A Multimedia Perspective on the Study of Cognitive Biases in Large Language Models

Markus Schedl
Johannes Kepler University Linz and Linz Institute of Technology
Linz, Austria
markus.schedl@jku.at

Shahed Masoudian
Johannes Kepler University Linz
Linz, Austria
shahed.masoudian@jku.at

Antonela Tommasel
Johannes Kepler University Linz
Linz, Austria
UNCPBA and CONICET
Buenos Aires, Argentina
antonela.tommasel@jku.at

Abstract

Cognitive biases are commonly defined as systematic deviations from rationality in human judgment and decision making. They have been studied in psychology for decades, with corresponding literature claiming to have found evidence for several hundreds of such biases. Driven by the success of large language models (LLMs), researchers have recently begun to investigate whether cognitive biases also emerge in LLMs. While some evidence has been found, respective studies are typically limited to textual interactions with the LLM under investigation, i.e., providing a text prompt aimed at uncovering a cognitive bias and receiving a textual answer. Despite today's widespread adoption of vision–language models (VLMs) and multimodal large language models (MLLMs), multimodal investigations of cognitive biases in generative artificial intelligence systems are rare. Addressing this gap, in the position paper at hand, we review cognitive biases identified in psychology whose study in LLMs requires multimodal data. Focusing on the text and image modalities, we discuss the identified biases, offer corresponding research hypotheses, and provide preliminary evidence.

CCS Concepts

• **Computing methodologies** → **Natural language processing**; **Cognitive science**; • **Applied computing** → **Psychology**.

Keywords

multimodal large language models, vision–language models, cognition, psychology, bizarreness effect, declinism, contrast effect, pareidolia, self-preference bias

ACM Reference Format:
Markus Schedl, Shahed Masoudian, and Antonela Tommasel. 2026. A Multimedia Perspective on the Study of Cognitive Biases in Large Language Models. In *The 5th International Workshop on Multimodal Human Understanding for the Web and Social Media (MUWS '26), November 10–14, 2026, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3841457.3841518

MUWS '26, Rio de Janeiro, Brazil

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/10.1145/3841457.3841518

1 Background and Motivation

Cognitive biases have been researched for decades in psychology, sociology, and economics. In psychology, they are commonly defined as systematic deviations from rationality and objectivity in human perception and reasoning, affecting judgment and decision making processes [12, 25]. While there is an ongoing debate among psychology scholars about what constitutes a cognitive bias and what does not [15, 29],[1] corresponding literature claims to have found evidence for more than 200 cognitive biases.

With the spiraling rise of generative large language models (LLMs), which have reached an unprecedented quality level in text generation and often provide human-like responses, a natural question arises as to which extent LLMs may be prone to similar cognitive biases as humans. As a result, research on natural language processing has begun studying analogous phenomena in LLMs [41, 45]. To provide a few examples, strong evidence for *framing effects* has been established by Malberg et al. [33] who showed that LLMs were more likely to positively evaluate a situation description when it was stated in a positive way compared to the equivalent information framed more negatively [33]. Alessa et al. [1] found significant changes in framing, with respect to their output's sentiment, when LLMs were prompted to generate summaries of product reviews and interviews. Xu et al. [51] revealed that LLMs suffer from unwarranted *overconfidence* in the correctness of their responses to question answering tasks, similar to the Dunning–Kruger effect in humans [37]. Schilcher et al. [42] uncovered *positional effects* when tasking LLMs to generate descriptions for a list of topics. A topic's position in the prompt affected structural and coherence features of the created descriptions and, in some cases, even resulted in the topic's omission.

Unsurprisingly, the vast majority of research on cognitive biases in LLMs follows a single-modal approach, where both prompt and response are purely textual. This is inline with the fact that most psychological tests for cognitive biases rely on textual instructions to study subjects in the form of vignettes [6]. In cases where the psychological experimental setup includes exposure to visual stimuli, it is commonly transformed into a textual description when adapted for the assessment of LLMs. In both cases, the resulting study setup neglects the multimedia capabilities of current multimodal large language models (MLLMs).

