Impersonating the Crowd: Evaluating LLMs' Ability to Replicate Human Judgment in Misinformation Assessment

David La Barbera
University of Milano-Bicocca
Milan, Italy
david.labarbera@unimib.it

Riccardo Lunardi
University of Udine
Udine, Italy
riccardo.lunardi@uniud.it

Mengdie Zhuang
University of Sheffield
Sheffield, United Kingdom
m.zhuang@sheffield.ac.uk

Kevin Roitero
University of Udine
Udine, Italy
kevin.roitero@uniud.it

Abstract

Large Language Models (LLMs) are increasingly used to replicate human decision-making in subjective tasks. In this work, we investigate whether LLMs can effectively impersonate real crowd workers when evaluating political misinformation statements. We assess (i) the agreement between LLM-generated assessments and human judgments and (ii) whether impersonation skews LLM assessments, impacting accuracy. Using publicly available misinformation assessment datasets, we prompt LLMs to impersonate real crowd workers based on their demographic profiles and evaluate them under the same statements. Through comparative analysis, we measure agreement rates and discrepancies in classification patterns.

Our findings suggest that while some LLMs align moderately with crowd assessments, their impersonation ability remains inconsistent. Impersonation does not uniformly improve accuracy and often reinforces systematic biases, highlighting limitations in replicating human judgment.

CCS Concepts

• **Computing methodologies** → **Natural language generation**;
• **Information systems** → **Crowdsourcing**.

Keywords

Large Language Models, Crowdsourcing, Misinformation

ACM Reference Format:
David La Barbera, Riccardo Lunardi, Mengdie Zhuang, and Kevin Roitero. 2025. Impersonating the Crowd: Evaluating LLMs' Ability to Replicate Human Judgment in Misinformation Assessment . In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR) (ICTIR '25), July 18, 2025, Padua, Italy*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3731120.3744581

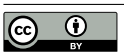
1 Introduction

The remarkable capabilities of LLMs [44] have sparked interest in their potential to replicate human decision-making processes [38]. A particularly promising application is their ability to impersonate human annotators in subjective tasks. Prior work studies LLMs' ability to impersonate individuals' assessments based on provided demographic and ideological information and suggests that LLMs can impersonate specific individuals [3, 5, 11, 15]. However, the extent to which LLMs can faithfully replicate human judgments remains an open question in many domains, particularly in politically sensitive domains such as misinformation assessment.

Crowdsourcing has proven to be a reliable method for performing misinformation detection, providing accurate and robust labels across diverse tasks [1, 19, 29]. A key strength of the crowd lies in its heterogeneity, which allows for the so called "wisdom of the crowd" [14]: crowd workers bring different backgrounds, biases, and perspectives to the task, resulting in a distribution of judgments rather than a single ground truth [33]. This diversity is not a weakness, but a feature, as it reflects the complex nature of truth perception in sociopolitical contexts and multiple works showed that aggregated crowd values are more effective than single judgements [14].

Given this, the potential of LLMs to impersonate crowd workers offers two main advantages. First, if LLMs can match the overall effectiveness of the crowd in assessing veracity, they could potentially serve as scalable and cost-efficient replacements. Second, and more importantly, if LLMs can successfully emulate a range of individual crowd workers, including their biases and variability in judgment, we could leverage them to reproduce the same diversity observed in human annotation. This would allow us to conduct experiments that are both faster and cheaper, while still preserving the richness of crowd-based studies. Ultimately, our work aims to assess whether LLMs can align with crowd assessments in both outcome and variability, thereby enabling their use as proxies in the analysis of polarized domains like political misinformation.

To address this, our study examines the effectiveness of LLMs in impersonating crowd workers when judging the veracity of political statements. Building on prior work in crowdsourced fact-checking [19, 29], we prompt LLMs with demographic data collected from real crowd workers and instruct them to assess the same political statements previously evaluated by the crowd. This setup allows for a systematic analysis of whether LLMs align with human judgments and to what extent they reproduce patterns of crowd assessments—including potential biases and disagreements across different personas.



ICTIR '25, Padua, Italy

ACM ISBN 979-8-4007-1861-8/2025/07
https://doi.org/10.1145/3731120.3744581

Furthermore, we investigate whether the act of impersonation introduces systematic shifts in LLM assessments. Given that LLMs are prompted to impersonate crowd workers, their internal reasoning and output may be subject to deviations from the assessment provided without any impersonation. Understanding these shifts is crucial for assessing the impact of LLM-based impersonation on downstream tasks.

Specifically, we focus on the following Research Questions (RQs):

RQ1 How much agreement is observed between LLM-generated assessments and the judgments of the crowd workers they are trying to impersonate?

RQ2 Does impersonation skew or shift LLM assessments, and what is the effect on accuracy scores?

Our results found that some LLMs show higher agreement with crowd assessments, but fail to capture the diversity of human judgments. In fact, some LLMs tend to provide the same assessments regardless of being instructed to impersonate crowd workers, suggesting limitations in their ability to reproduce individual annotator variability. We released all the data and code used to reproduce our experiments[1].

The remainder of this paper is structured as follows. We first provide the related work of instructing personas in LLMs and using LLMs for misinformation detection in Section 2. In Section 3, we first introduce three crowdsourced datasets used in this study. We also introduce the six LLMs tested and the strategies employed to design the prompts, together with the measurements used to capture the accuracy and bias of the model impersonation. The results with detailed discussions are reported in Section 4. Conclusion and future work are provided in the last section.

2 Background and Related Work

We summarize the literature on instructing personas in LLMs and using LLMs for misinformation detection.

2.1 Personas and LLMs

Researchers have been experimenting with the use of LLMs to impersonate human with different personas through carefully engineered prompts. The personas used to prompt LLMs typically encompass a set of traits such as personality characteristics [4, 5, 17, 22, 25, 40], behaviors such as lifestyle [32], habits [11, 34, 40, 42, 47], perspectives such as political views [2], and demographic attributes such as educational background, occupation, and income level [2, 8, 11, 22]. Conditioning LLMs with personas has showed promise in a variety of use cases, such as diversifying outputs [11], introducing a consistent style for specific tasks [31, 34, 40], and minimizing model hallucinations. However, few studies investigated the ability of LLMs to accurately replicate specific set of personas. For instance, Argyle et al. [2] found that LLMs exhibit high accuracy in predicting U.S. presidential voting preferences of various demographic groups using survey data. Wang et al. [40] demonstrate LLMs can effectively impersonate role-play characters, with human participants confirming that the LLMs generated text aligned with the intended personality traits. Our work contributes to this line of research by detailing the effectiveness of LLMs when impersonating crowd workers in misinformation assessment, using three distinct datasets collected from crowd workers.

