”Has 'identity contagion' or 'persona absorption' been observed between different large language models?”

Studies have found no evidence of direct persona transfer between language models, despite observing individual persona absorption and cross-model prompt effects.

Abstract

Large language models adopt assigned personas when prompted. Experimental studies show that persona-modulating cues change outputs markedly—one study reports a 42.5% increase in harmful completions in GPT-4, while others note up to a 33% drop in Theory of Mind performance and shifts in social biases. Five studies document robust persona absorption within models, and two report that adversarial prompts produce similar effects across different models. Interactions between agents sometimes yield linguistic and personality alignment, yet none of the studies provide direct evidence that a model adopts another's persona solely through interaction—that is, full identity contagion remains unobserved.

Key points include:

1. Persona absorption occurs reliably within individual models.
2. Persona-inducing prompts can transfer effects (e.g., jailbreak capabilities) across models.
3. Agent interactions lead to alignment of linguistic and personality traits without clear persona transfer.

Paper search

Using your research question ””Has 'identity contagion' or 'persona absorption' been observed between different large language models?””, we searched across over 126 million academic papers from the Semantic Scholar corpus. We retrieved the 50 papers most relevant to the query.

Screening

We screened in papers that met these criteria:

- **Multiple Model Interaction**: Does the study examine direct interactions between two or more distinct large language models?
- **Behavioral Outcomes**: Does the study document specific behavioral or output outcomes from the model-to-model interactions?
- **Observable Changes**: Does the research demonstrate measurable changes in model responses or behaviors following exposure to other AI models?
- **Empirical Evidence**: Does the study present experimental data or empirical evidence of model-to-model influence patterns?
- **Transfer Mechanisms**: Does the study analyze specific mechanisms (e.g., prompting, fine-tuning) through which model-to-model influence occurs?
- **Behavioral Analysis**: Does the study include analysis of model behavior or outputs (beyond purely technical or architectural aspects)?
- **Study Type**: Is the study either primary research with empirical data OR a systematic review/meta-analysis of model interactions?

We considered all screening questions together and made a holistic judgement about whether to screen in each paper.

Data extraction

We asked a large language model to extract each data column below from each paper. We gave the model the extraction instructions shown below for each column.

- **Research Approach for Investigating Identity Contagion/Persona Absorption**:

  Identify and describe the primary research methodology used:

- Classify the type of study (e.g., experimental, observational, theoretical)
- Describe the specific approach to investigating persona/identity effects in LLMs
- Note any unique experimental design elements

Look for:

- Explicit statements about methodology in the methods or introduction sections
- Specific experimental protocols used to test persona absorption
- Experimental conditions or manipulation techniques

If multiple approaches are used, list them in order of prominence. If information is unclear or incomplete, note "methodology not fully specified".

Examples of acceptable answers:

- "Experimental study using persona prompting across multiple LLMs"
- "Mixed-method approach combining quantitative performance testing and qualitative persona analysis"

- **Language Models and Personas Examined**:

  List ALL language models used in the study:

- Provide full model names (e.g., "GPT-4-Turbo", not just "GPT-4")
- Include version numbers if specified
- Note the number of models tested

Additional details to extract:

- Specific versions or release dates
- Any preprocessing or fine-tuning applied to models
- Source of the models (e.g., OpenAI, Meta)

If models are not explicitly named, write "Models not specified".

Formatting: Use a bulleted list or comma-separated list Example:

- ChatGPT-3.5
- GPT-4-Turbo
- Llama 2
- Mistral

- **Persona Induction Methods**:

Describe the specific methods used to induce or assign personas to language models:

- List types of personas used (e.g., demographic, personality traits, professional roles)
- Note the specific prompting techniques
- Capture the range or diversity of personas tested

Look for:

- Explicit descriptions of persona assignment in methods section
- Examples of actual prompts used
- Any psychological frameworks or personality models referenced

If multiple methods are used, list them in order of prominence.

Formatting: Provide a concise summary with specific examples where possible. Example: "Used 19 diverse personas across 5 socio-demographic groups, including racial and professional identities, induced through explicit role-playing prompts"

- **Key Findings on Identity Contagion/Persona Effects**:

  Extract the primary findings related to persona absorption or identity effects:

- Quantitative performance changes
- Qualitative observations about model behavior
- Statistically significant effects
- Unexpected or notable outcomes

Prioritize:

- Percentage or magnitude of performance changes
- Specific domains or tasks most affected
- Variations across different models

Look in results and discussion sections. If no clear findings, write "No significant findings reported".

