Stable Behavior, Limited Variation: Persona Validity in LLM Agents for Urban Sentiment Perception

Neemias B da Silva*, Rodrigo Minetto*, Daniel Silver†,Thiago H Silva*†
*Universidade Tecnológica Federal do Paraná (UTFPR), Curitiba, Brazil
†University of Toronto, Toronto, Canada
Email: neemiasbuceli@alunos.utfpr.edu.br, rminetto@utfpr.edu.br, {dan.silver,th.silva}@utoronto.ca

***Abstract*—Large Language Models (LLMs) are increasingly used as proxies for human perception in urban analysis, yet it remains unclear whether persona prompting produces meaningful and reproducible behavioral variation. We investigate whether distinct personas influence urban sentiment judgments generated by multimodal LLMs. Using a factorial set of personas spanning gender, economic status, political orientation, and personality, we instantiate multiple agents per persona to evaluate urban scene images from the PerceptSent dataset and assess within-persona convergence and cross-persona differentiation. Results show strong convergence among agents sharing a persona, indicating stable and reproducible behavior. However, cross-persona differentiation is limited: economic status and personality induce statistically detectable but practically modest variation, while gender shows no measurable effect and political orientation only negligible impact. Agents also exhibit an extremity bias, collapsing intermediate sentiment categories common in human annotations. As a result, performance remains strong on coarse-grained polarity tasks but degrades as sentiment resolution increases, suggesting that persona prompting does not capture fine-grained perceptual judgments. To isolate the contribution of persona conditioning, we additionally evaluate the same model without personas. Surprisingly, the no-persona model matches or exceeds persona-conditioned agreement with human labels across all task variants, suggesting that persona prompting may add limited annotation value in this setting.**

***Index Terms*—Persona Prompting, Urban Perception, Sentiment Analysis, Human-AI Alignment**

I. Introduction

Large language models (LLMs) are increasingly used as simulated agents to approximate human judgments [1]–[3], yet it remains unclear whether they can reproduce not only consistent but also sufficiently diverse judgments [4]–[6]. A common mechanism for introducing diversity in LLM agents is *attribute-based persona prompting*, in which attributes such as gender, economic status, personality, and political orientation are injected into prompts to guide model behavior [1], [5], [7]. Despite its growing adoption, a central question remains: *Do personas produce distinct, reproducible, and sufficiently granular patterns of human perceptual judgment, or merely create the appearance of diversity without substantial behavioral differentiation?* Addressing this question is critical for evaluating the validity of LLM-based agent simulations, particularly in light of concerns about the reduction of within-group variance [8], the lack of alignment with human psychometric baselines [9], and output variance across prompt templates [10].

This question is especially relevant for perceptual tasks in urban environments. Judgments about urban scenes—such as whether a place appears pleasant, unsafe, or neglected—vary across individuals and social contexts [11]. Prior work on LLM-based social simulation and persona prompting suggests that demographic and psychological attributes can influence model behavior and are commonly used to define synthetic agents [1], [5]. Motivated by this, we adopt gender, economic status, political orientation, and personality as controlled persona attributes to evaluate whether such prompts induce stable and distinguishable patterns of urban sentiment judgment.

Collecting annotations that reflect this diversity typically requires large numbers of human annotators, making dataset construction costly and time-consuming [12]. Multimodal LLMs (MLLMs) offer a potential alternative as *synthetic annotators*, capable of interpreting images and generating structured perceptual descriptions. However, their usefulness depends not only on producing stable outputs, but also on capturing variability at the granularity of human perception, particularly the ability to represent intermediate and ambiguous judgments.

In this paper, we empirically evaluate whether attribute-based personas induce distinguishable and stable behavior in MLLM agents performing urban image sentiment annotation. We construct a factorial design of 24 personas combining gender, economic status, political orientation, and personality traits, and instantiate 50 independent agents per persona to annotate 50 urban scene images from the PerceptSent dataset [11]. We analyze both *within-persona convergence* (consistency among agents sharing a persona) and *cross-persona differentiation* (differences across personas).

Our results show strong within-persona convergence, indicating stable and reproducible behavior. However, cross-persona differentiation shows complex patterns: economic status and personality show statistically detectable but practically modest variation in annotations, while gender shows no measurable effect; political orientation has only a negligible impact. Moreover, agents exhibit a systematic extremity bias, collapsing intermediate sentiment categories that are prevalent in human annotations. As a result, while performance remains strong on coarse-grained polarity judgments, it degrades as task resolution increases. These findings suggest that simple attribute-based persona prompting stabilizes outputs but fails to generate sufficiently fine-grained behavioral variation, limit-

ing its ability to capture the fine-grained perceptual judgments characteristic of human labelers.

This paper makes four main contributions:

- **A framework for evaluating persona validity in MLLM agents.** We propose a methodology that jointly measures within-persona convergence and cross-persona differentiation to assess whether persona prompts produce meaningful behavioral differences.
- **A controlled large-scale study of persona profiles in multimodal annotation.** We instantiate 1,200 persona-conditioned agents across 24 persona profiles to evaluate 50 urban scene images, analyzing sentiment predictions under a controlled factorial design. Also, we share code and generated data publicly[1].
- **Insights into the behavioral and resolution-dependent effects of persona prompting.** Our results show that economic status and personality drive detectable but limited cross-persona differentiation, while gender shows no measurable effect, and political orientation has only a negligible impact. More broadly, we identify a persistent tendency in MLLM agent judgments: while persona prompting stabilizes outputs and supports coarse-grained polarity judgments, it fails to produce fine-grained variation, collapsing intermediate sentiment categories and limiting its ability to capture fine-grained perceptual judgments. This reveals a resolution-dependent limitation of persona prompting.
- **A no-persona ablation suggests that persona conditioning may not improve annotation quality.** We run the same model on the same 50 images without any persona prompt, under two reasoning modes (with and without extended chain-of-thought (CoT)), with 5 independent repetitions per condition. The no-persona MLLM agents achieve agreement comparable to, and in some cases higher than, the 1,200 persona-conditioned MLLM agents. Furthermore, no-persona predictions show a substantially less extreme bimodal collapse, with Neutral class usage close to human rate (23–25% vs. 21%).