State-of-the-art MLLMs are capable of processing textual, visual, and auditory data, both at prompting and generation stages [23, 30]. This has opened new possibilities to investigate cognitive biases

[1] This dispute commonly revolves around the interpretation of rationality and is sometimes referred to as "Rationality War". More information can be found, e.g., in [16].

in LLMs, either by (1) extending experimental setups for already researched cognitive biases carried out on single-modal materials to include multimodal artifacts or (2) enabling the study of cognitive biases that could previously not be investigated in LLMs due to the requirement for visual or auditory stimuli. Addressing (1), among the few existing works, Huang et al. [20] evaluated MLLMs with respect to the influence of deceptive patterns in short videos causing cognitive biases, including *logical fallacies*. Kim et al. [27] studied *halo effects* in MLLM-supported hiring decisions, uncovering differences in MLLMs' ratings of candidates when prompts included or excluded pictures of potential candidates. As for (2), a notable example for a study that could benefit from including visuals is Asad et al. [2], which investigated body image *stereotypes* in a purely textual manner.

Against this background, the position paper at hand proposes to target both directions, which are still heavily underexplored, by leveraging highly capable MLLMs that have emerged recently, such as *Qwen3-VL* [3], *GLM-4.5V* [47], *InternVL3.5* [49], *Gemini 2.5 Pro* [9], *GPT-5* [43], and *Deepseek V4 Vision* [11]. In the following, we detail our perspective on the multimodal study of cognitive biases in LLMs and provide initial empirical evidence (Section 2). Subsequently, we point to grand challenges ahead (Section 3).

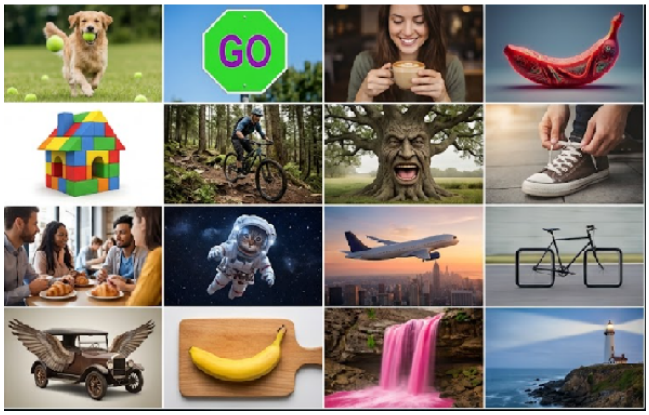
2 Cognitive Biases in Multimodal LLMs

To support our main proposal, i.e., the reconsideration of the study of cognitive biases as a multimodal endeavor, we first carefully reviewed the list of cognitive biases provided in the *Cognitive Bias Codex* [5] and identified those whose analysis in LLMs either necessitates or could be enriched by integrating multimodal data and models. We selected five of them (bizarreness effect, declinism, contrast effect, pareidolia, and self-preference bias) for further analysis. In the following, we introduce them and provide hypotheses for their possible manifestiations in MLLMs as well as initial empirical evidence. Please note that the conducted experiments are not meant to be exhaustive. They rather aim to support our perspective and serve as a starting point for subsequent in-depth studies. Therefore, the experimental setup was deliberately based on ad-hoc choices instead of comprehensive permutations of prompting strategies, models, and multimedia materials. In the experiments, we considered state-of-the-art models of the *GPT*, *Gemini*, *Claude*, and *Deepseek* families because of their widespread usage.

Bizarreness Effect

This effect refers to the human tendency to better remember unusual than mundane information [14]. Studying it includes sequential exposure of participants to mixed lists containing both ordinary and bizarre stimuli (e.g., words or images) [17]. Retrieved memory is then assessed through recall tasks in which participants describe or reproduce the items they have previously observed.

While the memory mechanism of MLLMs is fundamentally different from that of humans, we nevertheless hypothesize that MLLMs also react differently to strange than to common inputs. We therefore adapted the experimental setup from Gounden and Nicolas [18] by constructing a 4×4 image patch containing 16 images, of which 8 were deliberately bizarre and randomly interspersed among ordinary images. The bizarre images included a green "Go" sign, a red



(a) Image grid provided for analysis

(b) Generation with distractor

(c) Generation without distractor

Figure 1: Bizarreness Effect — Images provided for analysis (a) and generation results with distractor (b) and without distractor (c) regarding using *Gemini 3.5 Flash*.

banana, a tree with a face, an astronaut cat, a bicycle with square wheels, a flying car with wings, and a pink waterfall (Figure 1a).

