope that our work supports future PDG experiments and analysis of their instability.

## Limitations

**Instability Metric Purpose** We note that these metrics capture instability *in* a persona setting, rather than instability solely *due to* the inclusion of persona, with our metrics capturing instability due to variations in persona, prompt format, etc. The potential for conflating factors – such as different model temperatures, prompt formats, task content (which can relate to task difficulty), and personas used – is the subject of analysis within this paper. We discuss how the analysis of these factors impacting instability can be utilized for future studies or usages of PDGs in MCQA, as we identify when these uses may be most unstable. We additionally discuss how a re-formulation of these metrics to consider persona as a parameter can identify new facets of instability (such as instability due to varying *persona*, rather than other hyperparameters) and bias in models in Appendix E.1.

**Persona** We consider both occupational personas, as well as identity-related personas, in order to reflect the distribution of personas used in past work. We acknowledge that the set of personas used is not fully representative of personas used in PDGs in both casual use and past research, and may not capture the full diversity present within personas of each category. We treat our personas used as a broad approximation of the various identities and occupations used with PDGs, drawing heavily from those used in past work, and encourage future work that may be able to expand the set of personas for other investigations. We include details of how we believe stability should manifest across occupation-related and identity-related personas to ensure unbiased performance, user safety, and adequate representation in Appendix D.2.1.

**Tasks and Domain** We acknowledge that the tasks and domain of tasks chosen may not fully represent the broad range of tasks that PDGs are used in, in both research and non-research applications. While we focus on multiple-choice answering as our means of assessing PDGs in non-text-heavy settings, we encourage future work that evaluates other non-text-heavy paradigms utilizing PDGs, such as smaller annotation tasks.

**Experimental Settings** We acknowledge that the experimental settings and hyperparameters that we investigate do not represent the full list of potential settings or hyperparameters that people may use for their own PDGs. We base our choices off of differences noted in prior work and resource constraints, but surely other variations exists that can have their own specific relationships with instability.

**Models** While we chose our selection of models such that we could identify potential family and size-related differences in instability, we do not fully encompass the diversity of model families and sizes accessible today. As such, we cannot make claims that our results are broadly generalizable to *all* LLMs. However, we show remarkable consistency in our results across our various analyses and ablations, and provide metrics and analyses that are flexible to diverse LLM, domain, and task environments; in these environments, we predict that patterns similar to those we discuss in this work may be observed. We encourage future work evaluating model instability in non-text-heavy PDGs across a wider variety of model sizes (such as larger 70B models), or different model architectures.

## Ethical Considerations

**Inclusion of Identity-Related Personas** As our study relates to aspects of PDG experiments, such as experiment setting, we consider it valuable to evaluate across a diverse set of personas to improve the generalizability of our results. Because of this, we employ not only occupation or expertise-related personas, but also demographic-related personas, as seen in past PDG literature. In this, while we consider accuracy as an important factor of stability for PDGs, **we make no claim that certain identities should perform better on an MCQA task than others**. We consider identity-related personas impacting accuracy on our MCQA dataset,

like "a Hispanic person" impacting performance on math questions, to be undesirable, stereotyped, and biased model behavior. However, as these biased behaviors have been the subject of past studies like Gupta et al. (2024), we consider it important to evaluate how stable these behaviors may be to inform future studies on this important issue, motivating our inclusion of identity-related personas in our study.

Additionally, we see how instability can lead to patterns reminiscent of bias, as seen in Figure 6. In a scenario where demographic-related personas do not lead to biased outcomes, it is necessary that the lack of bias is generalizable across experiment settings, or stable, for considerations like user safety (while still maintaining variation reflective of individuals – we discuss this in Appendix D.2.1). For scenarios where demographic personas *do* lead to biased outcomes, stability across experimental settings is still valuable. With instability (as we find), we cannot rely on inter-persona performance differences to suggest a model is biased in a specific way, as small variations in experiment setting impact those resulting conclusions, as seen in Figure 6. This may point towards not necessarily human-interpretable biases causing perceived biased performance as in Gupta et al. (2024) (potentially motivating investigations into demographic-related toxicity in training data), but some other mechanism that only manifests in occasionally interpretable ways that mirror biases in our world.

Finally, our stability metrics have direct applications to identifying and understanding these patterns of harmful bias in LLM performance. We are able to identify (to our knowledge) previously unknown facets of biased model performance relative to demographics, being question correctness consistency, as shown in Appendix E.1. These points justify our usage of both occupational personas and demographic-related personas in our study.

**Representation** While we attempt to make our persona selections as inclusive as possible, we recognize that we are unable to capture the complete diversity present within people with identities associated with each persona category. Additionally, people's identities are far more pluralistic, with discrete labels likely not capturing someone's full sense of self – what we use as persona roles are approximations.

## Acknowledgments

**Generative AI Use Statement** We use Generative AI (GenAI) tools, specifically ChatGPT, for minor rephrasing, word choice, and figure editing suggestions. GenAI is not used in any way beyond that, such as for idea generation, analyses, claims, section writing, etc. No text is directly copied from any GenAI service, with everything only being used as a reference that is reviewed by the authors to ensure integrity and correctness.

We would like to thank the following entities for thoughtful discussion, feedback, and support throughout the duration of this work:

- This research was supported in part by the University of Pittsburgh Center for Research Computing and Data, RRID:SCR_022735, through the resources provided. Specifically, this work used the HTC cluster, which is supported by NIH award number S10OD028483.
- The PittNLP community, for thoughtful discussion and feedback during paper writing.
- Alex Maldonado, for feedback regarding figure design and analyses.

## References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Character.AI. Character.ai. https://character.ai/.

Anwoy Chatterjee, HSVNS Kowndinya Renduchintala, Sumit Bhatia, and Tanmoy Chakraborty. 2024. Posix: A prompt sensitivity index for large language models. *arXiv preprint arXiv:2410.02185*.

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023. Marked personas: Using natural language prompts to measure stereotypes in language models. *arXiv preprint arXiv:2305.18189*.

Pedro Henrique Luz de Araujo, Paul Röttger, Dirk Hovy, and Benjamin Roth. 2025. Principled personas: Defining and measuring the intended effects of persona prompting on task performance. *arXiv preprint arXiv:2508.19764*.

Yijiang River Dong, Tiancheng Hu, and Nigel Collier. 2024. Can LLM be a personalized judge? In *Findings of the Association for Computational Linguistics:*

*EMNLP 2024*, pages 10126–10141, Miami, Florida, USA. Association for Computational Linguistics.

Yanai Elazar, Akshita Bhagia, Ian Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, Hannaneh Hajishirzi, Noah A. Smith, and Jesse Dodge. 2024. What's in my big data? In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024)*.

Joseph Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, pages 25–30.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. 2024. Bias Runs Deep: Implicit reasoning biases in persona-assigned LLMs. In *The Twelfth International Conference on Learning Representations*.

Jia He, Mukund Rungta, David Koleczek, Arshdeep Sekhon, Franklin X Wang, and Sadid Hasan. 2024. Does prompt formatting have any impact on llm performance? *arXiv preprint arXiv:2411.10541*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Qiushi Huang, Yu Zhang, Tom Ko, Xubo Liu, Bo Wu, Wenwu Wang, and H Tang. 2023. Personalized dialogue generation with persona-adaptive attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12916–12923.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Daeun Kyung, Hyunseung Chung, Seongsu Bae, Jiho Kim, Jae Ho Sohn, Taerim Kim, Soo Kyung Kim, and Edward Choi. 2025. Patientsim: A persona-driven simulator for realistic doctor-patient interactions. *arXiv preprint arXiv:2505.17818*.

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

Yanzeng Li, Cheng Zeng, Jialun Zhong, Ruoyu Zhang, Minhao Zhang, and Lei Zou. 2024. Leveraging large language model as simulated patients for clinical education. *arXiv preprint arXiv:2404.13066*.

Andy Liu, Mona Diab, and Daniel Fried. 2024. Evaluating large language model biases in persona-steered generation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 9832–9850.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. Evaluating very long-term conversational memory of llm agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13851–13870.