The remainder of this paper is organized as follows. Section II reviews related work. Section III describes the methodology. Section IV presents experimental results. Section V discusses limitations and ethical considerations. Section VI concludes the paper.

II. Related Work

Recent advances in LLMs have motivated their use as computational proxies for human behavior, with studies showing they can reproduce stylized experimental findings and support survey research, behavioral simulation, and content analysis [2], [7], [13]. At the same time, prior work highlights important limitations, including bias, reduced human variability, and challenges for psychometric validity and reproducibility [8], [9]. While some studies find partial alignment between LLMs and human judgments in specific tasks [14], it remains unclear whether LLM-based agents can reliably substitute for human participants.

More closely related is work on attribute-based persona prompting, where demographic or psychological attributes guide model behavior. Prior studies show that personas can affect outputs [10], [15], while raising concerns that synthetic personas may distort underlying belief structures [16]. Previous work mainly evaluates cross-group differences in text generation tasks, with limited attention to within-persona stability, meaningful cross-persona differentiation, or improvement in agreement with human labels relative to a neutral prompt, particularly in multimodal perceptual tasks. Work specifically on MLLMs as synthetic annotators for perceptual and affective tasks remains sparse, with most studies focusing on text-only models [11], [12], [17]. Addressing these gaps is the focus of this work.

III. Methodology

The methodology comprises three phases (Figure 1): (i) a *persona design* constructs a balanced factorial set of attribute-based personas; (ii) an *annotation* phase that runs two parallel conditions on the same 50 urban scene images: the main *persona* condition, in which 1,200 persona-conditioned agents annotate each image, and a *no-persona ablation*, in which the same model annotates each image with a neutral observer prompt under two reasoning modes; and (iii) an *evaluation* phase that measures within-persona convergence and agreement with human ground truth across all conditions.

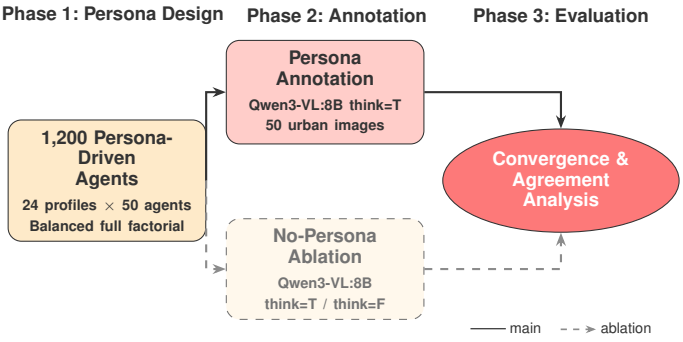


Fig. 1: High-level methodology overview.

A. Persona Design

We construct a balanced full factorial design across four persona attributes (Table I). We selected three commonly studied sociodemographic attributes (gender, economic status, and political orientation) and included personality as an additional attribute, following the suggestion that persona effects may depend not only on demographic cues but also on psychological traits that shape variability in agent behavior [5]. These archetypes were chosen to span contrasting cognitive and affective orientations. The number of levels per attribute was kept small to maintain a tractable full factorial design and keep the experimental workload computationally feasible.

[1]https://github.com/neemiasbsilva/mllm-persona-evaluation.

TABLE I: Persona attributes and category levels.

Attribute	Categories	Levels
Gender	Male, Female	2
Economic Status	Low income, High income	2
Political Spectrum	Progressive, Conservative	2
Personality	Analytical, Empathetic, Pragmatic	3

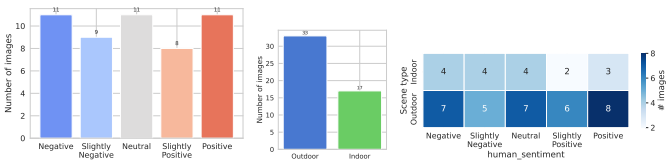
This design yields $2 \times 2 \times 2 \times 3 = 24$ unique personas. Each persona is instantiated with 50 independent agents, resulting in 1,200 agents in total. Agents are generated independently from the same persona prompt without shared context or conversation history. Residual variation within a persona arises only from low-temperature (0.1) floating-point non-determinism.

Each persona is defined by a set of four attributes injected into a structured system prompt. For example: *Gender: Male; Economic status: Low income; Political spectrum: Progressive; Personality archetype: Analytical.*

B. Image Dataset

The *PerceptSent* dataset [11] contains 5,000 urban scene photographs, each annotated by five human workers with: (i) a sentiment score on a 5-point ordinal scale (Negative, Slightly Negative, Neutral, Slightly Positive, Positive); and (ii) free-text perception labels drawn from a closed vocabulary.