2.2 LLMs and misinformation assessment

LLMs have been used in misinformation assessment [7, 9, 22, 27, 36, 46] by performing fact-checking, retrieving evidence to support the assessment, simulate the propagation of fake news, and generating explanations in a narrative style to facilitate understanding. Research in this area has focused primarily on two key goals. One is to improve the ability of LLMs to provide accurate assessments, enabling a cost-effective and reliable way of detecting misinformation [10, 20, 26, 36]. The other is to engineer LLMs to produce human-like reactions, offering insights into how the general population perceives the world and understanding human biases [11] and simulate the propagation of misinformation [22]. Recent work such as DELL [37] and GenFEND [24] leverage LLMs to generate diverse user reactions, or comments based on abstract user profiles, with the goal of enhancing detection performance and simulating user feedback where it is unavailable. Our work contributes to this area of research not only by evaluating the accuracy of LLMs in misinformation detection when impersonating human annotators, but also by examining how their assessments shift compared to those without impersonation and in relation to human judgments.

3 Methodology

In this section, we first introduce three openly available crowdsourced datasets used in this study. These datasets contain crowdworkers' judgment of political statements and demographic data to build persona profiles. We also introduce the six LLMs tested and provide a detailed description of the strategies used to design the prompts. The multiple prompts used in this study are available in the supplementary material. The measurements used to capture the accuracy and bias of the model impersonation are presented in the end.

3.1 Data

We rely on three crowdsourced datasets derived from PolitiFact[2] [39], collected in similar settings across different years. Each dataset includes demographic information from U.S.-based crowd workers, allowing us to replicate worker assessments and directly compare results across datasets.

Each dataset contains 120 statements fact-checked by PolitiFact, labeled with a six-level ordinal scale as ground truth: *Pants on Fire*, *False*, *Mostly False*, *Half True*, *Mostly True*, and *True*. Some examples of statements can be seen in Table 1. Each statement is independently evaluated by 10 workers, resulting in 1,200 individual truthfulness judgments per dataset. Workers are each assigned six statements, one per ground truth level.

In particular, the dataset used in IPM, introduced by La Barbera et al. [19], consists of 1,200 judgments collected using the six-level scale. The SIGIR_6 and SIGIR_100 datasets, both from Roitero et al. [29], include judgments made using a six-level and a hundred-level (0–100) scale, respectively. Roitero et al. [29] also collected an additional set of 1,200 judgments using a three-level scale, which we exclude due to the ambiguity of its middle category.

[1] https://osf.io/es6um/

[2] https://www.politifact.com/

Table 1: Sample of PolitiFact statements.

Speaker	Party	Date	Statement	Ground Truth
Ted Nugent	Republican	June 14, 2022	Three mass shootings were meant to distract from Hillary Clinton controversies.	*Pants on Fire*
Charlie Crist	Democrat	July 17, 2022	The average cost for health insurance in Florida went from about 600 a month for an individual to about 150 a month.	*False*
Doug Holder	Republican	July 13, 2022	Nearly 25 percent of all automobile accidents are caused by texting while driving.	*Mostly False*
Joni Ernst	Democrat	May 28, 2022	3 million per day in your tax dollars are being spent to guard unused border wall materials.	*Half True*
Mike Gallagher	Republican	April 12, 2022	Inflation has gone up every month of the Biden presidency and just hit another 40 year high.	*Mostly True*
Levar Stoney	Democrat	June 16, 2022	In Virginia, Black people are eight times (8X) more likely than white people to die of gun homicide.	*True*

To enable comparison, we binarize both crowd and ground truth labels: for the six-level scale, we group True, Mostly True, and Half True as True; Mostly False, False, and Pants on Fire as False. This binarization preserves the ordinal nature of the scale and reflects common practice. For the hundred-levels scale, we use 50 as the threshold: judgments higher or equal than 50 are True, while the ones lower to 50 are False.

In addition to the truthfulness labels, we use worker-provided demographic data to build persona profiles. Each persona represents one worker and includes its own political ideology, party affiliation, age, education, income, environmental views, and opinions on U.S. border policies. The full dataset includes 600 unique personas, 200 per dataset.

3.2 Large Language Models

We evaluate the effectiveness of several LLMs in impersonating crowd workers for fact-checking tasks, relying on models with varying sizes and from different families. Specifically, we employ Llama-3.1-8B-Instruct, Llama-3.1-70B-Instruct [13], Mistral-7B-v0.3-Instruct, Mistral-Small-Instruct-2409 [16], Gemma-3-4b-it [35] and Qwen2.5-32B-Instruct [28]. For each model and dataset, we build a persona for each worker using collected demographic data. Then, we present the LLM with the same 6 statements assessed by that worker, instructing it to evaluate the truthfulness of the statements while impersonating the worker.

To investigate how the specific formulation of this instruction impacts the LLM's impersonation performance, we design and test multiple distinct prompting strategies. These range from a baseline "*Naive*" approach, which simply instructs the model to act as the described person, to more complex framings. All prompts require the LLM to output a JSON object containing a truthfulness label and a textual explanation justifying the assessment from the assigned persona's perspective. The prompting techniques vary significantly in their approaches. One group focuses on simulating the internal cognitive process behind the judgment (*Internal Monologue*), explicitly asking the LLM to generate the internal reasoning leading to the conclusion [43]. Another set of prompts emphasize adopting a specific viewpoint or mental framework. This group includes the "*Perspective Taking*" [41], "*Cognitive Empathy*" [21] and "*Decision Making*" [23] prompts. Further strategies aim for a more direct or personal identification with the persona. The "*Immersive First-Person*" prompt use first-person language to encourage direct embodiment of the identity [6], while the "*Moral Dilemma*" present the task as a difficult decision influenced by the persona's values and background [30]. Finally, the "*Contextual Persona*" [48] and "*Think Aloud*" [45] prompts contextualize the evaluation within a simulated study setting. By employing this diverse set of prompts, we aim to understand how variations in instruction influence the LLM's ability to replicate the nuanced and often biased judgments of human crowd workers when assessing political information. All the prompts are available in the supplementary material.

We test each prompt on all of the considered models for each dataset and unique worker, thus collecting 162 sets of 1200 model-generated truthfulness assessments that are directly comparable to the annotations made by crowd workers.