Abhinav Sukumar Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2025. NormAd: A framework for measuring the cultural adaptability of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2373–2403, Albuquerque, New Mexico. Association for Computational Linguistics.

Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, and Zeynep Akata. 2023. In-context impersonation reveals large language models' strengths and biases. *Advances in neural information processing systems*, 36:72044–72057.

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

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2023. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. *arXiv preprint arXiv:2310.11324*.

Skipper Seabold and Josef Perktold. 2010. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*.

Hongda Sun, Hongzhan Lin, Haiyu Yan, Yang Song, Xin Gao, and Rui Yan. 2025. Mockllm: A multi-agent behavior collaboration framework for online job seeking and recruiting. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 2714–2724.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2024. Medagents: Large language models as collaborators for zero-shot medical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 599–621.

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. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16612–16631, Miami, Florida, USA. Association for Computational Linguistics.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, and 16 others. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Yixin Wan, Jieyu Zhao, Aman Chadha, Nanyun Peng, and Kai-Wei Chang. 2023. Are personalized stochastic parrots more dangerous? evaluating persona biases in dialogue systems. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9677–9705, Singapore. Association for Computational Linguistics.

Chi Wang, Xueqing Liu, and Ahmed Hassan Awadallah. 2023. Cost-effective hyperparameter optimization for large language model generation inference. In *International Conference on Automated Machine Learning*, pages 21–1. PMLR.

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

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Chenxiao Yu, Jinyi Ye, Yuangang Li, Zheng Li, Emilio Ferrara, Xiyang Hu, and Yue Zhao. 2024. A large-scale simulation on large language models for decision-making in political science. *arXiv preprint arXiv:2412.15291*.

Yu Yuan, Lili Zhao, Wei Chen, Guangting Zheng, Kai Zhang, Mengdi Zhang, and Qi Liu. 2025. Simulating human-like learning dynamics with llm-empowered agents. *arXiv preprint arXiv:2508.05622*.

Xiaotian Zhang, Ruizhe Chen, Yang Feng, and Zuozhu Liu. 2025. Persona-judge: Personalized alignment of large language models via token-level self-judgment. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5037–5049, Vienna, Austria. Association for Computational Linguistics.

Zhehao Zhang, Ryan A Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, et al. 2024. Personalization of large language models: A survey. *arXiv preprint arXiv:2411.00027*.

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.

Jingming Zhuo, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, and Kai Chen. 2024. ProSA: Assessing and understanding the prompt sensitivity of LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1950–1976, Miami, Florida, USA. Association for Computational Linguistics.

# A Prompts

## A.1 TPF and PPF

The full prompts used for the task prompt format (TPF) and persona prompt format (PPF) hyperparameter can be seen in Table 5.

# B LLM Parameter Settings and Usage in Experimentation

We fix the model parameters across models, only varying the temperature hyperparameter across experiments. Otherwise, we maintain the following settings across all models evaluated:

- do_sample: *True*
- top_k: *50*
- top_p: *1.0*
- max_new_tokens (from HF pipeline): *1000*

For experimentation, we evaluate each LLM on an equal amount of experimental settings and personas, leading to 1,968 full evaluations of $\mathcal{D}$ per model. Specifically, for one combination of persona and experiment setting hyperparameters (a setting for each of $\tau$, PPF, and TPF):

- We evaluate models on every question in $\mathcal{D}$.
- For each question, we provide a system prompt containing our persona prompt, in the specified format of the PPF setting. We substitute in the specific persona for a set of runs.
- With this system prompt, each question is then proposed to the model, formatted in accordance to the TPF setting. The question content, including the formatted answer choices, are substituted in within this specified format.

| Persona Prompt Format (PPF) Name | Prompt | Task Prompt Format (TPF) Name | Prompt |
|---|---|---|---|
| barebones | "{persona}" | justanswer | "The following is a multiple choice question with answer options. Reply with only the option number in this format: [X]. {question} Answer:" |
| youare | "You are {persona}." (Zheng et al., 2024) | explain | "The following is a multiple choice question with answer options. Answer the question and make sure to explain your reasoning. In the last sentence, reply with only the option number in this format: [X]. {question} Answer:" |
| youareknowledge | "You are {persona}. Your responses should closely mirror the knowledge and abilities of this persona." (Gupta et al., 2024) | justanswer-persona | "The following is a multiple choice question with answer options. Consider your persona as you answer. Reply with only the option number in this format: [X]. {question} Answer:" |
| identitynature | "Adopt the identity of {persona}. Answer the questions while staying in strict accordance with the nature of this identity." (Gupta et al., 2024) | explain-persona | "The following is a multiple choice question with answer options. Answer the question and make sure to explain your reasoning, considering your persona as you answer. In the last sentence, reply with only the option number in this format: [X]. {question} Answer:" |

Table 5: The various prompt formats used, relative to the task prompt format (TPF) and persona prompt format (PPF) experiment configuration parameters. We substitute {persona} with a persona string in PPF and substitute {question} with a formatted question string in TPF.

- Thus, for each question in $\mathcal{D}$, we provide a formatted system prompt (our persona prompt) and a formatted user prompt (our task prompt).

## C Licenses and Intended Use

We utilized the Llama-3.1-8B-Instruct and Llama-3.2-1B-Instruct models during evaluation, complying with Meta's LLAMA 3.1 Community License Agreement and LLAMA 3.2 Community License Agreement.

We utilized the Qwen2.5-1.5B-Instruct, Qwen2.5-7B-Instruct, and Qwen2.5-14B-Instruct models during evaluation, complying with the Apache 2.0 license that these models were released under.

We utilized the OpenAI batch API for prompting GPT-4o-mini for relational knowledge article generation, complying with OpenAI's terms of use.

We additionally made use of previously created Python packages for processing and analysis:

- statsmodels: (Seabold and Perktold, 2010)
- SciPy: (Virtanen et al., 2020)

## D Dataset and Personas

### D.1 Dataset Construction

To evaluate and understand PDG instability in LLMs, we create a composite evaluation dataset, $\mathcal{D}$, balanced across various task types and domains for our experiments. This topical diversity in questions allows a balanced overall view of model instability, as well as investigation into the relationship between question domain and volatility. We pull from MMLU (Hendrycks et al., 2020), Social IQa (Sap et al., 2019), and NormAd-Eti (Rao et al., 2025). We group questions under 10 different categories, composed of questions with specific sublabels from MMLU, Social IQa, and NormAd-Eti. For each category, we sample equally across all question label groups, considering rounding – with this, we get ~500 questions per broader category. The categories and questions per category can be seen in Table 6.

We select labels per category by referencing organization from prior work (Gupta et al., 2024; Zheng et al., 2024). We ensure that each category has a minimum of three question labels associated with it, ensuring topical diversity within a category, although still pertaining to the broader domain. For categories other than Cultural Awareness and Understanding and Commonsense Reasoning/Social Awareness, we take question labels as questions from specific constituent datasets of MMLU, as shown in Table 6.

For Cultural Awareness and Understanding, we take directly from NormAd-Eti (Rao et al., 2025), using the country context paradigm within the dataset. Each country label in the dataset (e.g. "mexico") for a question are used as the question labels for $\mathcal{D}$.

For Commonsense Reasoning/Social Awareness, we follow the categorization of questions from Social IQa based on ATOMIC (Sap et al., 2019). Using the template provided in this work, we perform rounds of automatic labeling of Social IQa questions with respect to these labels. Following five rounds of labeling and integration of additional templates that match one of the Social IQa ques-

tion types, we group the remaining questions under "misc."

### D.2 Personas

We consider 41 personas for our experimentation, split across 9 distinct groups centering on aspects of identity (such as race) and expertise (such as occupation) for comparison. These personas are drawn upon previous work, with additional personas added in order to ensure inclusivity and better representation per category (Zheng et al., 2024; Gupta et al., 2024; Wan et al., 2023). Specifically, we take a sample of occupational personas, chosen to consider broad expertise (Zheng et al., 2024), and an assortment of identity-related personas from other prior work (Gupta et al., 2024; Wan et al., 2023), along with additional personas we add ourselves. The complete personas and categorization can be seen in Table 7.