We first prompted *Gemini 3.5 Flash* to analyze the composite image while asking a mathematical question. The rationale was that the mathematical task would act as an interference step, shifting the model away from the immediate visual details of the image while preserving any latent representations that remained salient. In a subsequent prompt, we asked the model to generate a single image based on the three most significant visual elements from the image it had analyzed previously. Figure 1b presents the image generated by the model. We observe that the selected objects are solely drawn from the set of bizarre images. Notably, the effect still persisted when repeating the experiment without the distractor task, as evidenced in Figure 1c. Although preliminary, this result suggests that unusual visual stimuli may produce stronger attention activations in multimodal models, analogous to the increased memorability of bizarre stimuli reported in human cognition studies.

Declinism and Rosy Retrospection

These cognitive biases are defined, respectively, as the perception that the world or society is declining [44] and to remembering past events more positively than they were actually experienced [35]. Rosy retrospection can therefore be regarded as a cause of declinism. In the context of the multimodal music domain, such effects have been found in song lyrics; e.g., Parada-Cabaleiro et al. [40] showed that emotions reflected in lyrics have become more negative over the last decades. However, this work considers lyrics only as purely textual data.

In a multimodal setting, we hypothesize that image generation may be affected by such declinism tendencies in a way that, when prompted to create images from the past, the results will generally convey or induce more positive emotions compared to prompts that explicitly ask for a present setting or prompts with no explicit temporal request. Whether such effects, if existing, are the result of biased training data or induced during training, finetuning, or inference, needs further investigation.[2]

Anecdotal evidence is provided in Figure 2 which illustrates the results of two neutral prompts to *Gemini 3.5 Flash*: "Please create an image that shows the world [in the 1960s | today]."[3] While both images depict a similar scene of a city, the lower one may appear more dystopian, requiring people to wear masks and only starting at their smartphones.

Figure 2: Declinism — Images created by *Gemini 3.5 Flash* prompted to depict the world "in the 1960s" (above) and "today" (below).

Contrast Effect

This effect describes the intensification or diminution of a stimulus' perception when compared with a simultaneously or recently observed contrasting stimulus, thereby enhancing the differences between the compared items. Contrast effects have been foundationally researched in the perception of color contrasts [22] and product pricing [36].

[2]Limited technological possibilities in the past, esp. digital photography, may cause (1) fewer pictures being available from earlier times for training VLMs and (2) available images being taken in particularly emotion-laden settings, i.e., fewer careless taking of highly disposable pictures.

[3]As a side note, it also seems to suffer from an outdated interpretation of "today".

We hypothesize that similar effects also manifest when MLLMs are presented with two contrasting visual stimuli concurrently. Adapted from Kenrick and Gutierres [26], who studied contrast effects on women's attractiveness, we established a simple multimodal experimental setup to investigate the presence of contrast effects in MLLMs' judgments. We first presented *Gemini 3.5 Flash* with an image of a "normal" room (Figure 3, middle) and asked it to rate the room's "physical and aesthetic quality ... on a strict integer scale from 1 to 10". When evaluated in isolation, the room received a rating of 5. When presented alongside a luxury room (Figure 3, right), it again received a rating of 5. However, when presented alongside a poor-quality room (Figure 3, left), its rating increased to 6.[4] Although preliminary, these results suggest that exposure to contrasting inputs can influence the model's evaluation, indicating the presence of contrast effects in multimodal judgments.

Figure 3: Contrast Effect — Images given to *Gemini 3.5 Flash* for rating. The middle picture depicts a normal room while the room on the left serves as negative contrast and the one on the right as positive contrast.

Pareidolia

This phenomenon refers to the human tendency to perceive a vague stimulus as significant and impose a meaningful interpretation [24]. Prominent evidence has been found for perceiving animals, faces, or objects in cloud formations [32, 53].

We hypothesize that MLLMs could show similar behavior, while probably not as the dominating response but a secondary reaction to an image. To test this hypothesis, we first created an image showing a cloudy sky, intentionally making some clouds in the right part appear similar to the trunk and head of an elephant, as shown in Figure 4. Subsequently, we provided the picture to several MLLMs, applying the ad-hoc prompt "What can you see in this image?". Interestingly, the answers varied substantially: *Claude Sonnet 5* provided a very detailed analysis, speculating about the types of clouds and even filters that had likely been used to take the picture. It also detected "a question mark or a small funnel/tail shape". In contrast, *Gemini 3.5 Flash* answered concisely, clearly identifying a "distinct, twisting cloud formation that resembles a question mark or a stylized dragon/snake shape". *Deepseek V4 Vision* instead perceived a "distinct "S" shape or a jagged, meandering curve ... a contrail (the condensed water vapor left behind by a high-flying airplane)". In summary, all investigated MLLMs provided

[4]To test for possible position effects, we also changed the order of the rooms and asked the same question. We observed that the ratings pertaining to the 3 rooms remained the same, results therefore still held.

some pareidolic speculations, even when not explicitly asked to search for unusual patterns in the clouds.