3.3 Impersonation Accuracy and Bias

We measure the agreement between models and workers to determine whether impersonation introduces systematic shifts. We quantify agreement as the proportion of model-generated answers that exactly match the corresponding worker's six judgments. Additionally, we compute Krippendorff's α [18] separately for workers and models to evaluate their internal agreement consistently as done in crowdsourcing settings [19, 29].

Moreover, we define bias direction as the difference between worker and model assessments. A positive bias indicates greater leniency, classifying statements as more truthful, while a negative bias suggests a tendency from the LLMs to provide more negative judgments as compared to the crowd. Bias manifests when the model systematically deviates from the workers' judgments, either by assigning more extreme labels than the workers (amplifying their stance) or by providing more moderate or less assertive responses than the workers. Aggregating bias scores across workers allows us to assess whether models consistently lean toward leniency or strictness, revealing potential systematic distortions caused by impersonation. We also measure the accuracy of both LLMs and crowd workers on each dataset by aggregating the 10 judgments per statement using the mean [19, 46] to assess the overall effectiveness of both groups. Complementary, we compute

the accuracy for the base LLM models, without impersonation, to measure any differences in effectiveness.

Finally, we define two measures: correctness and consistency. Correctness is measured separately for assessments produced by crowd workers and LLMs with impersonation, assessing whether each group correctly classifies a statement based on their averaged assessment compared to the ground truth. We first compute the average judgment across the 10 crowd workers for a statement. If the averaged assessment aligns with the ground truth, the crowd is considered correct for that statement. The same process applies to LLMs with impersonation. Consistency measures the variability within the 10 assessments provided per statement, determining whether workers or LLMs assigned identical ratings. Greater variability in ratings indicates lower consistency, while repeated identical ratings suggest higher consistency. For a given statement, if all 10 assessments are identical, we consider the assessments consistent; otherwise, they are deemed inconsistent.

For the crowd low consistency scores are expected, as diverse assessments is the underlying assumption of the "wisdom of the crowd" [14]. On the contrary, high consistency in assessments produced by LLMs with impersonation may indicate rigid decision-making and artificial stability, failing to accurately replicate human judgment patterns.

4 Results

4.1 RQ1: Do LLMs Accurately Impersonate Crowd Worker Judgments?

4.1.1 Agreement. To answer **RQ1**, we first look at the agreement between models and worker assessments. Our analysis found that agreement varies significantly across models and datasets. We analyze worker–model agreement across all *(model, prompt, dataset)* combinations, computing the mean and standard deviation of per-worker agreement scores. The results show that certain prompts tend to foster higher alignment between model responses and human judgments. For instance, with the prompt *Internal Monologue*, models like Llama-3.1-70B and Mistral-Small achieved average agreement scores of 0.62 and 0.63, respectively (standard deviation $\approx$ 0.22), while *Naive* consistently produced high agreement across models, with Llama-3.1-8B reaching 0.62 (std = 0.20) and Mistral-Small reaching 0.60 (std = 0.21).

In contrast, prompts such as *Cognitive Empathy* and *Contextual Persona* yielded lower agreement: Qwen2.5-32B scored 0.39 and 0.39 respectively (std $\approx$ 0.26), and Mistral-Small scored 0.40 and 0.39 (std $\approx$ 0.26), indicating greater ambiguity or disagreement among annotators.

Model-wise, Llama-3.1-70B showed more stable performance, with agreement scores consistently above 0.60 for easier prompts, while smaller models like Qwen2.5-32B and Mistral-7B often dropped below 0.45 for several prompts, suggesting a gap in interpretability or alignment with human reasoning. Overall, these results highlight the significant impact of both prompt framing and model scale on alignment with human assessments.

To test for statistical significance on which model and prompt are more effective, we applied Tukey's HSD test to compare per-worker agreement scores across all *(model, prompt)* pairs. Out of 1,431 pairwise comparisons, 836 (58.42%) were statistically significant at $p = 0.05$, highlighting substantial variability across conditions. Interestingly, smaller models such as Qwen2.5-32B and Gemma-3-4b accounted for the largest number of significant differences (164 and 162 respectively), outperforming other models in many comparisons. Prompts eliciting the most consistent differences included *Decision Making* (149 significant wins), *Cognitive Empathy* (133), and *Contextual Persona* (133), suggesting these scenarios provoke stronger model-to-model disagreement or variability in human-model alignment.

The highest differences in agreement scores (up to $\Delta = 0.2338$) consistently involved Mistral-Small under the *Internal Monologue* prompt, which outperformed or underperformed significantly depending on the comparison, indicating that this configuration yields either very high or very low alignment depending on context. Overall, these results suggest that both prompt type and model architecture significantly impact the consistency of human agreement with model-generated answers.

Next, we analyze the distribution of agreement scores between LLMs and crowd workers, as shown in Figure 1. The figure presents the relative frequency of agreement values, computed between impersonated LLM outputs and individual crowd judgments, across different prompt formulations (columns) and models (rows). Each curve corresponds to a different dataset, enabling us to disentangle model- and prompt-specific effects from dataset-specific trends. We report the results for three models and six representative prompts, selected to provide a more concise overview. The remaining combinations were excluded for brevity, as they showed similar trends.

Across the plots, we observe that model behavior has a stronger influence on the agreement distribution than the dataset content. The distributions for each dataset tend to follow similar shapes within the same model, suggesting that the way in which a model is instructed (via prompts) and the model's own architecture and capabilities have a greater effect on alignment with the crowd than the particular statements or domains used for evaluation.

Among the models, Llama-3.1-70B stands out as the most aligned with human judgment. Its agreement distributions are consistently peaked in the mid-to-high range (typically between 0.5 and 0.7), especially under prompts like Internal Monologue, Perspective Taking, and Contextual Persona. These sharp, concentrated peaks suggest that Llama-3.1-70B not only matches the average crowd judgment more frequently, but does so with low variance across statements. Its scores are robust across datasets and appear relatively insensitive to minor prompt variations, reflecting strong and stable impersonation capabilities.

Qwen2.5-32B also exhibits strong alignment, although its agreement patterns are more sensitive to prompt formulation. Under prompts such as Decision Making and Contextual Persona, Qwen2.5-32B shows high concentrations in the 0.6–0.7 range, indicating confident and consistent alignment with crowd workers. However, for more open-ended prompts like Think Aloud, the agreement distribution becomes flatter, revealing some instability. This suggests that while Qwen2.5-32B is capable of high agreement, its impersonation fidelity depends more heavily on the structure and clarity of the prompt.