A stratified subset of 50 images is drawn proportional to sentiment $\times \frac{indoor}{outdoor}$ to ensure balanced coverage of both affective polarity and spatial context. Figure 2 shows details of the selected sample. The human modal sentiment distribution of the subset is approximately uniform across the five classes (8–12 images per class), with a slight over-representation of Negative scenes consistent with the full dataset's human label distribution. The indoor/outdoor split in the subset reflects the strong compositional skew of the source dataset: 92.0% of the 5,000 PerceptSent images depict outdoor scenes (4,598 outdoor vs. 402 indoor), a bias inherent to how urban scene collections are typically assembled (e.g., street, views, public spaces, buildings). Stratified sampling, therefore, preserves this ratio rather than artificially balancing it, ensuring the subset remains representative of the population from which conclusions will be drawn. The experiment targets $1{,}200 \times 50 = 60{,}000$ annotation triples.



(a) Human modal sentiment (b) Scene Type (c) Sentiment × scene type (image count)

Fig. 2: Characteristics of the studied dataset.

C. Annotation Pipeline

The annotation pipeline is implemented as a three-stage *LangGraph*[2] workflow with conditional retry logic (Figure 3).

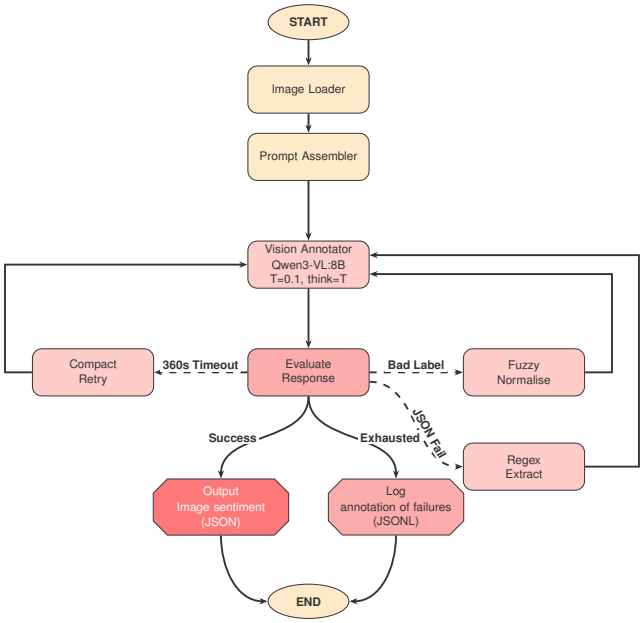


Fig. 3: Three-stage annotation pipeline comprising image loading, prompt assembly, and vision annotation. The Eval & Retry cluster handles timeout, JSON parsing, and label errors with up to 3 retries before logging failures.

1) Model Configuration: All annotations use *Qwen3-VL:8B* and are run locally via Ollama. See Table II for configuration details.

TABLE II: Vision model configuration.

Parameter	Value
Model	`Qwen3-VL:8B`
Temperature	0.1
Chain-of-thought	Enabled (`think=True`[3])
Seed	42
Context window	4,096 tokens
Max output tokens	Unlimited (−1)
Timeout	360 s per call

2) Pipeline Nodes: Image Loader reads JPEG images from local storage and encodes them as base64 strings. Images are pre-cached in memory before the annotation loop to avoid repeated disk I/O.

Prompt Assembler constructs the system prompt by injecting the four attribute tags into the following template:

```
You are a person with the following persona profile:
- Gender: {gender}
- Economic status: {economic_status}
- Political leaning: {political_spectrum}
- Personality archetype: {personality}

CRITICAL RULES:
1. You are this person. You are NOT an AI.
2. Evaluate the image exactly as this person would.
3. Do not break character.
```

[2]https://www.langchain.com/langgraph

[3]`think=False` was also evaluated in the no-persona ablation condition (Section III-D)

```
TASK --- IMAGE SENTIMENT ANNOTATION: Look at the image and respond
with a JSON object and nothing else.

The JSON must have exactly these four keys:
"sentiment" : one of {Negative, SlightlyNegative, Neutral,
SlightlyPositive, Positive}
"perceptions" : a JSON array of 1--5 labels chosen EXCLUSIVELY from
the PERCEPTION VOCABULARY listed below --- do NOT invent new labels
"caption" : a 1--2 sentence objective description of what is shown in
the image, written independently of your persona
"justification" : a single sentence in YOUR voice explaining your
sentiment score

PERCEPTION VOCABULARY (593 labels --- choose 1--5 that best match the
image).

Example response: {"sentiment": "Negative", "perceptions":
["Violence", "Debris/ Destruction"], "caption": "A damaged street
with rubble and broken glass scattered across the pavement.",
"justification": "Looks like the same mess they always leave after
those marches downtown."}

Return ONLY the JSON object. No markdown, no code fences, no
additional text.
```

The annotation instruction needs a JSON response with the field `sentiment`, which represents a 5-class ordinal.

Vision Annotator sends the assembled prompt and base64 image to the *Qwen3-VL:8B* model via the Ollama API. A multi-layer error-handling strategy with up to *3 parse retries*: (i) timeouts, retried with a compact JSON-only prompt; (ii) JSON parse errors, repaired via regex extraction and correction instructions; (iii) label normalization via fuzzy matching. Annotations exhausting all retries are logged and skipped.

3) Evaluate Response: The pipeline uses semaphore-bounded async concurrency to evaluate the response. On startup, existing *JSONL* output files are read to skip completed `(persona_id, image_id)` pairs, enabling fault-tolerant resumption after crashes. Successful responses and failed annotations are written to separate JSONL files, allowing concurrent runs of different conditions without file locking conflicts.

