What Do Little Machines in Spring 2026 Know About Buddhist History?

Marcus Bingenheimer, Temple University

Abstract: The availability of open-weights language models has led to a profusion of language models trained for particular tasks. These relatively small, but increasingly powerful, models run on consumer hardware and can be used freely for the price of electricity. Because the training data for most models are not made public, it is difficult to say how much knowledge about a given domain is embedded in any model; even if we knew more about the training data, it is not always clear how much of it a model can externalize.

This paper is a first step towards developing a benchmark to test the historical knowledge of language models in the domain of Buddhist Studies. We test representative models regarding their knowledge of Chinese and Japanese Buddhist history on consumer hardware. A test bank, subdivided by historical period, with multiple choice questions (MCQs) is used with different model series. Ranking the answers produces a "winner" that "knows" most about Buddhist history. We find that, in Spring 2026, the Qwen series emerges as a winner for models in the 30B range, while the larger Kimi2.5 models lead in a cloud-based setup.

Keywords: LLMs, benchmarks, open weights models, Buddhist History, Chinese Buddhism, Japanese Buddhism.

Introduction

Recent years have seen a proliferation of benchmarks evaluating Large Language Models (LLMs) on a variety of different skills and knowledge domains. As a first step towards benchmarking language models for domain knowledge in Buddhist

 https://doi.org/10.66775/9a5wbp16

history, this paper tries to evaluate how much open-weights language models ("little machines")[1] know unaided about the history of Buddhism in East Asia.

Knowledge Rather than Skills

We will focus on evaluating the knowledge encoded in LLMs not on skills—being fully aware that the difference between the two is, to a degree, arbitrary. Without implying that neural nets (of any kind) simply "lookup" or somehow "find" objects in a database of facts, there still seems to remain a difference, if only in degree, between testing for knowledge and testing for skills. Simply knowing something feels different from figuring out an answer in a series of steps. Where benchmarks test skills rather than knowledge—like with math, coding, translation, or essay writing—an answer must be computed, constructed. There is usually no known, correct answer that could be adduced from memory. The closest equivalent for "memory" here is the knowledge that is encoded in LLMs as part of different stages of the training process.

So far, the domain knowledge of LLMs is comprehensively benchmarked only for a few, usually profitable domains, such as law and medicine. Historians, linguists, philosophers, architects, biologists, archaeologists, musicologists, geographers, anthropologists etc. might also want to learn how much these models actually know about their field of interest.

The provisional distinction between skill and knowledge proposed here extends to education. First the internet and now AI have resulted in a widening gap between what students can achieve alone, with pen and paper, and what they can do aided by computation. Although education needs to attend to both modes, in evaluation we usually test them separately. A closed book exam tests knowledge; a take-home essay tests the ability to use all available resources.

---

[1] I dispense with quotation marks for cognitive operations here and below. Obviously, we have built machines that imitate knowing and understanding and, in that sense, their "knowing" etc. differs from that of animals. However, semantics are malleable—a machine knows things like a car runs and a plane flies.

Open-weights Models Rather than Commercial Models

Why limit the evaluation to open-weights models? Open-weights models are the closed-book exam equivalent for model evaluation. There are several useful limitations to bare-weights models:

> 1. No resources beyond the weights: When asked a question, the APIs of commercial, frontier models (GPT, Claude, Gemini etc.) operate rather like people in a library, in the sense that they connect to a wider network of information to perform the task (answer). They either search the internet or consult dedicated databases built for them. Open-weights models run either locally or in the cloud, and, by default, do not connect to the internet, run agentic tasks, or draw on sources beyond themselves. What they know is encoded in their weights. Like a human in a closed book exam, they are unable to call a friend or consult a dictionary.
>
> 2. Guardrails, censorship, and hidden knowledge: Simply knowing something differs from performing the task of answering. A human might know something, but may not be able to answer because of outside coercion, or might not want to answer because of ethical reasons. Commercial models are usually constrained by guardrails on various levels: during the training process, at inference time, and on the application level. [2] Open-weights models are only subject to constraints hard-coded in during their training, such as training data filtering and reinforcement learning via human or AI feedback (RLHF/RLAIF). However, usually no guardrails are imposed at inference time or at the application level.[3]

---

[2] See Dong et al., "Safeguarding Large Language Models: A Survey," for an overview of current practice.