Mistral-7B shows the most uniform behavior across prompts but also the most moderate alignment. Its distributions are broader and centered around 0.4–0.6, indicating that it tends to mirror

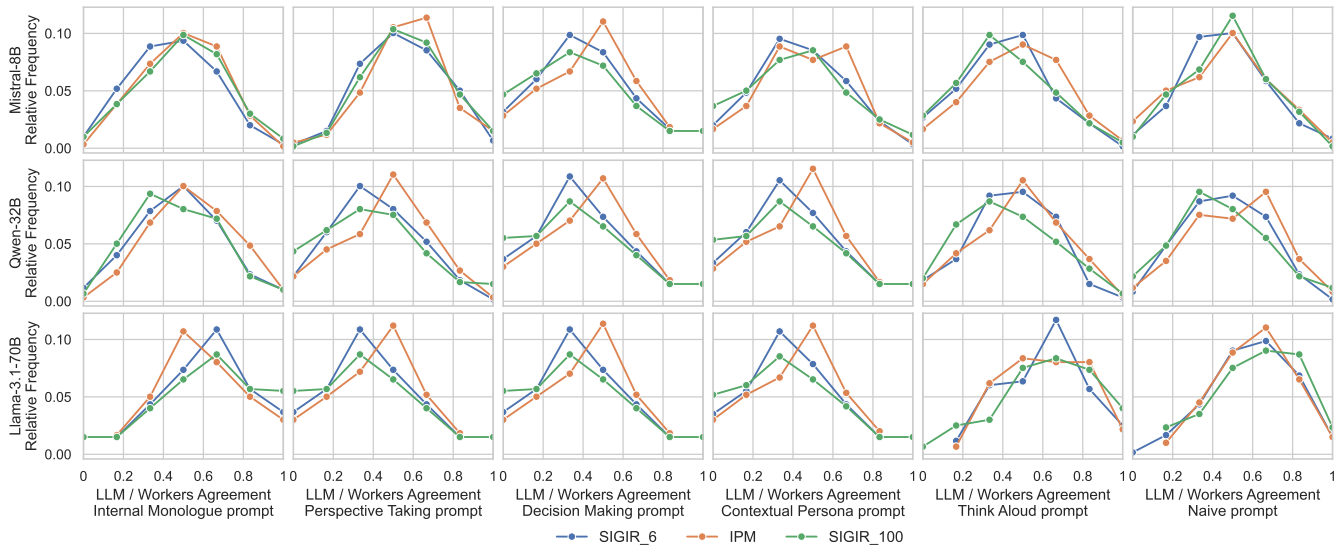


Figure 1: Distribution of agreement scores between a sample of models (rows) and crowd workers per prompt (columns).

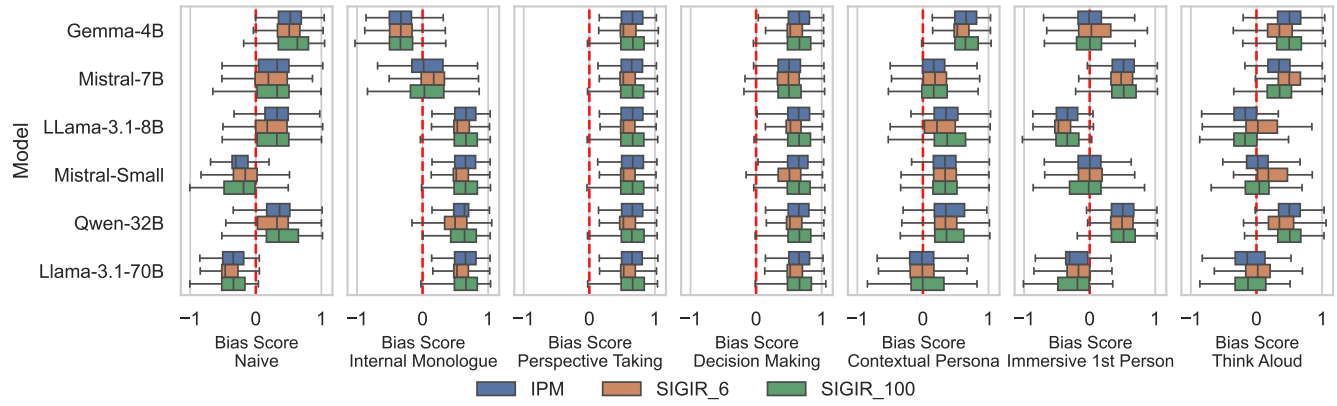


Figure 2: Bias direction for each model and dataset per prompt (columns).

the average crowd opinion without showing particularly strong confidence. Unlike Qwen2.5-32B, its performance is less sensitive to prompt type, but it rarely achieves the higher agreement rates seen in Llama-3.1-70B. This balance suggests that Mistral-7B is more general-purpose but less effective for high-fidelity impersonation of individual human judgments.

The impact of prompt formulation is also clearly visible in the plots. Prompts like *Internal Monologue* and *Perspective Taking* generally produce more peaked and consistent agreement distributions, especially for the stronger models. These prompts appear to encourage more structured, cognitively coherent responses that better align with how human annotators approach truthfulness tasks. Conversely, prompts such as *Naive* and *Think Aloud* often lead to flatter or more dispersed distributions, particularly for Mistral-7B and Qwen2.5-32B, indicating more variability and reduced alignment. This demonstrates that the clarity and structure of the prompt are essential in guiding LLMs toward human-like judgment patterns.

While dataset-specific patterns are present to some extent, they are far less prominent than model and prompt effects. All three datasets yield broadly similar agreement distributions when paired with the same model and prompt, reinforcing the conclusion that impersonation quality is primarily a function of model capacity and task framing rather than statement content. We performed pairwise statistical comparisons between all (*model*, *prompt*) combinations by analyzing the corresponding agreement distributions pooled over all datasets. For each pair, we applied the Mann–Whitney U test to determine whether the agreement values differed significantly. To account for multiple comparisons, we applied Bonferroni correction with $p = 0.05$. Out of 1,431 total comparisons, 859 (60.03%) were found to be statistically significant after correction.

To further understand the consistency of these impersonations, we computed Krippendorff's α to assess consistency among model answers across different crowd workers, and compared this to the agreement among the crowd themselves. As shown in Table 2, model impersonations under certain prompting strategies achieve much higher agreement than that observed among crowd workers: the average α for the crowd is 0.135 (IPM), 0.092 (SIGIR_100), and 0.094 (SIGIR_6), indicating very low inter-rater reliability. In contrast, models prompted with strategies such as *Perspective Taking* and *Decision Making* consistently reach $\alpha > 0.9$ across nearly all models, demonstrating a high degree of behavioral stability. This suggests that prompting can strongly constrain model behavior across crowd-provided contexts, potentially exceeding the internal consistency of human workers. On the other hand, prompts like *Cognitive Empathy* lead to low (and sometimes negative) agreement, especially for Qwen and Gemma, indicating that some strategies introduce more variance in model outputs than others.