Formatting: Use numerical data where possible, with context. Example: "80% of personas showed performance bias; up to 70% performance drop in some reasoning tasks"

Results

Characteristics of Included Studies

Study	Study Focus	Large Language Models Evaluated	Methodology Type	Key Findings	Full text retrieved
Shah et al., 2023	Persona modulation for black-box jailbreaks and transferability of harmful completions	GPT-4, Claude 2, Vicuna-33B	Experimental (automated persona modulation, harmful prompt generation)	Persona modulation increases harmful completions (42.5% in GPT-4); effects transfer to other models; high rates for xenophobia, sexism, disinformation	Yes
Frisch and Giulianelli, 2024	Personality consistency and linguistic alignment in interacting large language model agents	GPT-3.5-turbo	Experimental (persona prompting, collaborative writing, personality tests)	Creative personas show more consistency; linguistic alignment increases after interaction; creative agents adapt more	Yes
Hu et al., 2023	Social identity biases in large language models; effect of training data on in-group/outgroup bias	51+ large language models (GPT-2, GPT-3, Llama 2, Mistral, etc.)	Experimental (prompting, training data manipulation, comparison with human text)	Large language models exhibit social identity biases; fine-tuning increases bias; curation reduces it; does not directly address persona absorption	Yes

Study	Study Focus	Large Language Models Evaluated	Methodology Type	Key Findings	Full text retrieved
Collu et al., 2023	Adversarial persona prompting to bypass safety; defense mechanisms	GPT-3.5, GPT-3.5-turbo, Gemini-1.5-flash, Bing Chat, Bard, GPT-4o	Experimental (persona prompting, role-play, defense evaluation)	Adversarial personas bypass safety in most scenarios; defenses (multiple trustworthy personas) block up to 93% of jailbreaks; transferability across models	Yes
Tan et al., 2024	Effect of persona-based prompting (personality traits) on Theory of Mind (ToM) reasoning	Mistral 7B, Llama 2, Falcon 7B, Zephyr 7B Beta, GPT-3.5	Experimental (personality trait prompts, Theory of Mind tasks)	Dark Triad traits cause up to 33% performance drop in Theory of Mind tasks; agreeableness enhances performance; model-dependent sensitivity	Yes
Tan et al., "Personality Has An Effect", 2024	Effect of induced personalities (Dark Triad) on Theory of Mind reasoning	GPT-3.5, Llama 2, Mistral	Experimental (abstract only; personality trait prompts, Theory of Mind tasks)	Induced personalities significantly affect Theory of Mind reasoning; Dark Triad traits have large effects; higher variance in prompts increases controllability	No

Study	Study Focus	Large Language Models Evaluated	Methodology Type	Key Findings	Full text retrieved
Kovač et al., 2023	Perspective controllability; context-dependent value/personality expression	GPT-3.5-0301, GPT-3.5-0314, GPT-4-0314, OpenAssistant, Zephyr, StableVicuna, StableLM, etc.	Experimental (psychological questionnaires, context manipulation, qualitative/quantitative)	Large language models exhibit context-dependent values/personality; significant, unpredictable changes with perspective induction	Yes
Bernardelle et al., 2024	Mapping/manipulating political ideology using synthetic personas	No mention found	Experimental (abstract only; synthetic personas, ideological prompting)	Synthetic personas cluster left-libertarian; explicit prompting shifts models, but asymmetrically; inherent bias suggested	No
Gupta et al., 2023	Implicit reasoning biases in persona-assigned large language models	ChatGPT-3.5, GPT-4-Turbo, Llama-2-70B-Chat, ChatGPT-3.5-Nov	Experimental (persona assignment, 24 reasoning datasets, 19 personas)	80% of personas show performance bias; up to 70% performance drop; bias varies by model and group; GPT-4-Turbo least biased	Yes

Across the 9 included studies, we found the following patterns:

- Large language models evaluated:
  - GPT-4 (including variants) in 4 studies.
  - GPT-3.5 (including variants, including ChatGPT-3.5) in 5 studies.
  - Llama 2 in 3 studies.
  - Mistral in 3 studies.
  - Vicuna in 2 studies.

  - Zephyr in 2 studies.
  - Claude 2, Gemini, Bing Chat, Bard, Falcon, OpenAssistant, StableVicuna, and StableLM each in 1 study.
  - One study evaluated over 51 models, and in one study we didn't find mention of the specific models evaluated.
- Key findings:
  - Persona modulation or persona-based prompting increased harmful completions or enabled jailbreaks in 2 studies.
  - Defense mechanisms (such as multiple trustworthy personas) blocked up to 93% of jailbreaks in 1 study.
  - Transferability of persona-induced effects across models was reported in 2 studies.
  - Persona or persona-based prompting induced or modulated social, identity, or political bias in 6 studies.
  - Persona or personality trait prompts affected Theory of Mind (ToM) or reasoning performance in 3 studies.
  - Persona or context manipulation affected value or perspective controllability in 2 studies.
  - Model-dependent sensitivity or bias was reported in 2 studies.
  - All 9 included studies used experimental approaches; we didn't find mention of non-experimental methodologies in the included studies.
- Methodological notes:
  - All studies used experimental methods, with variations in persona prompting, role-play, or prompt manipulation.