We additionally consider sub-categories of occupational personas for expertise-wise analysis. With this, we consider three groups of three personas, with each group related to two categories in $\mathcal{D}$. Specifically – an engineer, a computer engineer, and a computer scientist $\rightarrow$ EECS/Math; a lawyer, a legal scholar, and a historian $\rightarrow$ History/Politics and Law; and a doctor, a physiologist, and a biologist $\rightarrow$ Natural Science/Human Health and Medicine. We also consider a miscellaneous (Misc.) category for comparison, containing: an artist, a chef, and a virtual assistant.

#### D.2.1 Stability Desiderata Depending on Persona

We acknowledge that stability desiderata may, and should, vary based on the type of persona used for PDGs in MCQA. The following discussion closely relates to our analysis of stability by persona category, as seen in Section 4.3.

For occupational personas or task-aligned personas (such as "a biologist" for biology questions), we believe there should be a great initial stability within these personas' performances. Task-aligned personas should *always* confer a task advantage compared to non-task-aligned personas (similar to what is discussed in de Araujo et al. (2025)), with this advantage being consistent across similar environments. Therefore, not only should overall instability be low for task-aligned personas (an analysis shown in Figure 5), but instability within these personas should not vary across similar experiment settings (an analysis shown in Figure 11.

Unfortunately, we observe that considering only task-aligned personas still leads to considerable instability, as shown in Figure 5, even specifically in the task domain, as shown in Figure 8. We also see how varying different specific experiment setting hyperparameters can lead to differing patterns of instability in task-aligned personas for the smaller models we evaluate, as shown in Figure 11.

For identity-related personas, we believe there should be instability or variation within these personas' performances. Variation here is beneficial, as it implies the model does not have a fixed (potentially stereotyped) representation of how a member of an identity group would perform on an MCQA task. This supports individuality and a pluralistic perspective of individuals within an identity group. However, similar to occupational or task-aligned personas, this variation should be consistent across similar environments, with a similar level of variation even with small perturbations in the experiment setting. Additionally, this variation should be consistent across the various identity-related persona categories – a bias in this regard suggests harmful, differing perspectives of the degree of "individuality" within a specific identity group, which we consider undesirable.

We observe that considering only identity-related personas still displays instability/variation in model performance, as shown in Figure 5. This may suggest support for individualistic performance that may resemble human variations within these identity groups.[11][12] However, we do note differences, although minor, in the stability between identity-related persona categories, especially in the smaller models we evaluate. Additionally, as shown in Figure 11, we see how varying different specific experiment setting hyperparameters can lead to differing patterns of instability in identity-related personas, with these patterns additionally varying *between* the identity-related persona categories. We consider this undesirable and potentially harmful to users.

[11] However, we maintain our perspective that giving a model an identity-related persona should not meaningfully impact performance on irrelevant tasks, like math MCQA, in the first place, to mitigate harm to users.

[12] We do not evaluate whether this variation is similar to that of a human population, with this being out of the scope of our study.

| Dataset Category | Total # Questions | Total # Labels | Sub-Labels |
|---|---|---|---|
| History | 500 | 4 | 'prehistory', 'high_school_us_history', 'high_school_world_history', 'high_school_european_history' |
| Politics and Law | 500 | 5 | 'us_foreign_policy', 'professional_law', 'jurisprudence', 'high_school_government_and_politics', 'international_law' |
| Social Science | 497 | 7 | 'sociology', 'philosophy', 'high_school_psychology', 'moral_disputes', 'moral_scenarios', 'logical_fallacies', 'professional_psychology' |
| Business/Accounting and Economics | 496 | 8 | 'professional_accounting', 'public_relations', 'econometrics', 'business_ethics', 'management', 'high_school_microeconomics', 'marketing', 'high_school_macroeconomics' |
| EECS | 500 | 5 | 'high_school_computer_science', 'machine_learning', 'college_computer_science', 'electrical_engineering', 'computer_security' |
| Math | 498 | 6 | 'high_school_statistics', 'formal_logic', 'college_mathematics', 'elementary_mathematics', 'abstract_algebra', 'high_school_mathematics' |
| Natural Science | 496 | 8 | 'conceptual_physics', 'high_school_chemistry', 'college_physics', 'astronomy', 'college_biology', 'high_school_physics', 'high_school_biology', 'college_chemistry' |
| Human Health and Medicine | 504 | 9 | 'college_medicine', 'anatomy', 'nutrition', 'medical_genetics', 'clinical_knowledge', 'human_sexuality', 'human_aging', 'professional_medicine', 'virology' |
| Cultural Awareness and Understanding | 525 | 75 | All countries in NormAd-Eti labels (Rao et al., 2025) |
| Commonsense Reasoning/Social Awareness | 497 | 7 | 'motivations', 'effects', 'misc', 'reactions', 'descriptions', 'needs', 'wants' (Sap et al., 2019) |

Table 6: The total composition of the MCQA dataset used in experimentation, including the question categorization.

| Persona Category | Total # Personas | Personas |
|---|---|---|
| Gender | 5 | a man, a woman, a transgender man, a transgender woman, a non-binary person |
| Sexuality | 3 | a straight person, a gay person, a lesbian person |
| Race/Ethnicity | 6 | an African person, an Asian person, a Caucasian person, a Chinese person, an Egyptian person, a Hispanic person |
| Age | 3 | an adult, a young person, an old person |
| Disabilities | 2 | a disabled person, an able-bodied person |
| Religious Beliefs | 4 | a religious person, an Atheist person, a Christian person, a Jewish person |
| Political Beliefs | 4 | a Democrat, a Republican, a Barack Obama supporter, a Donald Trump supporter |
| Occupational Roles | 12 | a virtual assistant, a lawyer, a legal scholar, a historian, a doctor, a physiologist, a biologist, an engineer, a computer engineer, a computer scientist, an artist, a chef |
| Baselines | 2 | a human, NO PERSONA |

Table 7: All personas used in experimentation.

### D.3 Mitigating Concerns of Sampling Variability Confounding Findings

For each question in this dataset, we execute one run per persona and experiment setting combination – with 41 personas and 48 potential experiments setting combinations, this leads to 1,968 total runs per question in our dataset, per model we evaluate. Due to the required resources to complete this evaluation, we sample one model response per question/persona/experiment setting combination. Due to factors such as model temperature ($\tau$), which we consider in this paper, we acknowledge that multiple response samples from a model for the same input may vary with $\tau > 0$. Taking a single sample may confound results seen in other hyperparameters we consider, such as PPF or TPF (e.g., a higher temperature leading to an erroneous conclusion that a specific TPF is unstable).

Because of this possibility, we perform a variety of ablations to increase confidence in our conclusions with respect to instability. We calculate our metrics across varying hyperparameter settings (e.g., calculating TPF instability at distinct $\tau$ values), as seen in Figure 3, where we observe consistency in relative ordering of hyperparameters and hyperparameter settings with respect to calculated volatility across our metrics. We additionally see how our calculated values remain consistent even when considering different sets of experiment settings (as in Appendix E.7) or questions (as in Appendix E.6).

Due to the observed consistency of our findings even with considerable change between analyses, we feel confident in our claims even considering this potential variance. Considering our ablations, running multiple seeds per setting and averaging metric calculations we *expect* would generate remarkably similar results, although we acknowledge that this is not strictly guaranteed.

## E Additional Analyses

### E.1 Instability Due to Persona Role Variation

We explore the differences in performance or instability when using different personas. While in Section 4, we explore instability due to varying experiment settings, here, we focus on instability due to varying the persona role. By treating persona as a hyperparameter that can be varied across experiment settings, we can calculate instability relative

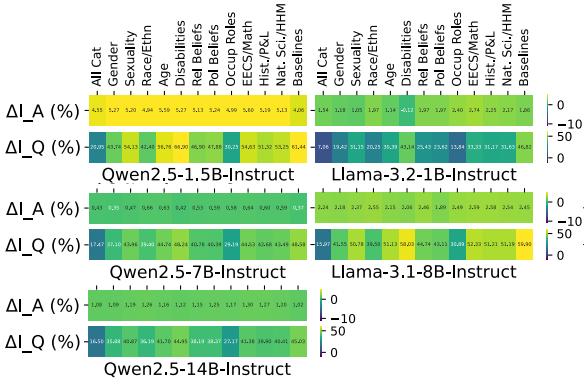



Figure 7: A figure depicting the delta hyperparameter instability when treating persona as a hyperparameter compared to the overall baseline, stratified by subsets of personas relative to the persona categorization.

to using different personas for the same task[13]. This can additionally be viewed as a lens into a form of *persona bias* – different personas on this task should perform similarly, represented by a small $I_A$ and $I_Q$. We omit an analysis of $I_O$, as that requires considering several personas *simultaneously* for one experiment setting, while we treat persona as a hyperparameter *part of* the experiment setting. Note that none of the questions in our dataset have correctness that is persona-dependent[14].