Figure 4: Pareidolia — An image showing a cloud formation whose right parts hint at a trunk of an elephant.

Self-Preference Bias and IKEA Effect

These highly related cognitive biases refer to the phenomenon that individuals value higher things they invested time or labor in. For instance, a study reported in Norton et al. [39] found participants were willing to pay a higher price for furniture they had helped building than for ready-made items. The common explanation of this effect is humans' desire to justify their efforts.

In the context of LLMs, it was found that most models rate texts they generated themselves higher than texts generated by other LLMs [8, 52]. To the best of our knowledge, such effects have not been studied for images. We therefore hypothesize that images created by one MLLM are rated more positively by that MLLM than similar images created by other MLLMs.

To test this hypothesis, we carried out a simple experiment: First, we considered the corresponding prompts of 5 randomly sampled images from the *OpenGPT-4o-Image* dataset [7], covering different categories.[5] Subsequently, we used the prompts to generate new images via the state-of-the-art models *Gemini 3.5 Flash* and *GPT 5.5 Instant*.[6] In a binary preference assessment task, we provided each model with each pair of images created with the same prompt. We investigated prompt-independent preference and prompt-conditioned preference. For the former, we asked the MLLMs "Which of these images do you prefer?" while the latter was assessed using the prompt "Which of these images do you prefer for "[prompt used for generating the image pair]"?". Interestingly, the results indicate a tendency for self-preference bias in *GPT 5.5* (80% of the preferred content was self-generated) but not in *Gemini 3.5* (40% self-preference), for the prompt-independent preference setup. For the prompt-conditioned experiment, results were even more pronounced: *Gemini 3.5* never preferred its own generations while *GPT 5.5* preferred its own creations in 80% of cases.

[5]Scientific Imagery, Culture and History, Spatial Reasoning, In-Image Text Rendering, and Style Control

[6]The images included in *OpenGPT-4o-Image* have been created using an older *GPT-4o* model and are substantially inferior to those created by state-of-the-art models. A comparison against those would therefore be misleading.

Other Relevant Cognitive Biases

In addition to the cognitive biases discussed above in more detail, we identified the following ones from the *Cognitive Bias Codex* [5] whose study could benefit from a multimodal perspective:

Cognitive Dissonance refers to the cognitive tension that humans feel when presented with contradictory information or actions [19]. This stimulates the mind to resolve the conflict. An example where such effects were studied in LLMs is Lehr et al. [28], who found that GPT-4o changed its attitude towards a person to resolve contradictory stimuli. However, their work was disputed, in particular by Cummins et al. [10]. In any case, Lehr et al. [28] relied purely on text, while MLLMs could be leveraged to study textual-visual, audio-visual, and even textual-audio-visual inconsistencies.

Halo Effect and Horn Effect describe humans' tendency to generalize initially perceived specific traits of a person in a positive or negative way, respectively [38, 48]. This indicates a spill-over of the "first impression" one gets of a subject to their other characteristics. For instance, Batres and Shiramizu [4] showed that physically attractive people tend to be perceived as having more positive personality traits than less attractive ones. In the context of LLMs, Kim et al. [27] studied halo effects and horn effects by simulating a candidate rating task for a given job ad. They prompted various LLMs and MLLMs with job applications and supplementary information of the candidates (either textual or visual), finding that image-based supplementary information had a larger influence on evaluations compared to text-only information. In visual settings, *LLaVA-OneVision* models resulted in a particularly pronounced halo effect, while the 26B version of *InternVL2.5* showed a considerable horn effect.

Salience Bias is defined as the tendency to focus one's attention to items with particularly outstanding, prominent, or emotion-laden features, even if they are irrelevant in the given context [21, 46]. In contrast to the bizarreness effect discussed above, the focus here lies on the *attention* given to items with salient characteristics instead of the ease of remembering them. Salience bias in MLLMs could be uncovered by examining their attention maps. Similarities between manifestations of this bias in humans and in MLLMs could be researched, e.g., by comparing MLLMs' attention maps with the output of detectors for salient regions [31, 34] or salient objects [13, 50], or with human annotations given either by explicit saliency labels or eyetracking data. If findings point to the existence of salience biases in MLLMs, the next step should be to investigate the ways in which they influence the models' responses.