4.1.2 Bias Direction. Figure 2 presents the direction and magnitude of model bias across prompts (columns), datasets (colors), and models (rows). Each boxplot represents the distribution of bias scores for a specific model-dataset combination under a given prompt. A bias score of zero indicates that the model's average judgment matches that of the crowd workers. Positive scores indicate leniency, where the model is more likely to label a statement as truthful than the average worker, while negative scores denote stricter evaluations. We excluded the results for three prompts to enhance plot readability, as they exhibited similar patterns to those reported.

Several patterns emerge from this plots. First, across most prompts, we see that the bias distributions are centered close to zero for many models, with limited deviation in the median score. In particular, Llama-3.1-70B, Qwen-32B, and Mistral-Small show relatively balanced behavior, often producing judgments that, on average, closely align with the crowd. This suggests that these models are generally well-calibrated and capable of emulating crowd behavior without introducing systematic skew in either direction.

However, variability across prompts and datasets reveals important differences. For instance, the *Naive* and *Think Aloud* prompts tend to produce more dispersed bias distributions, especially for smaller models such as Llama-3.1-8B and Gemma-4B. This indicates that when the instructions are under-specified or open-ended, models diverge more from crowd consensus, and do so inconsistently across statements. In contrast, prompts like *Internal Monologue*, *Perspective Taking*, and *Contextual Persona* produce tighter, more symmetric distributions, suggesting that structured and introspective prompts improve model stability and alignment.

From a dataset perspective, SIGIR_6 (orange) consistently shows narrower interquartile ranges compared to IPM and SIGIR_100, especially for well-performing models. This may suggest that the shorter and more controlled nature of SIGIR_6 statements yields more predictable model behavior and closer alignment with crowd judgments. Conversely, the broader ranges observed on IPM and SIGIR_100 might reflect greater statement diversity or subjectivity in the original crowd responses.

One of the more noticeable outliers is Mistral-7B, which shows a slight tendency toward leniency across several prompts and datasets, as its median bias values are often shifted above zero. Meanwhile, Gemma-4B displays larger interquartile ranges and higher variability across all prompts, indicating instability and inconsistent alignment. This reinforces the importance of model scale and architecture when attempting to simulate nuanced human evaluation tasks.

Interestingly, Llama-3.1-70B, while not uniformly centered around zero, maintains tight and symmetric distributions across most conditions, reaffirming its robustness not only in agreement (as previously seen) but also in bias behavior. It is the only model that consistently avoids extreme deviations across prompts and datasets. Finally, to conclude our analysis, we conducted a pairwise statistical comparisons of the bias scores across all *(Model, Prompt)* combinations, pooling data from all datasets. After applying Bonferroni correction with $p = 0.05$, 72.40% of the comparisons were found to be statistically significant, indicating substantial differences in bias levels between many configurations.

Overall, these results indicate that while large LLMs can produce judgments that are, on average, unbiased relative to the crowd, their consistency is affected by prompt formulation, dataset characteristics, and model capacity. Prompts that engage cognitive framing and perspective-taking (e.g., *Internal Monologue*, *Decision Making*) tend to elicit more balanced outputs. Dataset variability plays a role as well, with more heterogeneous collections leading to wider dispersion in model bias. These findings highlight that faithful crowd impersonation is not only about matching central tendencies but also about replicating the nuanced fluctuations and contextual sensitivities that characterize human annotations.

4.2 RQ2: Does Impersonation Skew LLMs Assessments and Affect Accuracy?

We now turn to **RQ2**. Table 3 reports the accuracy scores of both crowd workers and LLMs across three datasets, IPM, SIGIR_6, and SIGIR_100, under two settings: base (B), where models operate without persona conditioning, and impersonation (P), where models are prompted to simulate individual crowd worker personas. For impersonation, we report the mean and standard deviation of accuracy across all personas used.

Starting with the crowd worker baselines, we observe that the highest accuracy is achieved by the crowd on IPM (0.82), followed by SIGIR_100 (0.68), and finally SIGIR_6 (0.62). These values reflect the expected reliability of aggregated human judgments, particularly for datasets with more diverse content such as IPM. None of the LLMs in the impersonation condition manage to outperform the crowd on any dataset, underscoring the difficulty of fully replicating human-level accuracy through persona-based simulation.

Examining model effectiveness across conditions reveals a consistent pattern: impersonation often results in lower accuracy compared to the base setting. For instance, Llama-3.1-8B drops from 0.57 to an average of 0.53 on IPM, and from 0.58 to 0.52 on both SIGIR_6 and SIGIR_100. Similarly, Llama-3.1-70B shows a substantial decline on IPM, from 0.72 in the base setting to just 0.58 when impersonating workers. On the remaining two datasets, it also decreases from 0.58 to 0.54, indicating that even the most capable model in our pool does not benefit from persona conditioning in terms of overall classification accuracy.

Table 2: Average internal agreement Krippendorff's α per each model and prompting strategy, averaged across the three datasets.

Prompt	Gemma-4B	Mistral-7B	Llama-3.1-8B	Mistral-Small-24B	Qwen-32B	Llama-3.1-70B
Internal Monologue	0.27	0.53	0.38	0.29	0.48	**0.74**
Perspective Taking	0.62	0.70	**1.00**	0.16	0.53	**1.00**
Cognitive Empathy	0.00	**0.72**	0.62	0.27	0.04	0.04
Decision Making	**1.00**	0.68	**1.00**	**1.00**	0.68	0.67
Contextual Persona	0.12	**0.76**	0.39	0.40	0.22	0.23
Moral Dilemma	0.64	**0.81**	0.46	0.50	0.74	0.62
Immersive 1st Person	**1.00**	0.63	0.46	0.51	0.57	0.39
Think Aloud	**0.78**	0.72	0.10	0.53	0.66	0.44
Naive	0.66	0.75	0.55	0.58	**0.79**	0.65

Table 3: Accuracy comparison between crowd workers and models. For each model, accuracy scores are reported for the base model (B) and the model impersonating workers (P). For P, we report the average accuracy obtained across multiple prompting strategies, along with the standard deviation.