D. No-Persona Ablation Condition

To isolate the contribution of persona conditioning, we run the same *Qwen3-VL:8B* on the same 50 images without injecting any persona. The system prompt is replaced with a neutral observer instruction:

```
You are an impartial image annotator. Evaluate each image objectively,
without any personal or demographic lens.

[Same JSON annotation task and perception vocabulary as above.]
```

We evaluate two reasoning modes: `think=True` (extended CoT, identical to the persona setting) and `think=False` (direct response without an internal reasoning step), each run 5 independent times over the same 50 images (250 annotations per condition). The modal sentiment across the 5 runs is used as the condition's prediction for each image. These two conditions let us ask, independently, whether persona prompting adds annotation value beyond the model's intrinsic visual understanding, and whether CoT reasoning influences agreement with human labels when no persona is present.

E. Evaluation Metrics

We assess *within-persona convergence* using two measures. The *modal ratio* is the fraction of annotations within each (`persona_id`, `image_id`) group that match the modal sentiment label. A value of 1.0 indicates full consensus; the chance baseline for 5 classes is 0.2. The *sentiment variance* is computed over integer-encoded scores (Neg $=-2$ to Pos $=+2$) within each group. A value of 0.0 indicates that all MLLM agents assigned the same class, while higher values reflect greater dispersion.

Agreement with human ground truth is evaluated using 5-fold cross-validation over the Cartesian product of consensus thresholds σ and task formulations P. Figure 4 exemplifies how one particular image is annotated and why this particular image has been assigned for P_3 and σ_3 agreement. Here, $\sigma \in \{3, 4, 5\}$ denotes the minimum number of human annotators required to agree (higher σ yields stricter consensus), and $P \in \{P_2^-, P_2^+, P_3, P_5\}$ defines the task: binary negative polarity (P_2^-), binary positive polarity (P_2^+), 3-class (P_3), and 5-class (P_5).

0	1	3	1	0

Evaluator's votes

Problem setup: P_3
Dominance threshold: σ_3
Labeled as: Neutral
Human perceptions (# of votes):
low resolution/quality (3), everyday image (3), unpleasant colors (1), introspective (1), art/architecture (1).

Fig. 4: PerceptSent dataset example.

Since the annotation pipeline performs no model training or fine-tuning, the same model weights are used for every agent call. All images that satisfy the σ criterion are included in every fold. To estimate variability, we resample 60% of the annotation pool per image in each fold (716 annotations per image, from ≈1,194 total), recompute the modal sentiment, and evaluate performance. This yields between 2,868 and 35,850 annotations per fold, depending on σ. Metrics are averaged across folds.

Uncertainty is reported as 95% confidence intervals. This design progressively filters ambiguous images, ensuring that evaluation focuses on cases with stronger human agreement.

We report Macro F1 (primary metric, following the MLLM-sent framework [17]) and Cohen's κ (unweighted for binary tasks, linear for 3-class, and quadratic for 5-class).

IV. Experiments and Results

A. Annotation Statistics

The annotation run produced *59,708 successful annotations* from 1,200 unique agents across 50 images, covering all 24 persona profiles. Table III summarizes the pipeline statistics. The failure rate is 0.49% (292 annotations), with 98.2% of successful annotations completed on the first validation attempt. The no-persona ablation runs each produced exactly 250 annotations across the same 50 images with zero failures, yielding one annotation per image per condition.

Failures are demographically balanced (50.6% Male, 49.4% Female; 50.6% Low income, 49.4% High income) and concentrate on a small number of structurally complex scenes, ruling out MLLM agents-specific parsing effects and indicating that errors are primarily content-driven.

TABLE III: Annotation pipeline statistics across all three conditions.

Metric	Persona	No-pers. think=T	No-pers. think=F
Successful ann.	59,708	250	250
Failed ann.	292	0	0
Failure rate	0.49%	0%	0%
Unique agents	1,200	5	5
Personas / conditions	24	N/A	N/A
Unique images	50	50	50
Ann. / image	≈1,194	5	5

B. Predicted Sentiment Distribution

Figure 5 compares predicted sentiment distributions across all three conditions against the human ground truth. Persona prompting exhibits a pronounced bimodal collapse, with 77.8% of predictions concentrated at the polarity extremes (Negative + Positive), compared to 35.3% in human annotations. Intermediate categories are correspondingly underused, indicating an extremity bias relative to the central tendency commonly observed in human responses [18].

By contrast, the no-persona conditions are substantially better calibrated, reducing extreme-label usage (65–66%) and producing Neutral rates (23–25%) close to the human baseline (21%). This suggests persona conditioning may amplify rather than mitigate the bimodal collapse. Importantly, disagreement arises primarily from collapsed intermediate categories rather than polarity inversions: errors are largely adjacent in the 5-class task (Section IV-F), indicating a failure of granularity rather than scene comprehension.

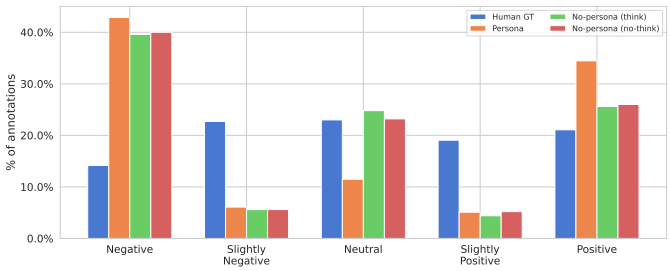


Fig. 5: Predicted sentiment distribution across all three conditions vs. human ground truth (GT).

C. Within-Persona Sentiment Convergence

MLLM agents instantiated with the same persona show strong within-persona convergence. For each of the $24 \times 50 = 1{,}200$ (`persona_id`, `image_id`) groups, we compute two complementary metrics: *modal ratio* and *sentiment variance*. Table IV summarizes these results across all groups.