Thematic Analysis

Persona Stability and Controllability

Study	Key Observations	Supporting Studies	Contradicting Evidence
Persona adoption is robust; large language models can be steered to adopt diverse personas, affecting outputs and reasoning.	Shah et al., 2023; Collu et al., 2023; Tan et al., 2024; Kovač et al., 2023; Gupta et al., 2023	Hu et al., 2023 (does not address persona absorption)	
Consistency of persona expression varies by model and persona type; creative personas show more stability.	Frisch and Giulianelli, 2024; Kovač et al., 2023		

Study	Key Observations	Supporting Studies	Contradicting Evidence
Persona traits can be controllably adjusted via prompts; higher variance in prompts increases controllability.	Tan et al., 2024; Tan et al., "Personality Has An Effect", 2024		
Context-dependent expression: large language models' values and traits shift unpredictably with context/perspective.	Kovač et al., 2023		

We found four main categories of persona-related phenomena in large language models, with the following number of supporting studies for each:

- Persona adoption/steering:5 studies (Shah et al., 2023; Collu et al., 2023; Tan et al., 2024; Kovač et al., 2023; Gupta et al., 2023)
- Consistency/stability of persona expression:2 studies (Frisch and Giulianelli, 2024; Kovač et al., 2023)
- Controllability of persona traits via prompts:2 studies (Tan et al., 2024; Tan et al., "Personality Has An Effect", 2024)
- Context-dependent/persona expression shifts:1 study (Kovač et al., 2023)

Kovač et al., 2023 was the only study supporting three different categories. Tan et al., 2024 appeared in two categories. We didn't find mention of supporting studies for context-dependent persona expression beyond Kovač et al., 2023, in the included studies.

One study (Hu et al., 2023) was cited as not addressing persona absorption, but we didn't find mention of other contradicting evidence.

In total, 7 unique studies were cited as supporting evidence across all categories, and 1 additional study was cited as contradicting evidence.

Inter-Model Identity Transfer

Study	Key Observations	Supporting Studies	Contradicting Evidence
Persona-modulating prompts are transferable: jailbreaks and adversarial personas effective across models.	Shah et al., 2023; Collu et al., 2023		

Study	Key Observations	Supporting Studies	Contradicting Evidence
Interaction between large language model agents leads to linguistic and personality alignment; creative personas adapt more.	Frisch and Giulianelli, 2024		
No direct evidence of "identity contagion" (i.e., persona transfer from one model to another via interaction); most effects are prompt-based or within-model.	All studies		

- Transferability of persona-modulating prompts:1 study reported that persona-modulating prompts (such as jailbreaks and adversarial personas) are transferable and effective across models.
- Linguistic and personality alignment:1 study described that interaction between large language model agents leads to linguistic and personality alignment, with creative personas adapting more.
- No direct evidence of identity contagion:In 1 study (summarizing all studies), we found an explicit statement that there is no direct evidence of "identity contagion" (i.e., persona transfer from one model to another via interaction); most observed effects are prompt-based or within-model.
- We didn't find mention of any studies providing supporting or contradicting evidence in the respective columns.

Unintended Identity Effects

Study	Key Observations	Supporting Studies	Contradicting Evidence
Persona assignment can reveal or amplify deep-rooted biases, leading to significant performance drops for certain groups.	Gupta et al., 2023; Tan et al., 2024; Collu et al., 2023		
Induced personas can dramatically increase harmful outputs (e.g., jailbreaks, disinformation).	Shah et al., 2023; Collu et al., 2023		

Study	Key Observations	Supporting Studies	Contradicting Evidence
Some models are more robust to persona-induced bias (e.g., GPT-4-Turbo), but all are affected to some degree.	Gupta et al., 2023; Tan et al., 2024		
Persona-based defenses (e.g., multiple trustworthy personas) can mitigate adverse effects, but effectiveness varies by model.	Collu et al., 2023		

We found four distinct categories of key observations regarding persona assignment in large language models, each supported by 1–3 studies:

- Persona assignment revealing or amplifying deep-rooted biases and causing performance drops:3 studies (Gupta et al., 2023; Tan et al., 2024; Collu et al., 2023).
- Induced personas increasing harmful outputs (such as jailbreaks or disinformation):2 studies (Shah et al., 2023; Collu et al., 2023).
- Some models being more robust to persona-induced bias, but all being affected to some degree:2 studies (Gupta et al., 2023; Tan et al., 2024).
- Persona-based defenses mitigating adverse effects, with effectiveness varying by model:1 study (Collu et al., 2023).