Dataset	Crowd	Gemma-3-4B		Mistral-7B		Llama-3.1-8B		Mistral-Small-24B		Qwen-2.5-32B		Llama-3.1-70B	
		B	P	B	P	B	P	B	P	B	P	B	P
IPM	0.82	0.55	(0.53, 0.04)	0.54	(0.57, 0.05)	0.57	(0.53, 0.05)	0.62	(0.58, 0.09)	0.57	(0.55, 0.05)	0.72	(0.58, 0.11)
SIGIR_6	0.62	0.51	(0.52, 0.02)	0.53	(0.54, 0.04)	0.58	(0.52, 0.04)	0.64	(0.56, 0.07)	0.61	(0.53, 0.03)	0.58	(0.54, 0.07)
SIGIR_100	0.68	0.50	(0.51, 0.02)	0.53	(0.54, 0.04)	0.58	(0.52, 0.04)	0.63	(0.55, 0.06)	0.57	(0.54, 0.04)	0.58	(0.54, 0.06)

A similar downward trend is observed for Mistral-Small, which in earlier versions of our experiments had shown promise under impersonation. Here, its accuracy drops in all three datasets, most notably from 0.64 to 0.56 on SIGIR_6. Mistral-7B shows marginal change, with a small gain on IPM (from 0.54 to 0.57) but no meaningful improvement elsewhere. Qwen and Gemma both follow the same pattern, with impersonation leading to lower average scores across the scores.

These findings challenge the assumption that simulating individual human perspectives via persona conditioning necessarily leads to better task performance. In our setting, impersonation tends to introduce noise, perhaps due to the added complexity of simulating diverse beliefs, backgrounds, or political stances. The resulting predictions may become less calibrated and more variable, contributing to a drop in accuracy.

An especially noteworthy finding comes from comparing SIGIR_6 and SIGIR_100. These datasets share the exact same set of factual statements but differ in the persona profiles assigned to the models. Despite this difference, the impersonation results for both datasets are strikingly similar, not just in average accuracy but also in standard deviation across prompts and models. This suggests that the particular choice of persona, at least within the range of those represented in our crowd worker data, has minimal effect on the final model prediction. In other words, changing the persona does not substantially alter the model's truthfulness classification, which raises important questions about how and when persona conditioning meaningfully affects outputs.

Taken together, these results suggest that impersonation, while an intriguing and increasingly popular technique for modeling individual-level subjectivity, does not consistently enhance factual accuracy in truthfulness assessment tasks. Instead, its effectiveness appears highly dependent on both the underlying model architecture and the dataset characteristics. Moreover, the finding that accuracy distributions remain stable even when persona profiles change implies that persona conditioning may have a more modest impact than previously assumed, at least when applied to tasks that rely on factual correctness rather than subjective interpretation. These insights raises question in adopting persona-based prompting as a default strategy for improving alignment with human judgments. While there may still be value in simulating demographic or ideological diversity to understand biases or explore disagreement patterns, our results show that such impersonation does not straightforwardly translate into better performance on objective truthfulness classification tasks.

We now examine the relationship between correctness and consistency in the assessments produced by both crowd workers and LLMs. Figure 3 presents a series of heatmaps, one for each model and dataset combination. Each cell counts the number of statements falling into one of four possible categories, defined by whether the assessment is correct (x-axis) and consistent (y-axis). Rows correspond to datasets, while columns correspond to annotators—starting with the crowd and followed by each LLM. We report a subset of LLMs because results for the others are very similar.

In the first column, representing crowd workers, we observe a prominent cluster in the top-right quadrant: statements that are both correct and inconsistent. This pattern reflects the natural variability of human annotation—crowd workers frequently arrive at

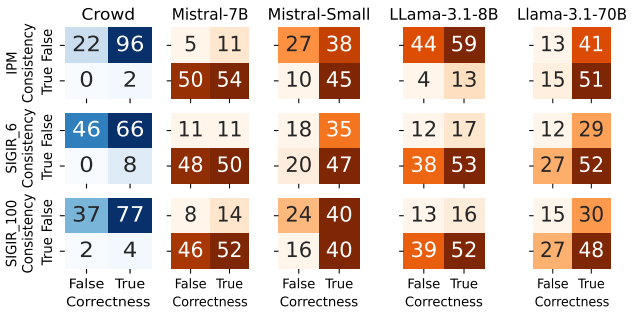


Figure 3: Comparing correctness and consistency in statement assessments produced by crowd workers and models.

correct conclusions, but their paths to these judgments vary substantially. This observation is consistent with prior work showing that human evaluators bring diverse perspectives to subjective tasks [12].

By contrast, the LLMs display markedly different consistency-correctness dynamics. For instance, Llama-3.1-70B (last column) tends to concentrate its predictions in the bottom-right quadrant, indicating that its assessments are mostly correct and consistent. This reflects a more stable and deterministic decision-making process compared to human annotators. However, this same model—and others like Llama-3.1-8B and Mistral-7B —also produce a sizable portion of statements in the bottom-left quadrant, meaning that their mistakes are made consistently. These systematic errors may reflect biases introduced by impersonation or limitations in the model's reasoning capabilities.

Mistral-Small (third column) and Llama-3.1-8B on the IPM dataset (fourth column, top row) exhibit a different failure mode. They show elevated counts in the top-left quadrant, corresponding to assessments that are incorrect and inconsistent. This suggests a lack of both reliability and accuracy, as these models frequently disagree with the ground truth and are unstable across prompts or persona conditions.

Taken together, these results highlight a key contrast between human and LLM behavior. While human crowd workers tend to be correct but inconsistent, mirroring the subjective, context-sensitive nature of truthfulness assessments, LLMs are often consistent, regardless of whether their predictions are right or wrong. This pattern implies that LLMs impose an artificial uniformity on their assessments, a trait that diverges from how real humans behave in such tasks. Effective impersonation, however, should not only capture the average judgment but also the variability inherent in human annotation. The failure to reflect this inconsistency suggests that while LLMs can simulate correctness, they fall short of mimicking the full spectrum of human reasoning patterns.

5 Conclusion and Future Work

In this work, we investigated whether LLMs can effectively impersonate crowd workers in the context of political misinformation assessment. Our study focused on two central research questions: the extent to which LLMs align with human assessments (**RQ1**), and the impact of impersonation on model performance (**RQ2**). By simulating individual crowd worker personas across multiple prompts, datasets, and models, we aimed to understand not just whether LLMs can approximate human judgments, but whether they can also reflect the variability and inconsistency that characterize real-world annotation.