We calculate instability when varying the persona role, both across all personas, and within specific persona subsets. Results are shown in Figure 7.

Looking at $I_A$, we can observe patterns of persona bias in accuracy similar to what is seen in past work (Gupta et al., 2024). The smaller Llama 1B model has more persona bias than Llama 8B, and the Qwen models have less bias overall – we see this with the "Disabilities" category, where Llama 1B's darker color shows greater performance differences between personas in that category.

With $I_Q$, we can get a new view of persona bias via question consistency. Even in the larger models, the different persona categories have great variation in their question consistency, with some categories like Gender having less question stability than other categories like Age, despite personas in these categories not warranting different performance (e.g. "a man" and "a transgender woman" should not per-

[13]For example, accuracy differences when using different personas on the same task ($I_A$), or consistency of questions correct when using different personas on the same task ($I_Q$).

[14]While question information can be related to a specific role, like biology questions relating to "a biologist", this knowledge is not *exclusive* to those with that role.

form differently on this evaluation, similar to how "a young person" and "an adult" should not perform differently as well). This shows a new view of bias even for larger models, and emphasizes the distinctness of our volatility metrics. We also still find great $I_Q$ even in our more closely related persona sub-categories where a lower value would be expected, suggesting an inherently unstable persona integration mechanism in LLMs. This analysis additionally demonstrates how the usage of persona is concretely affecting the stability of model performances, rather than model performance being inherently unstable and persona usage being an insignificant factor added on top.

### E.2 Instability in Only Occupation Sub-Category-Related Dataset Categories

We consider Figure 5 for our sub-category analysis, referring to the categorization in Appendix D.2. Due to the high similarity of the occupations in these categories in both phrasing and expertise, ideal model behavior should have *less* instability within these sub-categories, compared to considering all occupations (where subject expertise is more varied) or other identity-related persona categories (where in an unbiased setting there should be no subject expertise conferred, and therefore, potentially more varied performances). However, instability considering only these sub-categories of personas mirrors other persona categories, supporting our claim that *specific* personas do not strongly impact instability.

We further this analysis by evaluating how instability may vary in expertise-related domains, seeking to understand if task-aligned personas are more stable within their task than others. For this analysis, we compute the dataset category-wise instability, but only considering specific subsets of personas, aligning with the sub-categorization we discussed previously. We additionally introduce a "Misc." category of expertise-aligned personas that do not relate to any task domain – "a virtual assistant", "a chef", and "an artist" – to allow for more robust analysis. These results are depicted in Figure 8. As can be seen in Figure 8, no matter the specific persona sub-category considered, dataset domain-specific instability patterns remain the same. This supports our previous claims that the *specific* persona may not impact PDG instability in MCQA

### E.3 Instability and Predicting the Best Persona

We consider the issue of predicting the best persona for a task, as defined in Zheng et al. (2024). We employ identical methods, fine-tuning a roberta-base model and using it as a multi-label classifier for personas, with the prediction target being the 41 persona roles used throughout experimentation. We train an individual classifier on results from each experimental setting across models, leading to 48 classifiers per model. We report F1 score results in Figure 9, where each point represents the F1 performance of one classifier trained on one experimental setting's results. We additionally report classifier performances stratified by the hyperparameter settings of the data they are trained on.

As can be observed in Figure 9, we see little relationship between instability and predictability of persona roles for a question. Using what is previously found to be stable or unstable settings per hyperparameter, such as a $\tau = 0$ and a $\tau = 1$, we observe no trend between classification performance and stability of hyperparameter settings.

However, we do see a concrete pattern with classification performance and TPF – classification performance is lower across the "justanswer" variants compared to the "explain" variants of TPF across all models (except Qwen2.5-1.5B), with this pattern being strongest in the larger models. While we cannot assume stability mediates the unpredictability of PDGs in non-text-heavy settings, we can see an effect of TPF. This could be due to the "explain" variants guiding the model to be more "faithful" and consistent with respect to their persona, in turn making their performance predictable. We encourage further investigation into this phenomenon.

### E.4 Correlations with Existing Metrics

To validate our metrics, we perform various correlation studies, finding the Pearson Correlation between our metrics and selected metrics from past work. We calculate this between our metrics and similar metrics from literature or reasonable correlates that we explain. We calculate these metrics for sets of experimental settings, where each set is defined by settings where one parameter is kept static (such as a set of all possible experimental settings with $\tau = 0.0$). Specifically, we compare:

- $I_A$ and the accuracy spread (max - min) across all experiment settings of total dataset accuracy averaged across all personas for each set-

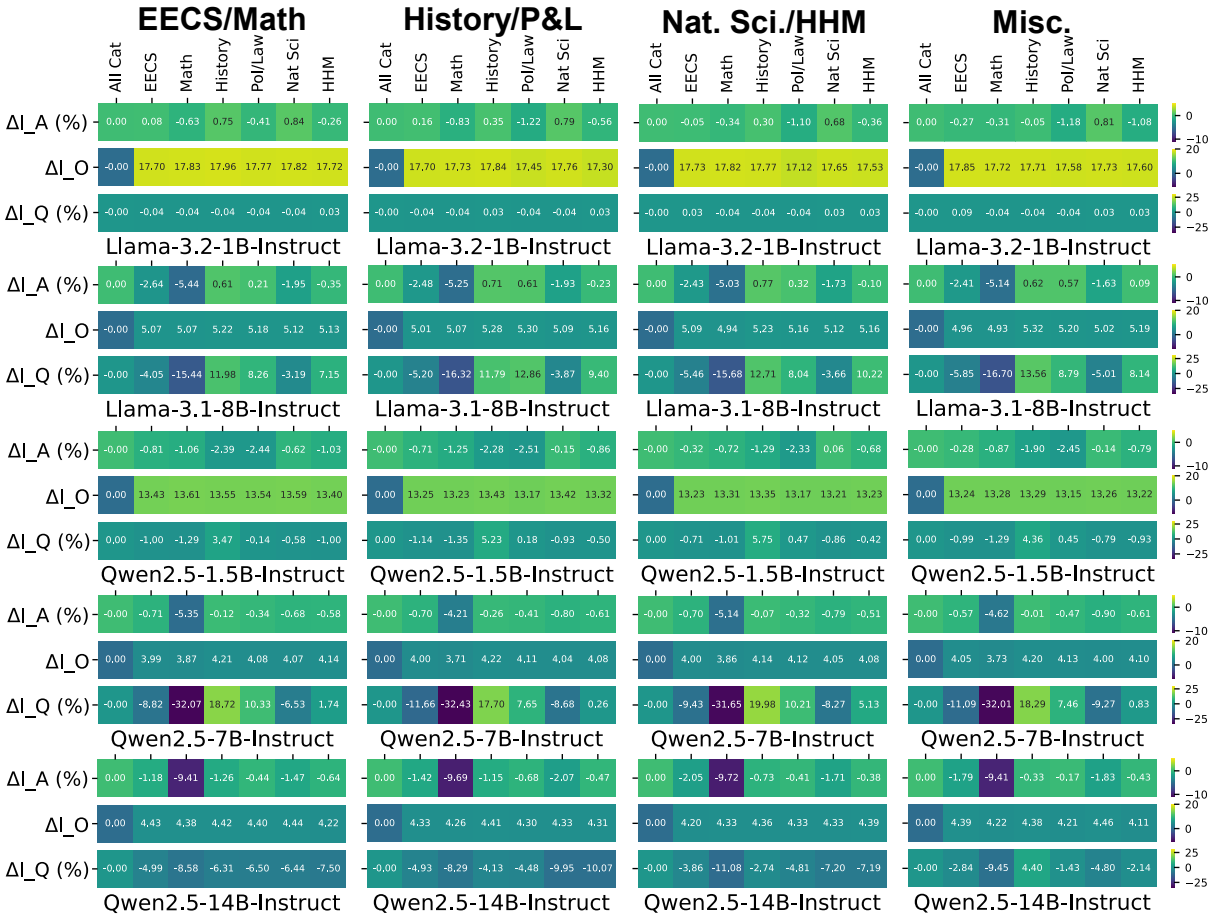



Figure 8: Instability results stratified by instability metric, by dataset category, by persona sub-category.