3 Conclusions and Grand Challenges

This position paper argued for reconsidering the emerging research direction of investigating cognitive biases in LLMs by taking a multimodal perspective, in particular, by leveraging the analytical and generative capabilities of modern MLLMs. Under this perspective, we reviewed the *Cognitive Bias Codex* to identify suited biases and examined five in more detail, discussing existing work that addressed them in text-only setups and providing hypotheses and initial evidence in textual-visual settings. While the conducted experiments support our hypotheses, additional and deeper investigations on a larger scale are required next. More generally, following this research path, the grand challenges we envision include:

- Creating a detailed taxonomy of cognitive biases relevant to study under a multimodal perspective, including but not limited to textual, visual, and auditory channels.
- Researching technologically and psychologically sound adaptations of human experiments to a multimedia-driven cognitive bias analysis in MLLMs.
- Beyond the single prompt–response interaction in our initial experiments, examining how cognitive biases evolve and influence each other in longer conversations between humans and MLLMs or even between different MLLMs.
- Mitigating detrimental effects of multimedia-informed cognitive biases in human–MLLM or MLLM–MLLM communication scenarios.
- Leveraging potentials of multimedia-informed cognitive biases in MLLMs, e.g., to improve their explainability, efficiency, or trustworthiness.

To conclude, we look forward to seeing exciting interdisciplinary research at the intersection of computer science and psychology, that investigates cognitive biases throughout the entire MLLM ecosystem from a multimedia perspective.

Acknowledgments

This research was funded in whole or in part by the Austrian Science Fund (FWF): 10.55776/COE12, 10.55776/DFH23, 10.55776/P36413; and by the State of Upper Austria and the Federal Ministry of Education, Science, and Research: LIT-2024-13-SEE-111.

References

[1] Abeer Alessa, Param Somane, Akshaya Thenkarai Lakshminarasimhan, Julian Skirzynski, Julian McAuley, and Jessica Maria Echterhoff. 2025. Quantifying Cognitive Bias Induction in LLM-Generated Content. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, Kentaro Inui, Sakriani Sakti, Haofen Wang, Derek F. Wong, Pushpak Bhattacharyya, Biplab Banerjee, Asif Ekbal, Tanmoy Chakraborty, and Dhirendra Pratap Singh (Eds.). The Asian Federation of Natural Language Processing and The Association for Computational Linguistics, Mumbai, India, 2890–2910. doi:10.18653/v1/2025.ijcnlp-long.155

[2] Narjis Asad, Nihar Ranjan Sahoo, Rudra Murthy, Swaprava Nath, and Pushpak Bhattacharyya. 2025. "You are Beautiful, Body Image Stereotypes are Ugly!" BIStereo: A Benchmark to Measure Body Image Stereotypes in Language Models. In *Findings of the Association for Computational Linguistics: ACL 2025*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 24471–24496. https://aclanthology.org/2025.findings-acl.1257/

[3] Shuai Bai, Yuxuan Cai, Ruizhe Chen, and et al. 2025. Qwen3-VL Technical Report. doi:10.48550/arXiv.2511.21631 arXiv:2511.21631 [cs.CV].

[4] Carlota Batres and Victor Shiramizu. 2023. Examining the "Attractiveness Halo Effect" Across Cultures. *Current Psychology* 42, 29 (2023), 25515–25519.

[5] Buster Benson. 2016. Cognitive Bias Cheat Sheet. https://buster.medium.com/cognitive-bias-cheat-sheet-55a472476b18

[6] J. S. Blumenthal-Barby and Heather Krieger. 2015. Cognitive Biases and Heuristics in Medical Decision Making: A Critical Review Using a Systematic Search Strategy. *Medical Decision Making: An International Journal of the Society for Medical Decision Making* 35, 4 (May 2015), 539–557. doi:10.1177/0272989X14547740

[7] Zhihong Chen, Xuehai Bai, Yang Shi, Chaoyou Fu, Huanyu Zhang, Haotian Wang, Xiaoyan Sun, Zhang Zhang, Liang Wang, Yuanxing Zhang, Pengfei Wan, and Yi-Fan Zhang. 2025. OpenGPT-4o-Image: A Comprehensive Dataset for Advanced Image Generation and Editing. doi:10.48550/arXiv.2509.24900 arXiv:2509.24900 [cs.CV].