[3] Recent research has observed another phenomenon that belongs here among the "unknown knowns": at times language models confidently state a wrong fact, while having access to the

3. Learning: Most humans tend to improve their results when they take the same test repeatedly. They learn, if not the specific, at least the general type of questions and come better prepared for the next test. Commercial APIs can remember conversations (more so if one pays for it), and all input data can, in principle, be used to train the next generation of the model, or be included in databases that are used by the model owners to improve the output at inference time. Open-weights models, on the other hand, only have a relatively limited context window, which is usually not retained when the model is disconnected. In the long run, open-weights models do not learn anything once the context buffer is gone and the next conversation starts anew. Such a lack of continuity and context is tragic for humans,[4] but useful for testing machines that are not supposed to learn answers and testing strategies.

4. Reasoning: The introduction of reasoning models in 2025 made LLMs more human-like in their reaction to multiple choice answers. The chances to answer MCQs correctly increases when "narrowing down" the options. Reasoning models do just that: they reason their way towards the correct answer of an MCQ, just as humans do. First the impossible, then the implausible answers are eliminated, then the most likely correct answer is given. Since we would like to test bare knowledge as much as possible, it would be ideal to be able to switch off this type of reasoning, but this is usually not an option. Some models allow parameters, like /set no-think or /no_thinking, but it is

correct answer. They have "hidden" knowledge, which they seem to be unable to fully externalize. See Gekhman et al., "Inside-Out: Hidden Factual Knowledge in LLMs"; Orgad et al., "LLMs Know More Than They Show"; and Simhi et al., "Trust Me, I'm Wrong: LLMs Hallucinate with Certainty Despite Knowing the Answer."

[4] Some humans suffer from a condition—anterograde amnesia—that forces them to live a relatively small context window (see e.g. the well documented case of Clive Wearing (described in Sacks, *Musicophilia: Tales of Music and the Brain*, 187–213), or the fictional Professor in Yokō Ogawa's *The Housekeeper and the Professor*).

unknown to what degree this prevents reasoning in the above sense in practice. In the end, this might be simply a concomitant to MCQs, which are indeed considered easier than straightforward questions because they do allow for an amount of reasoning over the probabilities of a limited set of possible answers.

To summarize: We should care about the capabilities of open-weight models, because these are widely available tools which are not directly controlled by third-party interests.[5] Like with humans in a closed book exam, open-weights models (by default) do not access heuristics beyond their own neural network. Like humans, these models exist within the limitations of their previous training. Conveniently different from humans, models can be tested repeatably, and, with the Temperature parameter set to zero, return the same responses. Commercial APIs, on the other hand, can be modified or discontinued by the service owners without notice. Lastly, like with humans, it is not possible—and probably not desirable—to stop models from reasoning over MCQs.

Existing Knowledge Benchmarks for Adjacent Domains

As of Spring 2026, there are dozens of general and specialized benchmarks, with new ones appearing every few months. In the space of Buddhism, 2025 saw advances in benchmarking LLMs on NLP tasks with Buddhist Sanskrit and Tibetan texts as well as in the evaluation of machine translation from Buddhist Chinese.[6] However, there is as of yet no benchmark evaluating a model's knowledge of

[5] This is not to say that open-weights models to not reflect political, social, or economic agendas as can be seen in the different answers that Chinese and American models give in response to "sensitive" questions. It will be interesting to observe in the future to what degree different nations manage to train models to present their own narratives.

[6] Hashiloni et al., "DharmaBench: Evaluating Language Models on Buddhist Texts in Sanskrit and Tibetan" and Nehrdich et al., "MITRA-zh-eval: Using a Buddhist Chinese Language Evaluation Dataset to Assess Machine Translation and Evaluation Metrics."

Buddhist history.[7] In fact, history in general is not exactly a focus of model evaluation.

Nevertheless, at least one history specific model benchmark, the History Seshat Test for LLMs (Hist-LLM), has been created on a subset of the Seshat Global History Databank. In keeping with the design of Seshat, the four options ask merely whether a particular feature can be said to have been present, inferred present, inferred absent, or absent in a particular time and place (e.g. hand-held firearms during the Tang).

More widely used is the Measuring Massive Multitask Language Understanding (MMLU) which is currently one of the most popular benchmarks. MMLU uses a four-option multiple choice format.[8] Its fifty-seven subjects include only five somewhat related to our context: high school US history (203 questions), high school world history (236 questions), philosophy (310 questions), prehistory (323 questions), and world religion (170 questions). A Chinese version of the MMLU, the CMMLU, has been created. The CMMLU translates and adapts the existing test and adds a large number of China specific subjects, including Chinese History (323 questions) and Chinese Literature (201 questions).