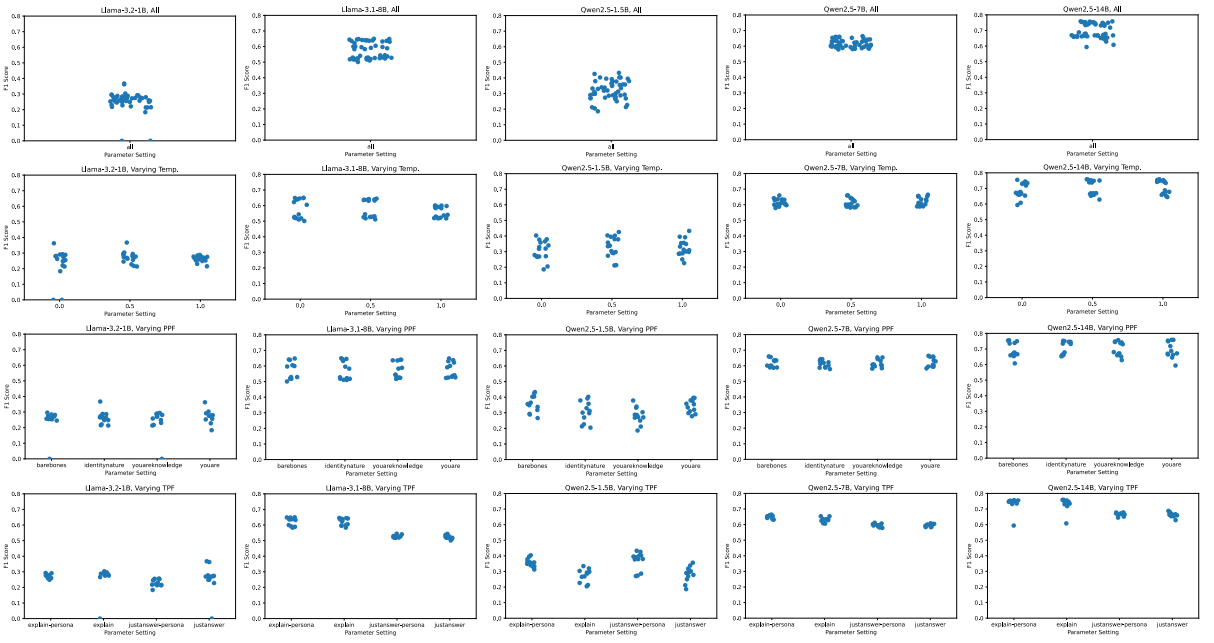



Figure 9: All classifier performances when trained across 48 experimental settings, additionally stratified by specific hyperparameter settings within the experimental settings.

ting. We compare against this metric from prior literature (Sclar et al., 2023) as they investigated robustness to differing prompt features with respect to performance, similar to our investigations into performance instability across experiment settings.

- $I_O$, without the multiplication of the average parity for a model, and the number of different best-performing personas on the total dataset across all experimental settings. We choose the number of best-performing personas as that is a direct experimental outcome (the best persona for a task) that we seek to capture with our metric. Due to reasons relating to initial persona parity for a model as discussed in the paper, we argue that the number of best-performing personas across experiment settings would also be affected by the same quality. Because of this, we remove the regularization term that would control for this quality in order to have more equivalent comparison.

- $I_Q$ and the Fleiss' Kappa (Fleiss, 1971) agreement of the binary vectors of question correctness for the total dataset between experiment settings for one persona, averaged across all personas. Here, we treat separate experiment settings for a persona as different "annotators". This is chosen as agreement between annotators is used in literature as a measure of instance-wise consistency/stability across annotators. Note that a higher agreement rate corresponds to a lower $I_Q$ (or a positive benefit from the baseline), so we would expect to observe a strong positive correlation if these two measurements are consistent, in contrast to the other pairs.

We report the p-value and r of this correlation, averaged across all models, for each of our metrics, in Table 8. We observe strong correlations.

### E.5 Alternative Formulations of $I_O$

We experimented with four formulations of $I_O$:

- A non-weighted standard deviation of ranks

- A non-weighted standard deviation of ranks, with additional multiplication of a model normalization term (as previously defined)

- A weighted standard deviation of ranks, weighted by the persona parity of each experimental setting

- A weighted standard deviation of ranks, weighted by the persona parity of each experimental setting, with additional multiplication of a model normalization term (as previously defined)

As we consider that ranking differences from states of parity "mean less" than those coming from states of a lack of parity (i.e. ranking fluctuations when all personas perform equally means less than ranking fluctuations when all personas perform very differently to one another), we want to weight a given ranking by the parity of the setting it comes from. With this, rankings that come from settings with great parity are weighted less in the final calculation. From this, we focus on the latter two formulations, using a weighted standard deviation with the parity of the experimental settings being used as the weights.

We then investigate the usage of the final model normalization term. Without this final model normalization term, we have the potential for the weighted standard deviation to be "washed out" – in cases where, across experimental settings, models have equivalent levels of high parity, all ranks will be weighted the same, potentially leading to the final $I_O$ score being inflated. To account for this, we add an additional model normalization term, integrating overall parity a model might have. This way, should the weighted standard deviation be "washed out", we are still able to integrate a view of model parity, which we deem as valuable for reasons previously stated.

With this, we wanted to evaluate whether this final model normalization term would "dominate" the final score. Thus, we investigate patterns of solely the model normalization term, and of the $I_O$ implementations – with this, we seek to judge whether the patterns of the formulation *with* the model normalization term completely follow those of solely the model normalization term, implying the model normalization term may be "too strong". We report heatmaps of the results in Figure 10.

From Figure 10, we can observe that, while A (solely the model normalization term) has similar patterns to B (weighted standard deviation with the model normalization term), they are not identical. We note key regions where they differ, such as $I_O$ for temperature across all models. From this,

| Instability Metric Pairs | $I_A \times$ Spread | | $I_O \times$ Num. Best. | | $I_Q \times$ Fleiss' Kappa | |
|---|---|---|---|---|---|---|
| *Models* | p | r | p | r | p | r |
| Llama-3.2-1B | 0.002 | -0.811 | 0.049 | -0.603 | 0.000 | 0.948 |
| Llama-3.1-8B | 0.005 | -0.779 | 0.111 | -0.508 | 0.000 | 0.954 |
| Qwen2.5-1.5B | 0.029 | -0.653 | 0.017 | -0.696 | 0.000 | 0.986 |
| Qwen2.5-7B | 0.019 | -0.688 | 0.040 | -0.624 | 0.000 | 0.990 |
| Qwen2.5-14B | 0.000 | -0.938 | 0.005 | -0.781 | 0.000 | 0.990 |
| *Mean* | 0.011 | -0.774 | 0.044 | -0.642 | 0.000 | 0.974 |

Table 8: Depicting the p-value and r of correlations between each pair of metrics across all models.