[8] Zhi-Yuan Chen, Hao Wang, Xinyu Zhang, Enrui Hu, and Yankai Lin. 2025. Beyond the Surface: Measuring Self-Preference in LLM Judgments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (Eds.). Association for Computational Linguistics, Suzhou, China, 1653–1672. doi:10.18653/v1/2025.emnlp-main.86

[9] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, and et al. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. doi:10.48550/arXiv.2507.06261 arXiv:2507.06261 [cs.CL].

[10] Jamie Cummins, Malte Elson, and Ian Hussey. 2025. Cognitive Dissonance in Large Language Models Is Neither Cognitive nor Dissonant. *Proceedings of the National Academy of Sciences* 122, 35 (Sept. 2025), e2517912122. doi:10.1073/pnas.2517912122

[11] DeepSeek-AI, Anyi Xu, Bangcai Lin, and et al. 2026. DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence. doi:10.48550/arXiv.2606.19348 arXiv:2606.19348 [cs.CL].

[12] Rolf Dobelli. 2013. *The Art of Thinking Clearly: Better Thinking, Better Decisions.* Hachette, UK.

[13] Shixuan Gao, Pingping Zhang, Tianyu Yan, and Huchuan Lu. 2024. Multi-Scale and Detail-Enhanced Segment Anything Model for Salient Object Detection. In *Proceedings of the 32nd ACM International Conference on Multimedia* (Melbourne VIC, Australia) *(MM '24)*. Association for Computing Machinery, New York, NY, USA, 9894–9903. doi:10.1145/3664647.3680650

[14] Lisa Geraci, Mark A. McDaniel, Tyler M. Miller, and Matthew L. Hughes. 2013. The Bizarreness Effect: Evidence for the Critical Influence of Retrieval Processes. *Memory & Cognition* 41, 8 (Nov. 2013), 1228–1237. doi:10.3758/s13421-013-0335-4

[15] Gerd Gigerenzer. 2000. *Adaptive Thinking: Rationality in the Real World.* Oxford University Press.

[16] Gerd Gigerenzer. 2025. The Rationality Wars: A Personal Reflection. *Behavioural Public Policy* 9, 3 (July 2025), 495–515. doi:10.1017/bpp.2024.51

[17] Yannick Gounden, Fabien Cerroti, and Serge Nicolas. 2017. Secondary distinctiveness effects: Orthographic distinctiveness and bizarreness effects make independent contributions to memory performance. *Scandinavian Journal of Psychology* 58, 1 (2017), 9–14. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/sjop.12338 doi:10.1111/sjop.12338

[18] Yannick Gounden and Serge Nicolas. 2012. The impact of processing time on the bizarreness and orthographic distinctiveness effects. *Scandinavian Journal of Psychology* 53, 4 (2012), 287–294. doi:10.1111/j.1467-9450.2012.00945.x

[19] Eddie Harmon-Jones (Ed.). 2019. *Cognitive Dissonance: Reexamining a Pivotal Theory in Psychology* (2 ed.). American Psychological Association. https://www.jstor.org/stable/j.ctv1chs6tk

[20] Jen-tse Huang, Chang Chen, Shiyang Lai, Wenxuan Wang, Michelle R Kaufman, and Mark Dredze. 2026. Probing Multimodal Large Language Models on Cognitive Biases in Chinese Short-Video Misinformation. In *Findings of the Association for Computational Linguistics: ACL 2026*, Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (Eds.). Association for Computational Linguistics, San Diego, California, United States, 12675–12696. doi:10.18653/v1/2026.findings-acl.616

[21] Martin P. Inderbitzin, Alberto Betella, Antonio Lanatá, Enzo P. Scilingo, Ulysses Bernardet, and Paul F. M. J. Verschure. 2013. The Social Perceptual Salience Effect. *Journal of Experimental Psychology: Human Perception and Performance* 39, 1 (2013), 62–74. doi:10.1037/a0028317

[22] Dorothea Jameson and Leo M. Hurvich. 1964. Theory of Brightness and Color Contrast in Human Vision. *Vision Research* 4, 1 (May 1964), 135–154. doi:10.1016/0042-6989(64)90037-9

[23] Yizhang Jin, Jian Li, Tianjun Gu, Yexin Liu, Bo Zhao, Jinxiang Lai, Zhenye Gan, Yabiao Wang, Chengjie Wang, Xin Tan, and Lizhuang Ma. 2025. Efficient Multimodal Large Language Models: A Survey. *Visual Intelligence* 3, 1 (Dec. 2025), 27. doi:10.1007/s44267-025-00099-6

[24] Karl Ludwig Kahlbaum. 1866. Die Sinnesdelierien. *Allgemeine Zeitschrift für Psychiatrie und psychisch-gerichtliche Medizin* 23 (1866), 1–86.