Other benchmarks geared towards Chinese language models are Chinese Generation Evaluation (CG-Eval) and the Large-scale Holistic Multi-subject Knowledge Evaluation (LHMKE). CG-Eval includes 400 middle and high school level Chinese history questions, but not in four-option format.[9] It mostly consists of term explanations and short essay questions, which are to evaluated by frontier models. Likewise, LHMKE goes beyond MCQs to test a wider variety of skills, relying on GPT 3.5 and 4 as scorer.[10] The data is said to be sourced from standard exams of different educational levels (elementary and high school, college, and

---

7 Currently, the most sustained effort in model development in Buddhist Studies is the Dharmamitra project: https://dharmamitra.org/.

8 Hendrycks et al., "Measuring Massive Multitask Language Understanding."

9 Dataset here: https://huggingface.co/datasets/Besteasy/CG-Eval/tree/main (2025-10).

10 Liu et al., "LHMKE: A Large-scale Holistic Multi-subject Knowledge Evaluation Benchmark for Chinese Large Language Models."

"career development"), but the github repo does not contain the data promised (p. 10,477).

Finally, the largest current Chinese LLM benchmark is Xiezhi, the "ever-updating benchmark," which promises around 250,000 questions, of which some 7% (approximately 17,500 questions) pertain to Chinese history.[11] It does include MCQs as well as comprehension questions.

Dataset

The current benchmark is designed to test open-weights models about the history of Buddhism in China and Japan. As of Spring 2026, it consists of 210 MCQs in a four-option format with groundtruth. The questions themselves are not very difficult if one has access to professional reference works. Without access to reference works, on the other hand, we expect that professional historians would struggle to score more than 75% with these relatively random but detailed questions.

The Chinese part of the test bank is organized in five temporal bands of thirty questions each: 1: Han to Northern and Southern Dynasties (c. 50–600 CE),[12] 2: Sui and Tang (c. 580–920 CE), 3: Song and Yuan (c. 900–1400 CE), 4: Ming (c. 1350–1650 CE), 5: Qing and Early Republic (c. 1640–1950 CE). The Japanese part has so far twenty questions in three bands: Asuka and Nara (c. 570–800 CE), Heian (c. 800–1150 CE), and Kamakura (c. 1150–1350 CE). This allows us to evaluate the diachronic distribution of the embedded knowledge. In principle, a model could score highly on a particular period but lack knowledge about others.

Example questions:

[11] Gu et al., "Xiezhi: An Ever-Updating Benchmark for Holistic Domain Knowledge Evaluation."

[12] All dates given for the bands are approximate and do not conform to the exact dates for dynasties or periods. This is because religious developments and biographies at times span different periods.

> “Which Chinese dynasty saw the first major suppression of Buddhism?” Han 漢, Northern Wei 北魏, Liang 梁, Sui 隋.
>
> “Which Song emperor first supported Buddhism but later restricted monastic landholdings?” Taizu 宋太祖, Taizong 宋太宗, Huizong 宋徽宗, Xiaozong 宋孝宗.
>
> “Which Ming emperor officially restored the ordination platform at Mount Wutai 五台山?” Hongwu 洪武帝, Yongle 永樂帝, Jiajing 嘉靖帝, Wanli 萬曆帝.
>
> “Which of these four was a famous poet-monk of the late Qing?” Hanshan 寒山, Daomin 道忞, Jing'an 敬安, Hu Shi 胡適.

As the examples show, most questions and answers contain CJK characters. We assume that, for this domain, a test that gives names and terms in Pinyin only is not meaningful. The overall English design aims to connect the test with the international scholarship as well as preparing the extension of the benchmark to other regions (Korea, Vietnam, Southeast Asia, Tibet, India, etc.) which will come with their own languages and scripts.

The test bank is intended as a first step towards a more extensive domain-specific benchmark for Buddhist Studies. Ideally, a future Buddhist Studies benchmark would be comprised of questions not only about different regions but also different approaches (philosophy, anthropology, philology, etc.). To prevent data contamination, the test bank is currently available only on request for academic use.

Experiments

Two series of experiments were conducted in October 2025 and April 2026 using Ollama as middleware. We focused on the most impactful open-weights model series: Deepseek, Gemma, Llama, Mistral, Qwen, Kimi, Glm, and Phi.

Results Fall 2025

We tested nineteen models with maximum 32b parameters locally with 150 questions regarding Buddhist history in China.