A mean modal ratio of 0.871 and median of 0.980 (with median sentiment variance of 0.000) indicate near-unanimous agreement within most groups, many of which achieve full consensus (modal ratio = 1.0). Although a minority of groups show disagreement, the vast majority exhibit full or near-full consensus, confirming stable and reproducible sentiment judgments across identically prompted MLLM agents. This high convergence is consistent with the near-deterministic setup (temperature 0.1, seed 42, fixed model weights and context window), with residual disagreement likely reflecting floating-point non-determinism.

TABLE IV: Within-persona convergence metrics. *Agents/group*: number of agents in each group.

Metric	Mean	± SD	Q1	Median
Modal ratio	0.871	±0.185	0.760	0.980
Sentiment variance	0.211	±0.405	0.000	0.000
Agents/group	49.6	±1.4	50	50

D. Do Personas Produce Distinct Judgments? Cross-Persona Differentiation

Establishing convergence within personas is only half of the answer; the other is whether *different* personas produce systematically different sentiments. Figure 6 shows the row-normalized sentiment distribution for all 24 personas. Some personas are more Negative-biased (top), while others are more Positive-biased (bottom). Notably, intermediate classes (Neutral, Slightly Negative, Slightly Positive) are largely absent across all personas, a direct consequence of the bimodal collapse noted above.

Despite this, the *relative ordering* of Negative-to-Positive proportions varies meaningfully across personas, indicating that persona composition modulates the MLLM agents' sentiment prior. A Kruskal–Wallis test over ordinal sentiment scores ($[-2, +2]$) grouped by the 24 personas rejects the null hypothesis of equal distributions ($H(23) = 489.52$, $p = 5.33 \times 10^{-89}$, $\varepsilon^2 = 0.0078$), indicating statistically significant differences. However, effect sizes are small, suggesting limited practical differentiation.

To identify which persona attributes drive this variation, we decompose sentiment distributions along each attribute (Figure 7). Factor-level Mann–Whitney U and Kruskal–Wallis tests over ordinal scores reveal which attributes produce statistically significant separation.

Economic status shows the largest effect: High-income personas have mean $\mu = -0.04$ (median $= 0$), while Low-income personas have $\mu = -0.32$ (median $= -1$), making it the only attribute that shifts the median below zero—indicating a qualitative polarity shift ($U = 482{,}309{,}610$, $p = 4.29 \times 10^{-77}$, $r = 0.071$).

Personality also shows a significant effect: all three archetypes share median $= 0$, but differ in mean ($H(2) = 84.57$, $p = 4.33 \times 10^{-19}$, $\varepsilon^2 = 0.0014$). Empathetic MLLM agents are the least Negative-biased ($\mu = -0.08$), while Analytical and Pragmatic MLLM agents are more Negative-leaning ($\mu = -0.21$ and $\mu = -0.25$, respectively).

Political spectrum is statistically detectable ($U = 432{,}646{,}050$, $p = 4.73 \times 10^{-11}$, $r = 0.025$), but with negligible effect size: Conservative ($\mu = -0.23$) and Progressive ($\mu = -0.13$) personas differ by only 0.10 points. This reflects the bimodal collapse: since MLLM agents overwhelmingly predict extreme labels (-2 or $+2$), all factor levels share

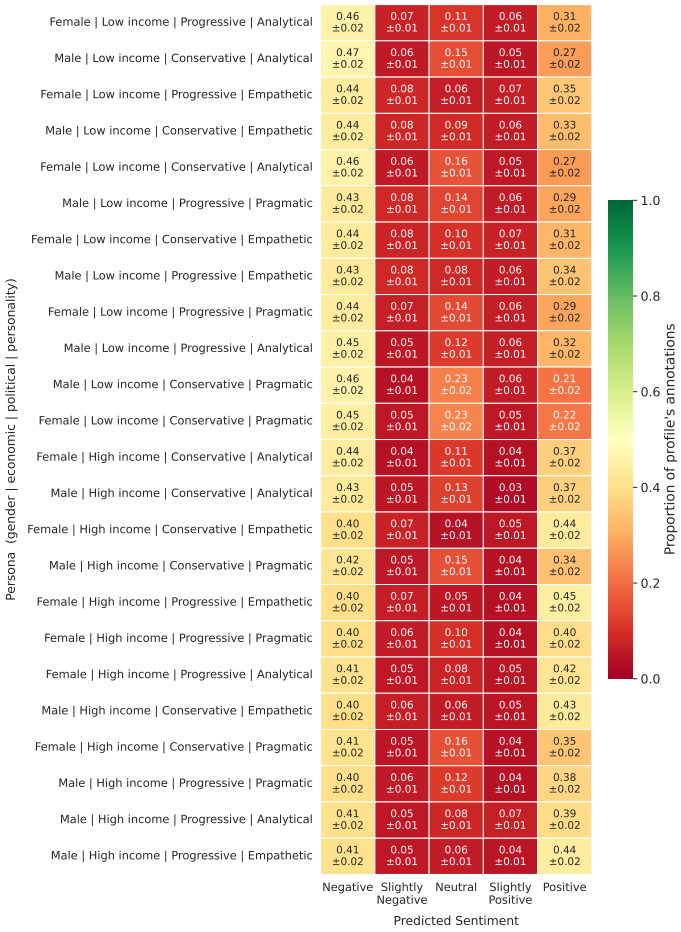


Fig. 6: Row-normalized sentiment proportion heatmap across all 24 persona profiles, sorted by descending Negative fraction. Each cell shows the fraction of annotations in a given sentiment class.