With respect to **RQ1**, our findings show that agreement between LLM-generated and crowd-sourced truthfulness judgments varies substantially across models and datasets. Among the models tested, Llama-3.1-70B and Mistral-Small exhibit the highest overall agreement with human annotators. However, closer inspection of the agreement distributions reveals that LLMs tend to produce more stable and narrowly peaked patterns compared to the broader, more dispersed agreement curves observed among the crowd. This is further supported by our analysis of internal consistency: while crowd workers display low Krippendorff's α, reflecting their natural diversity of perspectives, LLMs exhibit much higher internal agreement, indicating that their impersonations impose a kind of artificial uniformity not present in actual human behavior.

Turning to **RQ2**, we observe that adopting personas does not reliably lead to improved model accuracy. In many cases, persona-conditioned models perform worse than their base counterparts, particularly on datasets like IPM and SIGIR_100. Although some moderate gains are observed in isolated cases, these are not consistent across models or datasets. Moreover, the similarity in performance between SIGIR_6 and SIGIR_100 —datasets sharing the same statements but differing in persona profiles—suggests that impersonation may have limited impact on actual assessment outcomes.

These results collectively suggest that while LLMs are capable of producing plausible impersonations of individual workers in terms of style and structure, they fall short of replicating the more nuanced dimensions of human annotation namely, inconsistency, disagreement, and ambiguity. Impersonation, as currently implemented through prompt conditioning, encourages the model to behave in a more deterministic and rigid fashion, which may run counter to the very qualities that make human annotation valuable, especially in subjective or politically sensitive tasks.

Future work should focus on developing more expressive and flexible impersonation strategies that better capture the subtleties of human behavior. This could involve prompt designs that explicitly model uncertainty or conflict, fine-tuning on real crowd-labeled data to incorporate annotation noise and diversity, or leveraging ensemble methods to simulate disagreement. In addition, extending this line of research to other domains such as medical or legal annotation and across different cultural or linguistic contexts would provide deeper insights into the generalizability and robustness of LLM-based impersonation.

Acknowledgments

This work was partially supported by the ESF+ 2021/2027 Regional Program of the Autonomous Region Friuli Venezia Giulia (PPO 2023, Program No. 22/23 – LINE A: PhD programmes).

References

[1] Jennifer Allen, Antonio A. Arechar, Gordon Pennycook, and David G. Rand. 2021. Scaling up fact-checking using the wisdom of crowds. *Science Advances* 7, 36 (2021). https://doi.org/10.1126/sciadv.abf4393

[2] Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis* 31, 3 (2023), 337–351.

[3] Pietro Bernardelle, Leon Fröhling, Stefano Civelli, Riccardo Lunardi, Kevin Roitero, and Gianluca Demartini. 2024. Mapping and Influencing the Political Ideology of Large Language Models using Synthetic Personas. arXiv:2412.14843 [cs.CL] https://arxiv.org/abs/2412.14843

[4] Graham Caron and Shashank Srivastava. 2023. Manipulating the Perceived Personality Traits of Language Models. (Dec. 2023), 2370–2386. https://doi.org/10.18653/v1/2023.findings-emnlp.156

[5] Lucio La Cava and Andrea Tagarelli. 2024. Open Models, Closed Minds? On Agents Capabilities in Mimicking Human Personalities through Open Large Language Models. arXiv:2401.07115 [cs.AI] https://arxiv.org/abs/2401.07115

[6] Sijie Cheng, Zhicheng Guo, Jinawen Wu, Kechen Fang, Peng Li, Huaping Liu, and Yang Liu. 2024. EgoThink: Evaluating First-Person Perspective Thinking Capability of Vision-Language Models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 14291–14302. https://doi.org/10.1109/CVPR52733.2024.01355

[7] Anubrata Das, Houjiang Liu, Venelin Kovatchev, and Matthew Lease. 2023. The state of human-centered NLP technology for fact-checking. *Information processing & management* 60, 2 (2023), 103219.

[8] Angel Felipe Magnossão de Paula, J Shane Culpepper, Alistair Moffat, Sachin Pathiyan Cherumanal, Falk Scholer, and Johanne Trippas. 2025. The Effects of Demographic Instructions on LLM Personas. *arXiv preprint arXiv:2505.11795* (2025).

[9] Gianluca Demartini, Stefano Mizzaro, and Damiano Spina. 2020. Human-in-the-loop Artificial Intelligence for Fighting Online Misinformation: Challenges and Opportunities. *IEEE Data Eng. Bull.* 43, 3 (2020), 65–74.

[10] Xishuang Dong, Shouvon Sarker, and Lijun Qian. 2022. Integrating human-in-the-loop into swarm learning for decentralized fake news detection. In *2022 International Conference on Intelligent Data Science Technologies and Applications (IDSTA)*. IEEE, 46–53.

[11] Leon Fröhling, Gianluca Demartini, and Dennis Assenmacher. 2024. Personas with Attitudes: Controlling LLMs for Diverse Data Annotation. *arXiv preprint arXiv:2410.11745* (2024).

[12] Leon Fröhling, Gianluca Demartini, and Dennis Assenmacher. 2024. Personas with Attitudes: Controlling LLMs for Diverse Data Annotation. arXiv:2410.11745 [cs.CL] https://arxiv.org/abs/2410.11745

[13] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, and Alan Schelten ... Zhiyu Ma. 2024. The Llama 3 Herd of Models. arXiv:2407.21783 [cs.AI] https://arxiv.org/abs/2407.21783

[14] Jeff Howe. 2006. The rise of crowdsourcing. *Wired Magazine* 14, 6 (2006), 1–4. https://www.wired.com/2006/06/crowds/

[15] Tiancheng Hu and Nigel Collier. 2024. Quantifying the persona effect in llm simulations. *arXiv preprint arXiv:2402.10811* (2024).

[16] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. arXiv:2310.06825 [cs.CL] https://arxiv.org/abs/2310.06825

[17] Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2024. Evaluating and inducing personality in pre-trained language models. *Advances in Neural Information Processing Systems* 36 (2024).

[18] Klaus Krippendorff. 2008. Computing Krippendorff`s Alpha-Reliability. *UPENN Libraries* 1 (2008), 43. https://repository.upenn.edu/asc_papers/43

[19] David La Barbera, Eddy Maddalena, Michael Soprano, Kevin Roitero, Gianluca Demartini, Davide Ceolin, Damiano Spina, and Stefano Mizzaro. 2024. Crowdsourced Fact-checking: Does It Actually Work? *Information Processing & Management* 61, 5 (2024), 103792. https://doi.org/10.1016/j.ipm.2024.103792

[20] David La Barbera, Kevin Roitero, Stefano Mizzaro, et al. 2022. A Hybrid Human-In-The-Loop Framework for Fact Checking.. In *NL4AI@ AI* IA*. 13–23.