[25] Daniel Kahneman. 2017. *Thinking, Fast and Slow.* Penguin Books.

[26] Douglas T Kenrick and Sara E Gutierres. 1980. Contrast Effects and Judgments of Physical Attractiveness: When Beauty Becomes a Social Problem. *Journal of Personality and Social Psychology* 38, 1 (1980), 131.

[27] Kyusik Kim, Jeongwoo Ryu, Hyeonseok Jeon, and Bongwon Suh. 2025. Blinded by Context: Unveiling the Halo Effect of MLLM in AI Hiring. In *Findings of the Association for Computational Linguistics: ACL 2025*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 26067–26113. doi:10.18653/v1/2025.findings-acl.1338

[28] Steven A. Lehr, Ketan S. Saichandran, Eddie Harmon-Jones, Nykko Vitali, and Mahzarin R. Banaji. 2025. Kernels of Selfhood: GPT-4o Shows Human-like Patterns of Cognitive Dissonance Moderated by Free Choice. *Proceedings of the National Academy of Sciences* 122, 20 (May 2025), e2501823122. doi:10.1073/pnas.2501823122

[29] Tomás Lejarraga and Ralph Hertwig. 2021. How Experimental Methods Shaped Views on Human Competence and Rationality. *Psychological Bulletin* 147, 6 (June 2021), 535–564. doi:10.1037/bul0000324

[30] Chia Xin Liang, Pu Tian, Caitlyn Heqi Yin, Yao Yua, An-Hou Wei, Ming Li, Xinyuan Song, Tianyang Wang, Ziqian Bi, Ming Liu, Riyang Bao, and Pengbin Feng. 2026. A Comprehensive Survey and Guide to Multimodal Large Language

Models in Vision–Language Tasks. *Computation* 14, 6 (June 2026), 125. doi:10.3390/computation14060125

[31] Guang-Hai Liu and Jing-Yu Yang. 2019. Exploiting Color Volume and Color Difference for Salient Region Detection. *IEEE Transactions on Image Processing* 28, 1 (2019), 6–16. doi:10.1109/TIP.2018.2847422

[32] Jiangang Liu, Jun Li, Lu Feng, Ling Li, Jie Tian, and Kang Lee. 2014. Seeing Jesus in Toast: Neural and Behavioral Correlates of Face Pareidolia. *Cortex* 53 (April 2014), 60–77. doi:10.1016/j.cortex.2014.01.013

[33] Simon Malberg, Roman Poletukhin, Carolin M. Schuster, and Georg Groh. 2025. A Comprehensive Evaluation of Cognitive Biases in LLMs. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, Mika Hämäläinen, Emily Öhman, Yuri Bizzoni, So Miyagawa, and Khalid Alnajjar (Eds.). Association for Computational Linguistics, Albuquerque, USA, 578–613. doi:10.18653/v1/2025.nlp4dh-1.50

[34] Zhendong Mao, Yongdong Zhang, Ke Gao, and DongMing Zhang. 2012. A Method for Detecting Salient Regions Using Integrated Features. In *Proceedings of the 20th ACM International Conference on Multimedia* (Nara, Japan) *(MM '12)*. Association for Computing Machinery, New York, NY, USA, 745–748. doi:10.1145/2393347.2396302

[35] Mara Mather and Laura L. Carstensen. 2005. Aging and Motivated Cognition: The Positivity Effect in Attention and Memory. *Trends in Cognitive Sciences* 9, 10 (2005), 496–502. doi:10.1016/j.tics.2005.08.005

[36] Kent B. Monroe. 1973. Buyers' Subjective Perceptions of Price. *Journal of Marketing Research* 10, 1 (Feb. 1973), 70–80. doi:10.1177/002224377301000110

[37] Don A. Moore and Paul J. Healy. 2008. The Trouble with Overconfidence. *Psychological Review* 115, 2 (April 2008), 502–517. doi:10.1037/0033-295X.115.2.502

[38] Richard E. Nisbett and Timothy D. Wilson. 1977. The halo effect: Evidence for unconscious alteration of judgments. *Journal of personality and social psychology* 35, 4 (1977), 250.