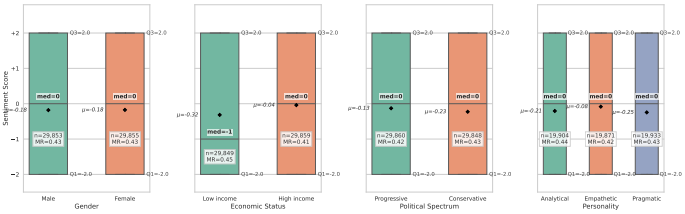


Fig. 7: Sentiment distribution by persona attribute ($[-2, +2]$ scale).

identical quartiles ($Q_1 = -2$, $Q_3 = +2$), leaving the mean as the only discriminating statistic. The statistical significance is therefore driven more by large sample size ($N = 60{,}000$) than by meaningful behavioral separation.

Gender shows no measurable effect ($U = 446{,}481{,}262$, $p = 0.666$, $r = 0.002$), with Male and Female personas indistinguishable ($\mu = -0.18$).

E. Persona vs. No-Persona: Does Conditioning Add Value?

Table V compares agreement with human ground truth across all three conditions at $\sigma = 3$ (all 50 images retained), which is the most informative threshold as it includes the full image set without requiring near-unanimous human consensus.

TABLE V: Agreement with human ground truth at $\sigma = 3$. Macro F1 and Cohen's κ (P_2: unweighted; P_3: linear; P_5: quadratic). No-persona uses 5 runs with modal aggregation. Confidence intervals use image-level bootstrap ($N = 1{,}000$).

Task	Metric	Persona (1,194 ann.)	No-pers. think=T	No-pers. think=F	Largest Δ over persona baseline
P_2^-	F1	0.807	**0.862**	**0.862**	+0.055
	κ	0.615	**0.725**	**0.725**	+0.110
P_2^+	F1	0.797	**0.836**	0.816	+0.039
	κ	0.597	**0.672**	0.634	+0.075
P_3	F1	0.588	**0.665**	**0.665**	+0.077
	κ	0.448	**0.507**	**0.507**	+0.059
P_5	F1	0.309	**0.433**	0.425	+0.124
	κ	0.791	**0.879**	0.865	+0.088

The results show that the no-persona condition matches or exceeds persona-conditioned agreement across all four task variants, despite using only 5 runs per image versus ≈1,194 persona-conditioned annotations. Both think=True and think=False conditions perform similarly, with the largest improvements reaching +0.055 F1 on P_2^-, +0.077 F1 on P_3, and +0.124 F1 on P_5 relative to persona prompting. The consistent outperformance across both reasoning modes confirms the result is not specific to the CoT setting.

This finding is particularly important given the asymmetry in annotation volume: the persona-conditioned system aggregates annotations from 1,200 agents, each evaluating an image once, yet this large-scale aggregation is matched or exceeded by only five no-persona inferences. The result suggests that the performance observed in the persona-conditioned system is attributable primarily to the MLLM's intrinsic visual understanding of the scene, rather than to any behavioral diversity introduced by persona conditioning. Persona prompting may introduce variability without improving signal in this setting.

It is important to note that the evaluation setups are not entirely symmetric: persona prompting aggregates ≈1,194 annotations per image, whereas no-persona uses 5 runs (5 annotations per image). Although this reduces asymmetry relative to a single run, the no-persona advantage remains consistent across both reasoning modes and all runs. A fully symmetric comparison would require substantially more no-persona repetitions and is left for future work.

F. Agreement with Human Ground Truth

Table VI reports point estimates with 95% confidence intervals for all 12 (σ, P) subsets, revealing a resolution-dependent pattern of agreement: strong polarity recognition but a systematic breakdown at finer sentiment granularity.

The results reveal a clear *polarity–granularity asymmetry*: the MLLM agent reliably distinguishes positive from negative

TABLE VI: Agreement metrics (60% annotation resample per image; 95% CI via image bootstrap, $N = 1{,}000$). n_{img}: images passing the σ filter; n_{ann}: annotations per evaluation. Cohen's κ is unweighted for P_2, linear-weighted for P_3, and quadratic-weighted for P_5. †: $\sigma = 5$ scores of 1.000 arise from extreme filtering and should be interpreted with caution.

Problem	σ	n_{img}	n_{ann}	**Macro F1** (95% CI)	**Cohen** κ (95% CI)
P_2^- (binary)	3	50	35,800	0.807 ± 0.12	0.615 ± 0.23
	4	41	29,356	0.842 ± 0.12	0.688 ± 0.22
	5	28	20,048	1.000† ± 0.00	1.000† ± 0.00
P_2^+ (binary)	3	50	35,800	0.797 ± 0.11	0.597 ± 0.22
	4	43	30,788	0.859 ± 0.10	0.720 ± 0.19
	5	24	17,184	1.000† ± 0.00	1.000† ± 0.00
P_3 (3-class)	3	45	32,220	0.588 ± 0.11	0.448 ± 0.21
	4	34	24,344	0.708 ± 0.18	0.687 ± 0.26
	5	18	12,888	1.000† ± 0.00	1.000† ± 0.00
P_5 (ordinal-5)	3	36	25,776	0.309 ± 0.09	0.791 ± 0.12
	4	16	11,456	0.514 ± 0.23	0.849 ± 0.17
	5	4	2,864	1.000† ± 0.00	1.000† ± 0.00

imagery, but this competence does not extend to intermediate sentiment categories.

Binary performance (P_2^-, P_2^+) is strong from the outset—Macro F1 of 0.807 and 0.797 at $\sigma = 3$, which confirms that coarse polarity discrimination is robust even without requiring strong inter-annotator consensus.

The 3-class task (P_3) marks the first substantive breakdown in agreement: at $\sigma = 3$, F1 drops to 0.588 despite the MLLM agents continuing to separate positive from negative scenes. The gap relative to binary performance is attributable almost entirely to the Neutral class, which is systematically misclassified.

