Can Structured Prompting Overcome Knowledge Bias in Vision-Language Models?

Kyrylo Kolesnichenko [1], Martas Ambraziunas[1]

[1] *Vilnius University, Muitinės g. 8, Kaunas, 44280, Lithuania*

Abstract

VLM models often rely on memorized knowledge rather than image analysis, getting only 17% accuracy on counterintuitive counting problems. Prior work demonstrated this on closed API SOTA models, but it remains unclear whether the same bias appears in smaller open-source models and whether more advanced prompting strategies can overcome it. We evaluate four open-source VLM with parameters ranging from 8B to 12B on the VLMBias benchmark and five prompting strategies: baseline, debiased, zero-shot chain-of-thought, compositional chain-of-thought, and double-check. Results showed that open-source models have similar to API-based models' patterns having accuracy and bias-aligned error rate of 17-22% and 65-85% respectively. We made 16 comparisons against the baseline and 6 remain statistically significant after Bonferroni correction. The tested strategies appeared to be model specific as they may cause opposite results for different models. These results suggest that prompting alone is not enough to reliably reduce knowledge bias in VLMs.

Keywords

Vision-language models, knowledge bias, prompting strategies, chain-of-thought, open-source models

1. Introduction

VLM models combine a vision encoder with a large language model, which allows them to answer the questions based on the image [12]. However, since the language model remembers huge amounts of information from the training data, this knowledge can overcome the visual analysis of the image. For example, if you show an Adidas logo with 4 stripes and ask the model how many stripes there are, the model might answer 3, based on the memorized knowledge and not image analysis.

Vo et al. [1] demonstrated this issue with the VLMBias benchmark, which consists of 2784 counterintuitive images in 7 domains. They tested 5 closed SOTA API-based models and received an average 17.05% accuracy on counting tasks and a bias-aligned error rate of 75.70%. Authors also tested 2 simple prompting strategies, debiased and double-check, that gave +1.87 and +2.70 percentage points improvement. However, Vo et al. did not test any open-source models and did not use more comprehensive prompting strategies like chain-of-thought [11, 2] and compositional methods [3].

In this work, we evaluate 4 open-source VLM models, with parameters ranging from 8B to 12B, on the full VLMBias benchmark with 5 prompting strategies of different complexity. We apply the McNemar test [10] and the Bonferroni correction to find statistical significance. Our main results:

- Open-source models show approximately the same accuracy and bias-aligned error rate as API-based models, 17-22% and 65-85%, respectively.
- Out of 16 prompt comparisons, only 6 showed statistical significance.
- No prompting strategy showed consistent results across all models. The same strategy can produce completely opposite effects on different models.
- These results suggest that prompting strategies alone are not enough to reliably reduce knowledge bias.

*IVUS 2026: Information Society and University Studies 2026, May 12, Kaunas, Lithuania

kyrylo.kolesnichenko@knf.stud.vu.lt (K. Kolesnichenko); martas.ambraziunas@knf.vu.lt (M. Ambraziunas)

0009-0001-2399-2159 (K. Kolesnichenko); 0009-0007-1177-5388 (M. Ambraziunas)

Copyright © 2026 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-06

CEUR-WS.org/Vol-4317/paper19.pdf

2. Related work

2.1. Knowledge bias in VLMs

Bias in large language models is one of the core challenges associated with their use [14]. Models are trained to answer even when they are not confident, which causes hallucinations [13]. In multimodal systems, it takes the shape of a knowledge bias, when the large language model part ignores the image and answers based on what it remembers. VLMBias [1] formalized it, showing that even SOTA closed models often answer with memorized values instead of analyzing the image.

2.2. Prompting and Chain of Thought reasoning

Wei et al. [11] showed that chain-of-thought prompting, where the model reasons step by step before answering, significantly improves LLM performance on complex tasks. Kojima et al. [2] showed that adding "Let's think step by step" works without any examples across different types of tasks. For multimodal tasks, Mitra et al. [3] proposed CCoT, a two-step approach where the model first describes what it sees and then answers the question. Wang et al. [4] showed that generating multiple answers and picking the most common one makes CoT more reliable. Lanham et al. [5] found that CoT reasoning is not always faithful as models get bigger; they start to decide the answer before generating the reasoning chain and using CoT only to justify an already chosen answer.