[39] Michael I. Norton, Daniel Mochon, and Dan Ariely. 2012. The IKEA Effect: When Labor Leads to Love. *Journal of Consumer Psychology* 22, 3 (2012), 453–460. doi:10.1016/j.jcps.2011.08.002

[40] Emilia Parada-Cabaleiro, Maximilian Mayerl, Stefan Brandl, Marcin Skowron, Markus Schedl, Elisabeth Lex, and Eva Zangerle. 2024. Song Lyrics Have Become Simpler and More Repetitive over the Last Five Decades. *Scientific Reports* 14, 1 (2024), 5531.

[41] Markus Schedl, Ralph Hertwig, Antonela Tommasel, and Shahed Masoudian. 2026. Cognitive Effects and Biases in Large Language Models. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 6: Tutorial Abstracts)*, Chenghua Lin, Aline Paes, and Rodrigo Wilkens (Eds.). Association for Computational Linguistics, Rabat, Morocco, 4–6. doi:10.18653/v1/2026.eacl-tutorials.2

[42] Patrick Schilcher, Dominik Karasin, Michael Schöpf, Haisam Saleh, Antonela Tommasel, and Markus Schedl. 2025. Characterizing Positional Bias in Large Language Models: A Multi-Model Evaluation of Prompt Order Effects. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (Eds.). Association for Computational Linguistics, Suzhou, China, 20643–20664. doi:10.18653/v1/2025.findings-emnlp.1124

[43] Aaditya Singh, Adam Fry, Adam Perelman, and et al. 2026. OpenAI GPT-5 System Card. doi:10.48550/arXiv.2601.03267 arXiv:2601.03267 [cs.CL].

[44] Oswald Spengler. 1928. *The Decline of the West.* Knopf.

[45] Yasuaki Sumita, Koh Takeuchi, and Hisashi Kashima. 2025. Cognitive Biases in Large Language Models: A Survey and Mitigation Experiments. In *Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing (SAC '25)*. Association for Computing Machinery, New York, NY, USA, 1009–1011. doi:10.1145/3672608.3707812

[46] Shelley E. Taylor and Susan T. Fiske. 1978. Salience, Attention, and Attribution: Top of the Head Phenomena. In *Advances in Experimental Social Psychology*, Leonard Berkowitz (Ed.). Vol. 11. Academic Press, 249–288. doi:10.1016/S0065-2601(08)60009-X

[47] GLM-V Team, Wenyi Hong, Wenmeng Yu, and et al. 2026. GLM-4.5V and GLM-4.1V-Thinking: Towards Versatile Multimodal Reasoning with Scalable Reinforcement Learning. doi:10.48550/arXiv.2507.01006 arXiv:2507.01006 [cs.CV].

[48] Edward L. Thorndike. 1920. A Constant Error in Psychological Ratings. *Journal of Applied Psychology* 4, 1 (1920), 25–29.

[49] Weiyun Wang, Zhangwei Gao, Lixin Gu, and et al. 2025. InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency. doi:10.48550/arXiv.2508.18265 arXiv:2508.18265 [cs.CV].

[50] Lanhu Wu, Zilin Gao, Hao Fei, Mong-Li Lee, and Wynne Hsu. 2025. LEAF-Mamba: Local Emphatic and Adaptive Fusion State Space Model for RGB-D Salient Object Detection. In *Proceedings of the 33rd ACM International Conference on Multimedia* (Dublin, Ireland) *(MM '25)*. Association for Computing Machinery, New York, NY, USA, 1013–1022. doi:10.1145/3746027.3754863

[51] Chenjun Xu, Bingbing Wen, Bin Han, Robert Wolfe, Lucy Lu Wang, and Bill Howe. 2025. Do Language Models Mirror Human Confidence? Exploring Psychological Insights to Address Overconfidence in LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 25655–25672. doi:10.18653/v1/2025.findings-acl.1316

[52] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23)*. Curran Associates Inc., Red Hook, NY, USA, 46595–46623.

[53] Liu-Fang Zhou and Ming Meng. 2020. Do You See the "Face"? Individual Differences in Face Pareidolia. *Journal of Pacific Rim Psychology* 14 (Jan. 2020), e2. doi:10.1017/prp.2019.27