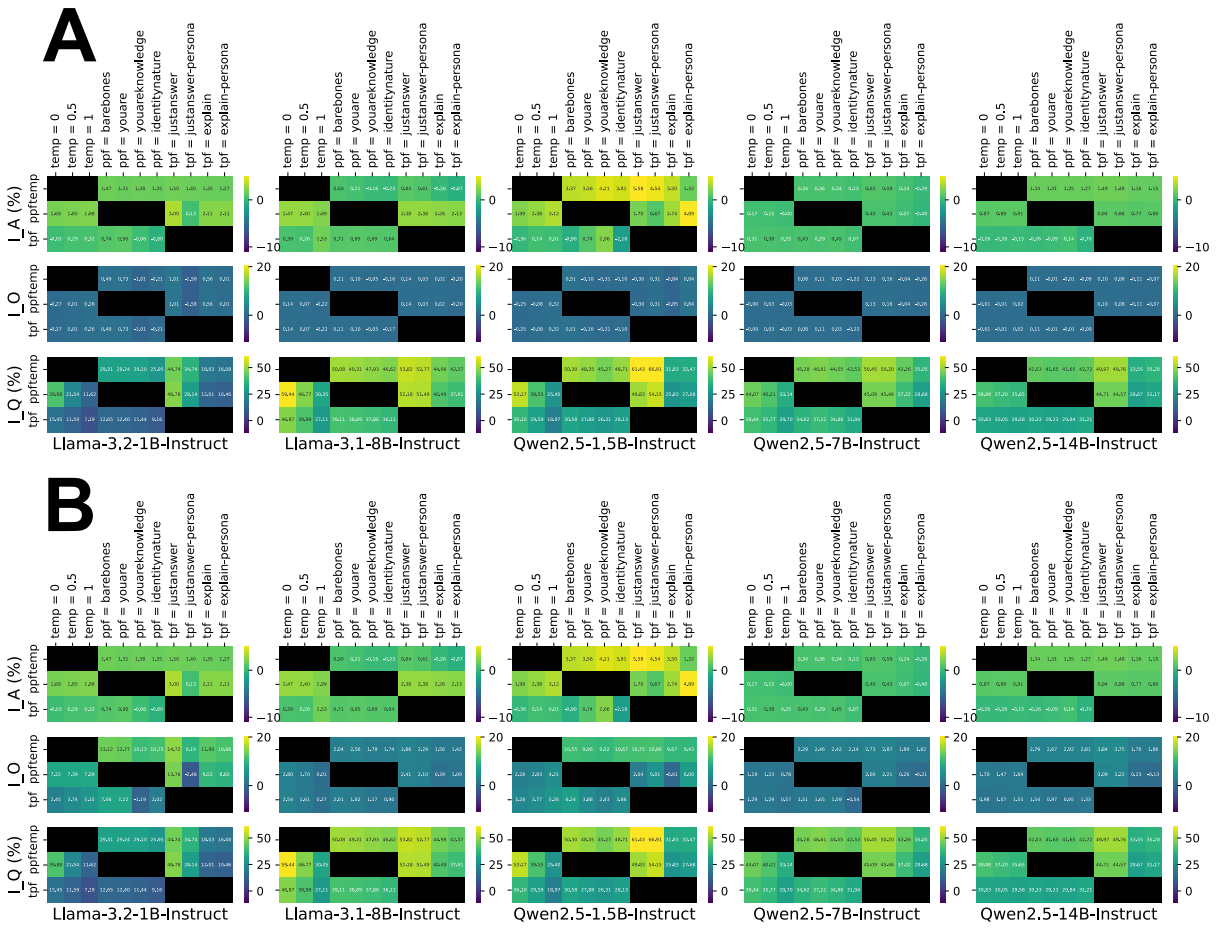



Figure 10: Comparing instability results across different formulations of $I_O$.

we go forward with the final formulation of $I_O$, calculating a weighted standard deviation of ranks across experimental settings, weighted by the parity of each setting, with the addition of a final model normalization term.

### E.6 Analyzing Metric Sensitivity to Question Count

We analyze the impact of the question set size on our metric results. For this analysis, we compute the baseline metrics (as in Table 1) using a randomly selected subset of 75% and 50% of the question set, balanced across dataset categories. We provide results in Table E.6, additionally with the current baseline results for comparison.

As can be seen in Table E.6, our metrics slightly vary with respect to the size of the question set. This aligns with our expectations; with a reduced number of questions, accuracy percentage differences are more substantial, reasonably affecting $I_A$ and $I_O$. Additionally, with a reduced question set, differences in the set of shared questions are similarly more substantial, affecting $I_Q$. While results vary slightly, it can be noted that they still stabilize around a similar value across dataset size/metrics across models, maintaining features such as the relative order of models for each instability metric, and the relative difference between models for each instability metric. To summarize: as the question set size changes, the metrics vary slightly but still follow the same trend across models and experiment settings.

### E.7 Analyzing Metric Sensitivity to Experiment Setting Count

We analyze the impact of the number of experiment settings on our metric results. For this analysis, we compute the baseline metrics using a reduced subset of experiment settings, considering only experiment settings across all temperature settings with TPF = justanswer, explain, explain-persona and PPF = barebones, youareknowledge, identitynature ("Subset A", 27 settings), and experiment settings with TPF = justanswer, explain-persona and PPF = barebones, identitynature ("Subset B", 12 settings). We provide results in Table 10, additionally with the current baseline results for comparison.

As can be seen in Table 10, our metrics vary with respect to the size of the experiment setting set, similar to that of the question set size. This aligns with our expectations. Similar to how, with other measures of spread like classic standard deviation, samples from a population of gradually increasing sizes approach an approximate population value, we see a similar pattern with our metrics. We maintain features such as the relative order of models for each instability metric. To summarize: as the experiment setting set size changes, the metrics vary slightly but approach similar values and still follow the same trend across models and experiment settings.

Additionally, it is important to note that these metric calculations are, of course, dependent on the composition of the sample, like with other measures of spread. While here, the subsets of experiment settings attempt to maintain an equal count of low/high instability hyperparameter settings, differing subsets may produce considerably different results (such as a subset of only low instability hyperparameter settings).

### E.8 Parsing Errors Across Hyperparameter Settings

We consider the percent of parsing errors to be the percent of questions marked incorrect due to an unparseable model response for a specific experiment setting and persona combination. We provide a table of parsing error fraction separated by model and by experiment setting hyperparameter value in Table 11. For each experiment setting hyperparameter value, we average the fraction of parsing errors across all experiment settings with that hyperparameter value for that specific model. For example, for Qwen2.5-1.5B and TPF = bb, we average the fraction of parsing errors for Qwen2.5-1.5B-Instruct across all experiments with TPF = barebones. We additionally provide a similar table, but with the fraction of incorrect questions (inclusive of incorrect answers and unparseable answers) along similar strata in Table 12.

We can note slight variations in parsing errors across hyperparameter settings (with smaller variations in larger models, as expected), with a few notable cases (such as Qwen2.5-1.5B comparing PPF = bb, ya to PPF = yak, in). We can note that, even in these notable cases, by comparing the total incorrect fraction to that of parsing errors, the difference in parsing errors does not account for that of the incorrect questions. In summary, while we note parsing errors, these errors are not enough to explain the noted performance differences that influence our final instability metrics.

| | All Questions Per Dataset Category (100%) | | | #=375 Per Dataset Category (∼75%) | | | #=250 Per Dataset Category (∼50%) | | |
|---|---|---|---|---|---|---|---|---|---|
| *Models / Metric* | $I_A$ | $I_O$ | $I_Q$ | $I_A$ | $I_O$ | $I_Q$ | $I_A$ | $I_O$ | $I_Q$ |
| Llama-3.2-1B | 3.697 | 18.791 | 99.960 | 3.684 | 19.379 | 99.952 | 3.715 | 20.311 | 99.955 |
| Llama-3.1-8B | 2.999 | 5.784 | 78.293 | 2.980 | 6.276 | 78.041 | 3.163 | 7.139 | 78.678 |
| Qwen2.5-1.5B | 6.586 | 14.305 | 98.576 | 6.714 | 15.058 | 98.592 | 6.660 | 15.291 | 98.633 |
| Qwen2.5-7B | 0.972 | 4.539 | 57.688 | 0.963 | 4.855 | 57.500 | 1.074 | 5.702 | 57.893 |
| Qwen2.5-14B | 1.623 | 4.812 | 52.828 | 1.671 | 5.278 | 52.739 | 1.647 | 5.746 | 52.451 |