3. Method

3.1. Models

We evaluate four open-source VLMs with parameters ranging from 8B to 12B:

- Qwen3-VL-8B-Instruct [6]
- InternVL3-8B [7]
- Gemma 3-12B-it [8]
- GLM-4.1V-9B-Thinking [9]

This combination of models was selected to compare different architectural approaches at a similar scale: Qwen3-VL uses ViT with dynamic resolution, InternVL3 is based on InternViT-300M, Gemma 3 uses SigLIP2 and GLM-4.1V-9B-Thinking AIMv2-Huge visual encoder. All models were instruction-tuned and publicly available. Notably, GLM-4.1V-9B is a reasoning model that generates a chain of thought before providing the final answer, which makes it specifically interesting for testing how reasoning interacts with knowledge bias. The main characteristics of the models are given in Table 1.

Table 1
Model specifications

Model	Parameters	Vision Encoder	Max Resolution
Qwen3-VL-8B-Instruct	8B	ViT	dynamic
InternVL3-8B	8B	InternViT-300M	448 x 448 tiles
Gemma 3-12B-it	12B	SigLIP2	896×896 crops
GLM-4.1V-9B-Thinking	9B	AIMv2-Huge	dynamic

3.2. Prompt conditions

We test five prompting strategies with different complexities, each taken from published work. The baseline prompt repeats the format used by Vo et al. [1] from the VLMBias benchmark and is a

simple counting question without additional instructions. The debiased prompt, also taken from Vo et al. [1], adds before the question an instruction to ignore prior knowledge and answer based only on the image. The zero-shot chain-of-thought prompt, based on the method of Kojima et al. [2], adds "Let's think step by step" before the question, making the model reason without any examples. Compositional chain-of-thought (CCoT), based on Mitra et al. [3], uses a two-step approach in which the model first makes a detailed description of what it sees on the image, and then answers the counting question using its own previously generated description as context. Double-check prompt, also taken from Vo et al. [1], asks the model to answer the question, then double-check and confirm its response. The exact prompts are given in Figure 1.



Figure 1: Prompt templates for each condition. {question} denotes the original VLMBias counting question.

3.3. Evaluations

We use VLMBias [1] as a benchmark; it consists of 2784 counting questions in seven domains: animals, logos, chess, game boards, optical illusions, flags, and patterned grids. Each question goes with an altered image in which the correct count is intentionally different from the common value. For the evaluations, we use exact match accuracy, which is defined as the percentage of questions, for which the model retrieved a numeric answer that matches the ground truth, and bias-aligned error rate, which is defined as the percentage of incorrect answers when the model's response matches the commonly memorized value.

We measure statistical significance between prompt strategies using the McNemar test [10], which gives a binary outcome for each pair. Every prompt strategy is compared to the baseline within each model. For multiple comparisons, we apply a Bonferroni correction with a significance threshold of

a = 0.05 / 4 = 0.0125 because there are four comparisons per model. We also report 95% confidence intervals computed with bootstrap using 10000 iterations.

The computations were run with vLLM[15] in different Google Colab environments with NVIDIA H100 and NVIDIA A100 GPUs. The generation limit was set to 512 tokens for Qwen3-VL-8B-Instruct, InternVL3-8B, and Gemma-3-12B-IT, and to 2048 tokens for GLM-4.1V-9B-Thinking. The maximum context length was 8192 tokens for all models.

4. Results

4.1. Overall accuracy

The exact-match accuracy results for four models and five prompt conditions are shown in Table 2.

Table 2
Exact match accuracy (%) across prompt conditions. Best result per model in bold.

Model	Baseline	Debiased	CoT	CCoT	Double-check
Qwen3-VL-8B-Instruct	19.32	20.19	**20.87**	19.65	19.25
InternVL3-8B	20.15	21.01	16.70	**22.59**	21.05
Gemma 3-12B-it	17.56	**19.11**	19.04	17.28	17.74
GLM-4.1V-9B-Thinking	22.95	20.73	**24.64**	20.01	23.31

Despite differences in the prompting strategies, all models show very low accuracy in the range from 16.70% to 24.64%, which shows that knowledge bias remains even after prompting. The best result from all model and prompt combinations was achieved by GLM-4.1V-9B-Thinking with CoT at 24.64%, while the worst was achieved by InternVL3-8B with the same condition at 16.70%, which is notable since both extremes come from the same strategy on different models. Overall, the received accuracy levels are comparable to the results of closed API-based models reported by Vo et al. [1], which suggests that this pattern is not tied to a particular model family.

4.2. Statistical significance

Out of 16 comparisons, only 6 showed statistical significance for the McNemar test with Bonferroni correction. Significant results are shown in Table 3. Qwen3-VL did not have a single significant comparison. InternVL3 CoT reduced the accuracy by 3.45 percentage points, and CCoT increased it by 2.44 percentage points. Gemma 3 debiased increased accuracy by 1.54 percentage points. GLM-4.1V-9B-Thinking had the most changes: CoT improved accuracy by 1.69 percentage points, while debiased and CCoT reduced it by 2.23 and 2.95 percentage points. No strategy worked consistently across all models.