[21] Yoon Kyung Lee, Inju Lee, Minjung Shin, Seoyeon Bae, and Sowon Hahn. 2023. Chain of Empathy: Enhancing Empathetic Response of Large Language Models Based on Psychotherapy Models. *ArXiv* abs/2311.04915 (2023).

[22] Yuhan Liu, Xiuying Chen, Xiaoqing Zhang, Xing Gao, Ji Zhang, and Rui Yan. 2024. From skepticism to acceptance: simulating the attitude dynamics toward fake news *(IJCAI '24)*. Article 873, 9 pages. https://doi.org/10.24963/ijcai.2024/873

[23] Manikanta Loya, Divya Sinha, and Richard Futrell. 2023. Exploring the Sensitivity of LLMs' Decision-Making Capabilities: Insights from Prompt Variations and Hyperparameters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 3711–3716. https://doi.org/10.18653/v1/2023.findings-emnlp.241

[24] Qiong Nan, Qiang Sheng, Juan Cao, Beizhe Hu, Danding Wang, and Jintao Li. 2024. Let Silence Speak: Enhancing Fake News Detection with Generated Comments from Large Language Models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management* (Boise, ID, USA) *(CIKM '24)*. Association for Computing Machinery, New York, NY, USA, 1732–1742. https://doi.org/10.1145/3627673.3679519

[25] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 2, 22 pages. https://doi.org/10.1145/3586183.3606763

[26] Yunke Qu, Kevin Roitero, David La Barbera, Damiano Spina, Stefano Mizzaro, and Gianluca Demartini. 2022. Combining human and machine confidence in truthfulness assessment. *ACM Journal of Data and Information Quality* 15, 1 (2022), 1–17.

[27] Dorian Quelle and Alexandre Bovet. 2024. The perils and promises of fact-checking with large language models. *Frontiers in Artificial Intelligence* 7 (2024), 1341697.

[28] Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 Technical Report. arXiv:2412.15115 [cs.CL] https://arxiv.org/abs/2412.15115

[29] Kevin Roitero, Michael Soprano, Shaoyang Fan, Damiano Spina, Stefano Mizzaro, and Gianluca Demartini. 2020. Can The Crowd Identify Misinformation Objectively? The Effects of Judgment Scale and Assessor's Background. In *Proceedings of the 43st International ACM SIGIR Conference on Research and Development in Information Retrieval*. (Xi'an, China (Virtual)) *(SIGIR '20)*. Association for Computing Machinery, 439–448. https://doi.org/10.1145/3397271.3401112

[30] Nino Scherrer, Claudia Shi, Amir Feder, and David M. Blei. 2023. Evaluating the Moral Beliefs Encoded in LLMs. arXiv:2307.14324 [cs.CL] https://arxiv.org/abs/2307.14324

[31] Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature* 623, 7987 (2023), 493–498.

[32] Yimin Shi, Yang Fei, Shiqi Zhang, Haixun Wang, and Xiaokui Xiao. 2025. You Are What You Bought: Generating Customer Personas for E-commerce Applications. *arXiv preprint arXiv:2504.17304* (2025).

[33] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast—but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (Honolulu, Hawaii) *(EMNLP '08)*. Association for Computational Linguistics, USA, 254–263. https://aclanthology.org/D08-1027/

[34] Haoyu Song, Wei-Nan Zhang, Jingwen Hu, and Ting Liu. 2020. Generating persona consistent dialogues by exploiting natural language inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 8878–8885.

[35] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, et al. 2025. Gemma 3 Technical Report. arXiv:2503.19786 [cs.CL] https://arxiv.org/abs/2503.19786

[36] Venktesh V, Abhijit Anand, Avishek Anand, and Vinay Setty. 2024. QuanTemp: A real-world open-domain benchmark for fact-checking numerical claims. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 650–660. https://doi.org/10.1145/3626772.3657874

[37] Herun Wan, Shangbin Feng, Zhaoxuan Tan, Heng Wang, Yulia Tsvetkov, and Minnan Luo. 2024. DELL: Generating Reactions and Explanations for LLM-Based Misinformation Detection. In *Findings of the Association for Computational Linguistics ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand and virtual meeting, 2637–2667. https://aclanthology.org/2024.findings-acl.155

[38] Pengda Wang, Zilin Xiao, Hanjie Chen, and Frederick L. Oswald. 2024. Will the Real Linda Please Stand up...to Large Language Models? Examining the Representativeness Heuristic in LLMs. In *Proceedings of the Conference on Language Models (COLM)*. https://arxiv.org/abs/2404.01461

[39] William Yang Wang. 2017. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Regina Barzilay and Min-Yen Kan (Eds.), Vol. 4. Association for Computational Linguistics, Vancouver, Canada, 422–426. https://doi.org/10.18653/v1/P17-2067

[40] Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, et al. 2024. Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 1840–1873.

[41] Alex Wilf, Sihyun Lee, Paul Pu Liang, and Louis-Philippe Morency. 2024. Think Twice: Perspective-Taking Improves Large Language Models' Theory-of-Mind Capabilities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 8292–8308. https://doi.org/10.18653/v1/2024.acl-long.451

[42] Chen Xu, Piji Li, Wei Wang, Haoran Yang, Siyun Wang, and Chuangbai Xiao. 2022. Cosplay: Concept set guided personalized dialogue generation across both party personas. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 201–211.

[43] Diji Yang, Jinmeng Rao, Kezhen Chen, Xiaoyuan Guo, Yawen Zhang, Jie Yang, and Yi Zhang. 2024. IM-RAG: Multi-Round Retrieval-Augmented Generation Through Learning Inner Monologues. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 730–740. https://doi.org/10.1145/3626772.3657760

[44] Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. 2024. Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond. *ACM Trans. Knowl. Discov. Data* 18, 6, Article 160 (April 2024), 32 pages. https://doi.org/10.1145/3649506

[45] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. arXiv:2210.03629 [cs.CL] https://arxiv.org/abs/2210.03629

[46] Xia Zeng, David La Barbera, Kevin Roitero, Arkaitz Zubiaga, and Stefano Mizzaro. 2024. Combining Large Language Models and Crowdsourcing for Hybrid Human-AI Misinformation Detection. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2332–2336.

[47] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too?. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Iryna Gurevych and Yusuke Miyao (Eds.). Association for Computational Linguistics, Melbourne, Australia, 2204–2213. https://doi.org/10.18653/v1/P18-1205

[48] 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*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 15126–15154. https://doi.org/10.18653/v1/2024.findings-emnlp.888