Table 9: Instability metric results, stratified by the number of questions considered per dataset category. “#" denotes the number of questions considered for each category.

| | All Experiment Settings (#=48) | | | Subset A (#=27) | | | Subset B (#=12) | | |
|---|---|---|---|---|---|---|---|---|---|
| *Models / Metric* | $I_A$ | $I_O$ | $I_Q$ | $I_A$ | $I_O$ | $I_Q$ | $I_A$ | $I_O$ | $I_Q$ |
| Llama-3.2-1B | 3.697 | 18.791 | 99.960 | 3.366 | 13.276 | 99.845 | 3.829 | 11.155 | 98.271 |
| Llama-3.1-8B | 2.999 | 5.784 | 78.293 | 3.121 | 6.126 | 72.863 | 3.210 | 5.528 | 61.083 |
| Qwen2.5-1.5B | 6.586 | 14.305 | 98.576 | 5.618 | 15.022 | 97.505 | 3.990 | 11.818 | 89.042 |
| Qwen2.5-7B | 0.972 | 4.539 | 57.688 | 1.070 | 5.132 | 54.071 | 1.303 | 5.112 | 43.860 |
| Qwen2.5-14B | 1.623 | 4.812 | 52.828 | 1.559 | 5.141 | 47.740 | 1.925 | 3.966 | 32.061 |

Table 10: Instability metric results, stratified by the number of experiment settings considered. “#" denotes the number of experiment settings for a given subset. "Subset A" represents all temperature settings with TPF = {justanswer, explain, explain-persona} and PPF = {barebones, youareknowledge, identitynature}. "Subset B" represents all temperature settings with TPF = {justanswer, explain-persona} and PPF = {barebones, identitynature}.

| *Model / Parsing Error* | Overall | Temp = 0.0 | Temp = 0.5 | Temp = 1.0 | PPF = bb | PPF = ya | PPF = yak | PPF = in | TPF = ja | TPF = ja-p | TPF = e | TPF = e-p |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Llama-3.2-1B | 0.090 | 0.086 | 0.086 | 0.097 | 0.088 | 0.076 | 0.092 | 0.102 | 0.010 | 0.174 | 0.069 | 0.106 |
| Llama-3.1-8B | 0.015 | 0.008 | 0.008 | 0.030 | 0.014 | 0.014 | 0.016 | 0.017 | 0.004 | 0.005 | 0.025 | 0.028 |
| Qwen2.5-1.5B | 0.189 | 0.182 | 0.184 | 0.200 | 0.132 | 0.161 | 0.255 | 0.205 | 0.236 | 0.099 | 0.272 | 0.148 |
| Qwen2.5-7B | 0.013 | 0.012 | 0.012 | 0.016 | 0.010 | 0.006 | 0.014 | 0.022 | 0.001 | 0.000 | 0.020 | 0.032 |
| Qwen2.5-14B | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.002 | 0.001 | 0.001 | 0.002 | 0.002 |

Table 11: The percent of questions marked incorrect due to parsing errors across all models. Results are stratified by hyperparameter setting, where results for a specific hyperparameter setting represent the average across all experiment settings with that specific hyperparameter setting.

| *Model / Parsing Error* | Overall | Temp = 0.0 | Temp = 0.5 | Temp = 1.0 | PPF = bb | PPF = ya | PPF = yak | PPF = in | TPF = ja | TPF = ja-p | TPF = e | TPF = e-p |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Llama-3.2-1B | 0.649 | 0.635 | 0.645 | 0.667 | 0.642 | 0.641 | 0.651 | 0.661 | 0.615 | 0.668 | 0.647 | 0.665 |
| Llama-3.1-8B | 0.367 | 0.350 | 0.357 | 0.394 | 0.364 | 0.366 | 0.367 | 0.372 | 0.381 | 0.384 | 0.351 | 0.352 |
| Qwen2.5-1.5B | 0.563 | 0.546 | 0.555 | 0.588 | 0.535 | 0.549 | 0.595 | 0.573 | 0.576 | 0.504 | 0.617 | 0.554 |
| Qwen2.5-7B | 0.322 | 0.318 | 0.321 | 0.328 | 0.319 | 0.316 | 0.322 | 0.331 | 0.323 | 0.323 | 0.320 | 0.323 |
| Qwen2.5-14B | 0.263 | 0.263 | 0.263 | 0.264 | 0.259 | 0.262 | 0.265 | 0.267 | 0.274 | 0.277 | 0.255 | 0.247 |

Table 12: The percent of questions marked incorrect across all models. Results are stratified by hyperparameter setting, where results for a specific hyperparameter setting represent the average across all experiment settings with that specific hyperparameter setting.

| | Accuracy | | |
|---|---|---|---|
| *Model* | *Non-CoT* | *CoT* | *Delta (CoT - Non-CoT)* |
| Qwen2.5-14B | 0.724 | 0.749 | 0.025 |
| Qwen2.5-7B | 0.677 | 0.678 | 0.001 |

Table 13: Depicting the average accuracy across all experiment settings with non-chain-of-thought (CoT) TPFs ("justanswer", "justanswer-persona") and CoT TPFs ("explain", "explain-persona"). The difference between these values is shown in the final column.

### E.9 Considering Implications of Larger Models

We consider the results for Qwen2.5-14B-Instruct, as seen throughout the paper (and notably in Tables 1 and 4). Here, we can observe that Qwen2.5B-14B-Instruct breaks the model size-related trend for $I_A$ and $I_O$, exhibiting greater instability in accuracy and outcome (represented by these metrics) than Qwen2.5-7B-Instruct.

However, we do not believe this is due to greater inherent instability in the model, but rather due to performance improvements due to chain-of-thought (CoT) prompting (Wei et al., 2022). While smaller models are noted to largely not benefit from CoT, larger models do, creating large accuracy differentials between non-CoT and CoT experiment settings. This can be observed in Table 13. We compare the average accuracy in non-CoT and CoT TPFs ({"justanswer", "justanswer-persona"} and {"explain", "explain-persona"}, respectively). When comparing Qwen2.5-7B and Qwen2.5-14B (the point where we observe our instability scaling trend breaks), we see a large difference in the benefit of CoT to model accuracy.

As we consider CoT-like prompts within our task prompt formats (the "explain" and "explain-persona" variants), we would see larger accuracy differentials between settings using these prompts and those that do not, contributing to a great $I_A$ (due to CoT generally improving larger model performance) and $I_O$ (due to changes in accuracy subsequently changing inter-persona rankings). We observe this in Table 14. When we remove consideration of CoT-like prompts (the "explain" and "explain-persona" TPF variants), we see our scaling trends largely restore, with Qwen2.5-14B and Qwen2.5-7B having nearly the same $I_A$, and Qwen2.5-14B having the lowest $I_O$ and $I_Q$. This supports our belief that models who benefit from CoT (like Qwen2.5-14B) would have greater calculated instability when comparing CoT to non-CoT setting. When we control for this factor, we see our scaling trend extend even to our largest model.

Aside from this, we note that Qwen2.5-14B-Instruct maintains all other existing patterns with models in the same family across all of our analyses. While we do not evaluate larger models due to resource constraints, we reasonably expect that larger Qwen2.5 models, for instance, would continue to maintain these trends.