Table 3
Statistically significant comparisons in percentage points relative to baseline.

Model	Prompt	pp	p-value
InternVL3-8B	CoT	-3.45	<0.0001
InternVL3-8B	CCoT	+2.44	0.0001
Gemma 3-12B-it	Debiased	+1.54	0.0011
GLM-4.1V-9B-Thinking	Debiased	-2.23	0.0003
GLM-4.1V-9B-Thinking	CoT	+1.69	0.0015
GLM-4.1V-9B-Thinking	CCoT	-2.95	<0.0001

4.3. Bias-Aligned Error Analysis

Bias-aligned error rate shows that a share of incorrect answers matches the remembered values. In all tested cases, the value stays high, from 64.94% to 84.77%. GLM-4.1V-9B-Thinking having the highest overall accuracy has the highest bias-aligned error rate, 82–85%. The biggest difference was in InternVL3 with CoT: bias-aligned rate dropped from 75.38% to 64.94%. As a result, the model does not start seeing better but starts making a wider set of mistakes. Results are shown in Figure 2.

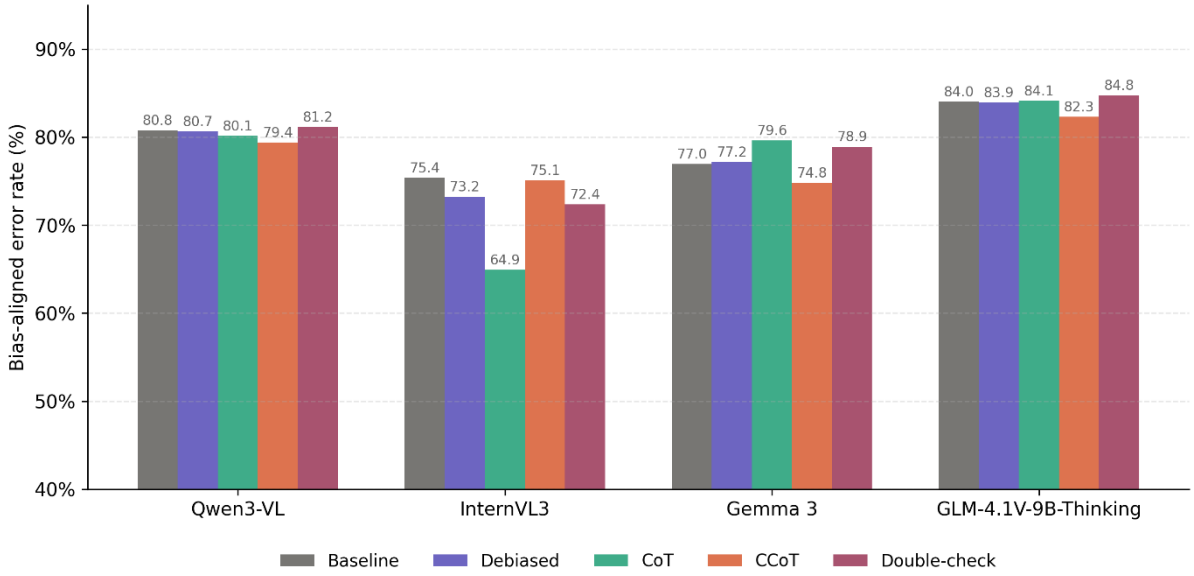


Figure 2: Bias-aligned error rate (%) across models and prompt conditions

4.4. Per domain analysis

Per-domain accuracy with baseline condition shows very different results. Animals and Chess Pieces received around zero, ranging from 0 to 1.83% and from 0 to 7.29% for all models, so prompting for these domains does not help at all. Optical Illusions have stable results at around 50% for all models, which is likely related to the perceptual complexity of the images rather than knowledge bias. Flags show the biggest difference between models, GLM-4.1V-9B-Thinking gets 38.75%, while Gemma 3 only 13.75%. Results are shown in Table 4.

Table 4
Exact match accuracy (%) by domain under baseline condition.