The 5-class task (P_5) makes this pattern explicit at the most informative threshold ($\sigma = 3$), with 36 images (all classes present). Macro F1 falls to 0.309. The quadratic-weighted κ of 0.791 at the same threshold reveals that errors are nevertheless adjacent rather than polarity inversions—ordinal direction is preserved, but intensity resolution fails, suggesting the model captures coarse ordering but not the finer distinctions required for human-like perceptual calibration. The MLLM agent effectively collapses the intermediate categories (Slightly Negative, Neutral, Slightly Positive) toward the polarity extremes. This result helps explain why persona effects remain limited: the sentiment regions where persona-conditioned differences would be expected to emerge are precisely those the model compresses.

The apparent F1 = 1.000 at $\sigma = 5$ across all tasks is an artifact of extreme filtering, not a performance ceiling. As the threshold rises, ambiguous and intermediate cases are progressively excluded: the Neutral class vanishes from P_3, and P_5 retains only four images from the polarity extremes. These perfect scores arise by construction, not by improved model capability, and should not be interpreted as evidence of strong agent capability.

Figure 8 contextualizes these results. It shows confusion matrices across agreement thresholds ($\sigma \in \{3, 4, 5\}$, rows) and problem types (P_2^-, P_2^+, P_3, P_5, columns).

In binary tasks, off-diagonal mass is minimal. In the 3-class setting, Neutral images are systematically mapped to Negative or Positive. In the 5-class case, errors are confined to adjacent categories, with no polarity inversions. This has direct consequences for persona evaluation. The intermediate categories (Slightly Negative, Neutral, Slightly Positive) are precisely where human annotators disagree most and where persona-driven variation would be expected to surface. If the model systematically collapses this region, persona-induced differences in fine-grained perceptual judgment cannot be expressed, regardless of whether they exist.

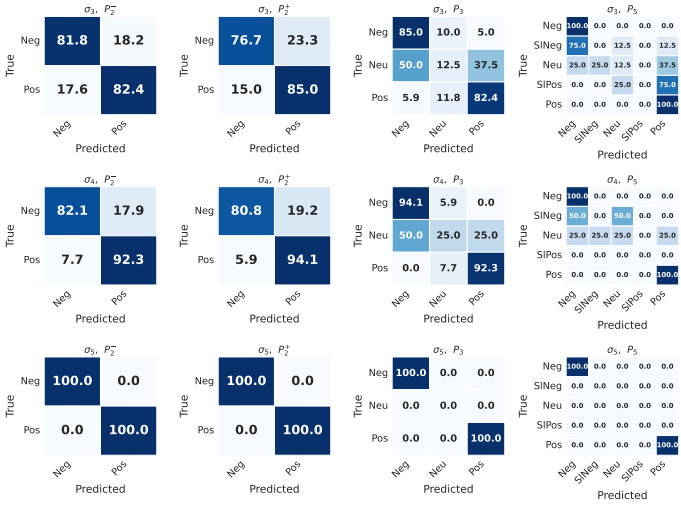


Fig. 8: Pooled out-of-fold confusion matrices across agreement thresholds ($\sigma \in \{3, 4, 5\}$, figure rows) and problem types (P_2^-, P_2^+, P_3, and P_5, figure columns). Row-normalized. Errors in the 5-class remain adjacent labels, consistent with granularity failure rather than polarity inversion. Perfect scores at σ = 5 arise from extreme filtering (see Section IV-F) and should not be interpreted as evidence of strong model capability.

V. Limitations and Ethical Considerations

This study has several limitations. First, persona representations are defined through simple attribute labels — structured key-value injections of gender, economic status, political orientation, and personality — which may oversimplify complex social identities. The limited differentiation observed across certain attributes (e.g., gender and political orientation) may reflect this abstraction rather than the absence of real-world variation. Richer alternatives, such as biographical narratives, psychological inventories, or few-shot behavioral exemplars, may better capture the lived experiences underlying real human perspectives and remain an open direction for future work.

Second, results are based on a single multimodal LLM (Qwen3-VL:8B) and a relatively small image sample (50 images). While the factorial design ensures controlled comparisons, broader datasets and multiple models are required to assess generalizability. Third, the observed extremity bias

suggests that model outputs may not faithfully reproduce human annotation behavior, particularly for fine-grained or intermediate sentiment categories.

From an ethical perspective, the use of synthetic personas raises concerns about misrepresentation and the potential reinforcement of stereotypes. Although persona attributes are used here as experimental controls, such representations should not be interpreted as reflecting real groups. Moreover, deploying LLM-based synthetic annotators in real-world settings may risk amplifying biases if not carefully validated against human data.

These considerations highlight the need for cautious interpretation and for future work incorporating richer contextual grounding and human validation.

VI. Conclusions and Future Directions

This paper investigated whether attribute-based personas induce distinguishable and stable behavior in MLLM agents performing urban image sentiment annotation. Using a controlled factorial design with 1,200 persona-conditioned agents across 24 persona profiles, we evaluated both within-persona convergence and cross-persona differentiation, and compared these results against a no-persona ablation.

Four main findings emerge. First, persona prompts produce strong within-persona convergence, indicating stable and reproducible judgments. Second, cross-persona differentiation is limited: economic status and personality induce statistically detectable but practically modest variation, while gender and political orientation contribute little meaningful separation. Third, the dominant failure mode concerns sentiment granularity rather than polarity discrimination, with agents collapsing intermediate categories despite preserving coarse ordinal structure. Fourth, despite the substantial annotation volume invested in persona conditioning, a five-run modal no-persona prediction matched or exceeded persona-conditioned agreement, suggesting the benefit of persona conditioning in this setting warrants further investigation — though reasoning mode and annotation volume asymmetry should be controlled in future work before drawing definitive conclusions.