## F Additional Figures and Figure Details

### F.1 Heatmap Label Keys

The following is the key for the column labels for Figure 3:

- **bb** = barebones, **ya** = youare, **yak** = youare-knowledge, **in** = identitynature, **ja** = justanswer, **ja-p** = justanswer-persona, **e** = explain, **e-p** = explain-persona

The following is the key for the heatmap labels for Figure 5:

- Dataset: **All Cat** = All Categories, **History** = History, **Pol/Law** = Politics and Law, **Soc Sci** = Social Science, **Bus/Acc/Econ** = Business/Accounting and Economics, **EECS** = EECS, **Math** = Math, **HHM** = Human Health and Medicine, **Culture** = Cultural Awareness and Understanding, **Com Sen** = Commonsense Reasoning/Social Awareness

- Persona: **All Cat** = All Categories, **Gender** = Gender, **Sexuality** = Sexuality, **Race/Ethn** = Race/Ethnicity, **Age** = Age, **Disabilities** = Disabilities, **Rel Beliefs** = Religious Beliefs, **Pol Beliefs** = Political Beliefs, **Occup Roles** = Occupational Roles, **EECS/Math** = Occupational roles related to the EECS/Math categories in $\mathcal{D}$, **Hist./P&L** = Occupational roles related to the History/Politics and Law categories in $\mathcal{D}$, **Nat. Sci./HHM** = Occupational roles related to the Natural Science/Human Health and Medicine categories in $\mathcal{D}$, **Baselines** = Baselines

| *Model* | Considering "explain" and "explain-persona" | | | Excluding "explain" and "explain-persona" | | |
|---|---|---|---|---|---|---|
| | $\downarrow I_A$ | $\downarrow I_O$ | $\downarrow I_Q$ | $\downarrow I_A$ | $\downarrow I_O$ | $\downarrow I_Q$ |
| Llama-3.2-1B-Instruct | 3.697 | 18.791 | 99.960 | 4.246 | 20.841 | 95.882 |
| Llama-3.1-8B-Instruct | 2.999 | 5.784 | 78.293 | 1.782 | 4.551 | 58.848 |
| Qwen2.5-1.5B-Instruct | 6.586 | 14.305 | 98.576 | 6.515 | 12.234 | 76.759 |
| Qwen2.5-7B-Instruct | **0.972** | **4.539** | 57.688 | **0.546** | 2.877 | 25.058 |
| Qwen2.5-14B-Instruct | 1.623 | 4.812 | **52.828** | 0.593 | **1.968** | **13.917** |

Table 14: A table depicting the baseline overall instability results for each model using our three instability metrics, comparing when considering settings with "explain" and "explain-persona" TPFs versus without. **Bolded** values indicate the lowest value per column.

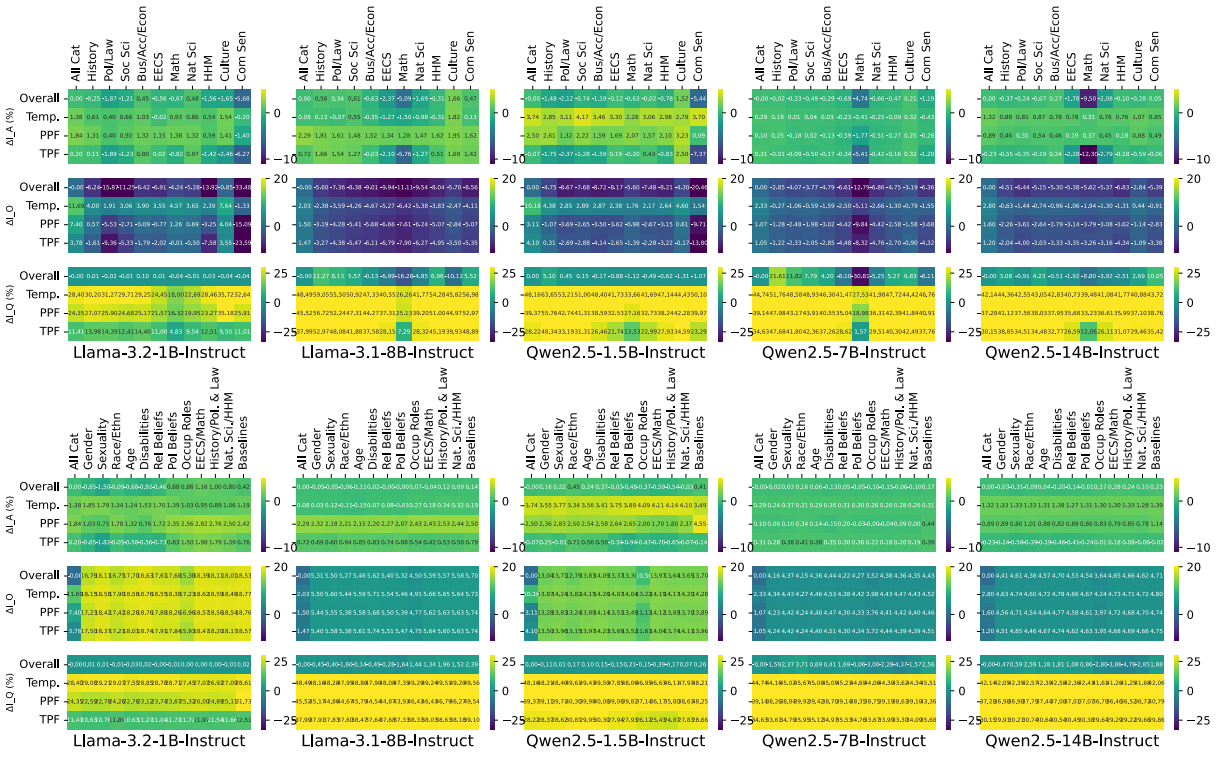



Figure 11: Hyperparameter instability results per metric, per model, stratified by dataset category and persona category.

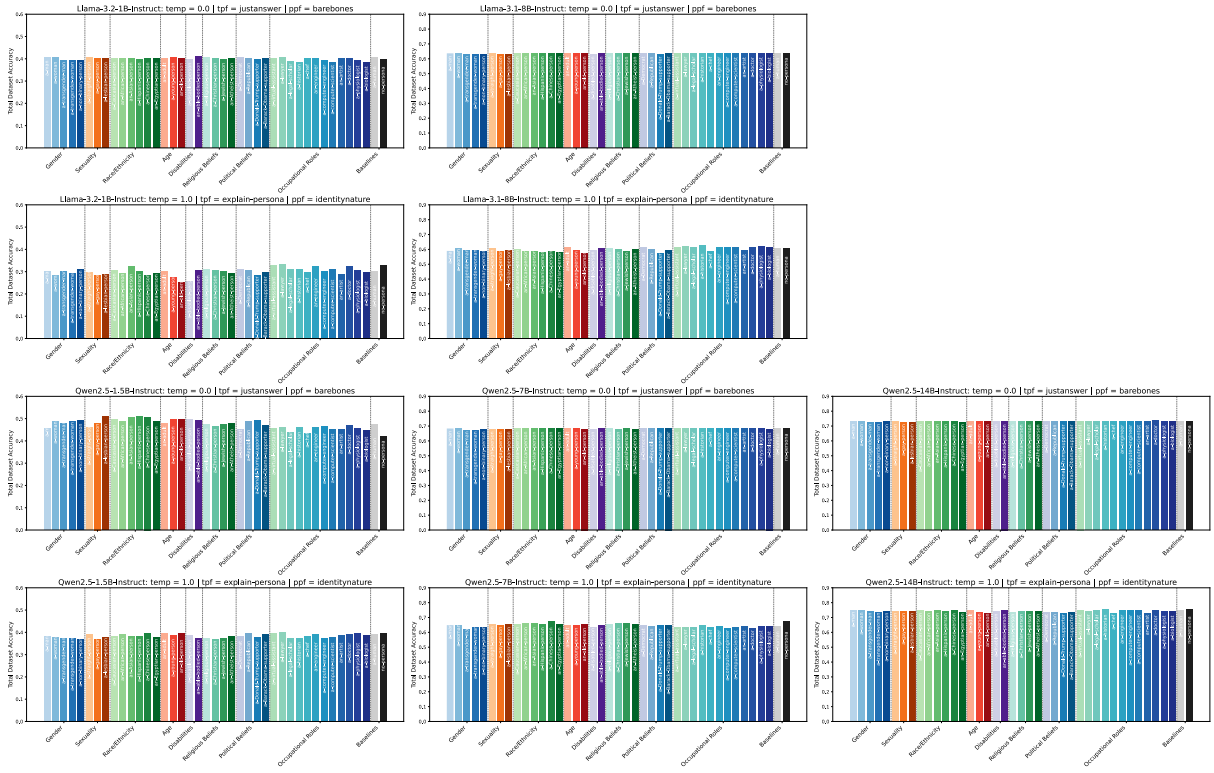



Figure 12: All inter-persona performance distributions between the "unstable" and "stable" experiment settings, as identified using our metrics, across all models.