Model	Animals	Chess Pieces	Flags	Game Boards	Logos	Optical Illusions	Patterned Grids
Qwen3-VL-8B-Instruct	1.83	0.00	17.92	7.74	5.07	50.88	14.29
InternVL3-8B	0.00	0.00	17.92	6.55	16.43	50.25	12.20
Gemma 3-12B-it	0.00	0.00	13.75	7.74	6.04	50.63	5.06
GLM-4.1V-9B-Thinking	0.18	7.29	38.75	20.24	0.24	54.80	16.37

5. Discussion and Conclusion

We tested four open-source VLM models with the VLMBias benchmark using five prompting strategies. Results showed the same pattern as closed API-based models [1], getting 17% to 22% accuracy and 65% to 85% bias-aligned errors. Out of 16 prompt comparisons 6, were statistically

significant. Significant effects are model-specific, and different prompting strategies can have opposite effects when applied to different models. For the domains with the strongest real-world associations, like Animals and Chess Pieces, all model-prompt combinations showed near-zero accuracy.

Our work limitations include a limited model size in the range from 8B to 12B parameters, responses were limited by token count, and temperature was set to 0 for all experiments. Future work could explore visual prompting approaches, test with larger and more diverse models and look into architectural solutions like stronger multimodal alignment, since prompting alone seems insufficient.

Overall, knowledge bias in VLMs appears to be a structural problem that prompting strategies alone cannot solve.

6. Declaration on Generative AI

During the preparation of this work, the authors used Claude (Anthropic) for coding assistance in the experiment pipeline, grammar and spelling checks, and basic proofreading. The authors reviewed and edited all AI-generated content and take full responsibility for the content of this publication.

7. References

[1] A. Vo, K.-N. Nguyen, M. R. Taesiri, V. T. Dang, A. T. Nguyen, D. Kim, Vision Language Models are Biased, in: Proceedings of the International Conference on Learning Representations (ICLR), 2026. URL: https://openreview.net/forum?id=DG4S2OlGQA.

[2] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large Language Models are Zero-Shot Reasoners, in: Advances in Neural Information Processing Systems 35, 2022. URL: https://arxiv.org/abs/2205.11916.

[3] C. Mitra, B. Huang, T. Darrell, R. Herzig, Compositional Chain-of-Thought Prompting for Large Multimodal Models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 14420–14431.

[4] X. Wang, J. Wei, D. Schuurmans, Q. V. Le, E. H. Chi, S. Narang, A. Chowdhery, D. Zhou, Self-Consistency Improves Chain of Thought Reasoning in Language Models, in: International Conference on Learning Representations, 2023. URL: https://openreview.net/forum?id=1PL1NIMMrw.

[5] T. Lanham, A. Chen, A. Radhakrishnan, et al., Measuring Faithfulness in Chain-of-Thought Reasoning, 2023. doi:10.48550/arXiv.2307.13702.

[6] Qwen Team, Qwen3-VL Technical Report, 2025. URL: https://github.com/QwenLM/Qwen3-VL. doi:10.48550/arXiv.2511.21631.

[7] W. Wang, Z. Gao, L. Gu, et al., InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency, 2025. doi:10.48550/arXiv.2508.18265.

[8] Gemma Team, Gemma 3 Technical Report, 2025. URL: https://storage.googleapis.com/deepmind-media/gemma/Gemma3Report.pdf. doi:10.48550/arXiv.2503.19786.

[9] GLM-V Team, GLM-4.5V and GLM-4.1V-Thinking: Towards Versatile Multimodal Reasoning with Scalable Reinforcement Learning, 2025. URL: https://github.com/zai-org/GLM-V. doi:10.48550/arXiv.2507.01006.

[10] Q. McNemar, Note on the sampling error of the difference between correlated proportions or percentages, Psychometrika 12 (1947) 153–157. doi:10.1007/BF02295996.

[11] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, D. Zhou, Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, in: Advances in Neural Information Processing Systems 35, 2022. URL: https://proceedings.neurips.cc/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

[12] F. Bordes, R. Y. Pang, A. Ajay, et al., An Introduction to Vision-Language Modeling, 2024. URL: https://ai.meta.com/research/publications/an-introduction-to-vision-language-modeling/.

[13] A. T. Kalai, O. Nachum, S. S. Vempala, E. Zhang, Why Language Models Hallucinate, 2025. URL: https://openai.com/index/why-language-models-hallucinate/.
[14] Y. Guo, M. Guo, J. Su, Z. Yang, M. Zhu, H. Li, M. Qiu, S. S. Liu, Bias in Large Language Models: Origin, Evaluation, and Mitigation, 2024. doi:10.48550/arXiv.2411.10915.
[15] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang, I. Stoica, Efficient Memory Management for Large Language Model Serving with PagedAttention, in: Proceedings of the 29th Symposium on Operating Systems Principles, SOSP '23, ACM, New York, NY, 2023, pp. 611–626. doi:10.1145/3600006.3613165.