Taken together, these findings reveal an important resolution-dependent limitation of simple attribute-based persona prompting: while it supports behavioral stabilization and coarse-grained polarity judgments, it does not generate sufficiently rich variation to capture fine-grained perceptual judgments characteristic of human annotators.

Several directions follow. First, richer persona representations grounded in social context and biographical narratives may better elicit meaningful variation. Second, prompt interventions aimed at mitigating extremity bias may improve calibration for intermediate sentiment categories. Third, broader validation across models, prompting strategies, and datasets is needed, including further study of whether persona prompting contributes diversity beyond accuracy.

More broadly, these findings raise questions about the adequacy of simple attribute-based personas as proxies for human diversity, while suggesting that their value may lie more in stabilizing behavior than in generating rich perceptual variation.

VII. Acknowledgments

This study is partially supported by the National Council for Scientific and Technological Development - CNPq (processes 314603/2023-9, 441444/2023-7, and 444724/2024-9) and INCT TILD-IAR (proc. 408490/2024-1).

References

[1] L. P. Argyle, E. C. Busby, N. Fulda, J. R. Gubler, C. Rytting, and D. Wingate, "Out of one, many: Using language models to simulate human samples," *Political Analysis*, vol. 31, no. 3, p. 337–351, 2023.

[2] G. Aher, R. I. Arriaga, and A. T. Kalai, "Using large language models to simulate multiple humans and replicate human subject studies," in *Proc. of ICML*, (Honolulu, Hawaii, USA), JMLR.org, 2023.

[3] J. S. Park, C. Q. Zou, A. Shaw, B. M. Hill, C. Cai, M. R. Morris, R. Willer, P. Liang, and M. S. Bernstein, "Generative agent simulations of 1,000 people," *arXiv preprint arXiv:2411.10109*, 2024.

[4] T. Beck, H. Schuff, A. Lauscher, and I. Gurevych, "Sensitivity, performance, robustness: Deconstructing the effect of sociodemographic prompting," in *Proc of EACL*, (St. Julian's, Malta), pp. 2589–2615, 2024.

[5] T. Hu and N. Collier, "Quantifying the persona effect in LLM simulations," in *Proc of ACL* (L.-W. Ku, A. Martins, and V. Srikumar, eds.), (Bangkok, Thailand), pp. 10289–10307, ACL, Aug. 2024.

[6] C. J. Li, J. Wu, Z. Mo, A. Qu, Y. Tang, K. I. Zhao, Y. Gan, J. Fan, J. Yu, J. Zhao, *et al.*, "Simulating society requires simulating thought," *ArXiv arXiv:2506.06958*, 2025.

[7] C. A. Bail, "Can generative ai improve social science?," *PNAS*, vol. 121, no. 21, p. e2314021121, 2024.

[8] A. Wang, J. Morgenstern, and J. P. Dickerson, "Large language models that replace human participants can harmfully misportray and flatten identity groups," *Nat Mach Intell*, vol. 7, no. 3, pp. 400–411, 2025.

[9] P. Wang, H. Zou, Z. Yan, F. Guo, T. Sun, Z. Xiao, and B. Zhang, "Not yet: Large language models cannot replace human respondents for psychometric research," *OSF: osf.io/preprints/osf/rwy9b_v1*, 2024.

[10] M. Lutz, I. Sen, G. Ahnert, E. Rogers, and M. Strohmaier, "The prompt makes the person(a): A systematic evaluation of sociodemographic persona prompting for large language models," in *Proc. of EMNLP*, (Suzhou, China), pp. 23212–23237, ACL, Nov. 2025.

[11] C. R. Lopes, R. Minetto, M. R. Delgado, and T. H. Silva, "PerceptSent - exploring subjectivity in a novel dataset for visual sentiment analysis," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, 2023.

[12] W. B. d. Oliveira, L. B. Dorini, R. Minetto, and T. H. Silva, "Outdoorsent: Sentiment analysis of urban outdoor images by using semantic and deep features," *ACM Trans. Inf. Syst.*, vol. 38, Apr. 2020.

[13] A. Filippas, J. J. Horton, and B. S. Manning, "Large language models as simulated economic agents: What can we learn from homo silicus?," in *Proc. of EC*, (New Haven, CT, USA), p. 614–615, ACM, 2024.

[14] W. De Neys and M. Raoelison, "Humans and llms rate deliberation as superior to intuition on complex reasoning tasks," *Communications Psychology*, vol. 3, no. 1, p. 141, 2025.

[15] A. Malik, N. Sabri, M. M. Karnaze, and M. ElSherief, "Are LLMs empathetic to all? investigating the influence of multi-demographic personas on a model's empathy," in *Proc. of EMNLP*, (Suzhou, China), pp. 24938–24959, ACL, Nov. 2025.

[16] C. Barrie and R. Cerina, "Synthetic personas distort the structure of human belief systems," *OSF: osf.io/preprints/socarxiv/n7fq8_v1*, 2026.

[17] N. B. da Silva, J. Harrison, R. Minetto, M. R. Delgado, B. T. Nassu, and T. H. Silva, "Multimodal llms see sentiment," *ArXiv: https://arxiv.org/abs/2508.16873*, 2025.

[18] R. Tourangeau and T. W. Smith, "Asking sensitive questions: The impact of data collection mode, question format, and question context," *Public opinion quarterly*, vol. 60, no. 2, pp. 275–304, 1996.