Across all observations, Collu et al., 2023 was cited most frequently (3 times), followed by Gupta et al., 2023 and Tan et al., 2024 (2 times each), and Shah et al., 2023 (1 time). We didn't find mention of any contradicting evidence reported for any observation, and we didn't find references to any other studies in this table.

Synthesis of Findings

Identity Contagion Patterns

- Persona absorption within individual large language models:The included studies report that persona absorption is robust within individual large language models, with evidence that persona-inducing prompts can alter outputs, reasoning, and safety-related behaviors.
- Transferability of persona-modulating prompts across models:Some included studies report transferability of persona-modulating prompts (such as jailbreaks and adversarial personas) across models, suggesting that similar vulnerabilities may exist in multiple large language models.
- Linguistic and personality alignment in agent interaction:Interaction between large language model agents can lead to linguistic and personality alignment, but direct evidence of "identity contagion" (i.e., persona transfer from one model to another via interaction) is limited to alignment phenomena rather than full persona adoption.

Influencing Factors:

- Model architecture and training:Factors such as Reinforcement Learning from Human Feedback (RLHF) and fine-tuning influence susceptibility to persona absorption and bias.
- Persona type:The type of persona (e.g., Dark Triad traits, adversarial roles, demographic identities) affects the magnitude and nature of absorption effects.
- Context and prompt design:Context and prompt design play a critical role in both intended and unintended persona effects.

Implications for Large Language Model Interaction

- Safety and defense mechanisms:The ease with which large language models can be steered into harmful or biased personas highlights the need for robust safety and defense mechanisms.
- Effectiveness of persona-based defenses:Persona-based defenses (such as multiple trustworthy personas) show promise but are not universally effective.
- Context-dependent and unpredictable persona absorption:The context-dependent and sometimes unpredictable nature of persona absorption complicates efforts to ensure consistent, safe, and unbiased large language model behavior.
- Monitoring and mitigation:Monitoring and mitigating unintended identity effects, especially bias amplification and performance degradation for marginalized groups, is essential for responsible large language model deployment.

References

F. Tan, G. Yeo, Fanyou Wu, Weijie Xu, Vinija Jain, Aman Chadha, Kokil Jaidka, Yang Liu, and See-Kiong Ng. "PHAnToM: Persona-Based Prompting Has an Effect on Theory-of-Mind Reasoning in Large Language Models." *International Conference on Web and Social Media*, 2024.

———. "PHAnToM: Personality Has An Effect on Theory-of-Mind Reasoning in Large Language Models." *arXiv.org*, 2024.

Grgur Kovač, Masataka Sawayama, Rémy Portelas, Cédric Colas, Peter Ford Dominey, and Pierre-Yves Oudeyer. "Large Language Models as Superpositions of Cultural Perspectives." *arXiv.org*, 2023.

Ivar Frisch, and Mario Giulianelli. "LLM Agents in Interaction: Measuring Personality Consistency and Linguistic Alignment in Interacting Populations of Large Language Models." *PERSONALIZE*, 2024.

Matteo Gioele Collu, Tom Janssen-Groesbeek, Stefanos Koffas, Mauro Conti, and S. Picek. "Dr. Jekyll and Mr. Hyde: Two Faces of LLMs." *arXiv.org*, 2023.

Pietro Bernardelle, Leon Fröhling, S. Civelli, Riccardo Lunardi, Kevin Roitero, and Gianluca Demartini. "Mapping and Influencing the Political Ideology of Large Language Models Using Synthetic Personas." *arXiv.org*, 2024.

Rusheb Shah, Quentin Feuillade--Montixi, Soroush Pour, Arush Tagade, Stephen Casper, and Javier Rando. "Scalable and Transferable Black-Box Jailbreaks for Language Models via Persona Modulation." *arXiv.org*, 2023.

Shashank Gupta, Vaishnavi Shrivastava, A. Deshpande, A. Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. "Bias Runs Deep: Implicit Reasoning Biases in Persona-Assigned LLMs." *International Conference on Learning Representations*, 2023.

Tiancheng Hu, Yara Kyrychenko, Steve Rathje, Nigel Collier, S. V. D. Linden, and Jon Roozenbeek. "Generative Language Models Exhibit Social Identity Biases." *Nature Computational Science*, 2023.