MODEL	Han to c. 589 CE	Sui and Tang (c. 590–950 CE)	Song and Yuan (c. 950–1370 CE)	Ming (c. 1370–1650 CE)	Qing and Republic (c. 1650–1950 CE)	TOTAL
gemma2:9b	10/30 (33%)	20/30 (67%)	17/30 (57%)	20/30 (67%)	14/30 (47%)	81/150 (54%)
gemma3:4b	8/30 (27%)	12/30 (40%)	14/30 (47%)	16/30 (53%)	10/30 (33%)	60/150 (40%)
gemma3:12b	19/30 (63%)	18/30 (60%)	22/30 (73%)	16/30 (53%)	18/30 (60%)	93/150 (62%)
gemma3:27b	15/30 (50%)	24/30 (80%)	22/30 (73%)	20/30 (67%)	15/30 (50%)	96/150 (64%)
phi3:3.8b	13/30 (43%)	17/30 (57%)	17/30 (57%)	16/30 (53%)	12/30 (40%)	75/150 (50%)
phi3:14b	10/30 (33%)	20/30 (67%)	21/30 (70%)	17/30 (57%)	13/30 (43%)	81/150 (54%)
phi4:14b	14/30 (47%)	20/30 (67%)	20/30 (67%)	19/30 (63%)	13/30 (43%)	86/150 (57%)
qwen2.5:7b	18/30 (60%)	22/30 (73%)	21/30 (70%)	19/30 (63%)	17/30 (57%)	97/150 (65%)
qwen2.5:14b	**23/30 (77%)**	25/30 (83%)	22/30 (73%)	**23/30 (77%)**	**21/30 (70%)**	**114/150 (76%)**
qwen2.5:32b	**25/30 (83%)**	**28/30 (93%)**	21/30 (70%)	**23/30 (77%)**	**23/30 (77%)**	**120/150 (80%)**
qwen3:8b	17/30 (57%)	24/30 (80%)	**24/30 (80%)**	17/30 (57%)	20/30 (67%)	102/150 (68%)
qwen3:14b	**21/30 (70%)**	**27/30 (90%)**	**25/30 (83%)**	**21/30 (70%)**	20/30 (67%)	**114/150 (76%)**
qwen3:32b	18/30 (60%)	**28/30 (93%)**	**23/30 (77%)**	20/30 (67%)	**21/30 (70%)**	110/150 (73%)
llama3:8b	14/30 (47%)	21/30 (70%)	18/30 (60%)	18/30 (60%)	13/30 (43%)	84/150 (56%)
llama3.1:8b	12/30 (40%)	24/30 (80%)	19/30 (63%)	17/30 (57%)	13/30 (43%)	85/150 (57%)
mistral:7b	9/30 (30%)	22/30 (73%)	16/30 (53%)	13/30 (43%)	12/30 (40%)	72/150 (48%)
mistral:8x7b	16/30 (53%)	23/30 (77%)	19/30 (63%)	23/30 (77%)	17/30 (57%)	98/150 (65%)
AVERAGE	15.4/30 (51%)	22.1/30 (74%)	20.1/30 (67%)	18.7/30 (62%)	16.0/30 (53%)	18.4/30 (61%)

MODEL	Han to c. 589 CE	Sui and Tang (c. 590–950 CE)	Song and Yuan (c. 950–1370 CE)	Ming (c. 1370–1650 CE)	Qing and Republic (c. 1650–1950 CE)	TOTAL
(deepseek-r1:8b)	8/30 (27%)	7/30 (23%)	10/30 (33%)	6/30 (20%)	11/30 (37%)	42/150 (28%)
(deepseek-r1:14b)	7/30 (23%)	8/30 (27%)	10/30 (33%)	6/30 (20%)	8/30 (27%)	39/150 (26%)

Table 1: Results of different language models Fall 2025

Discussion

Surprisingly, both versions of Deepseek-r1 tested were unable to deal with the MCQs and defaulted on option "a" to all MCQs (with some exceptions in the 8b version). On checking responses to individual questions in chat mode, it turned out that Deepseek struggled with parsing the mixture of English and Chinese characters in the MCQs. This might have been fixed with a different prompt, but for this particular round of tests, we accepted Deepseek merely as a useful control for random answers (25%).

Apart from Deepseek, the picture was reassuringly consistent: newer generations of a model series perform somewhat better than older ones, and larger model versions return more correct answers than smaller ones. Especially pronounced is a 17% improvement from Mistral 7b to the mixture of experts architecture Mixtral 7x8b by Mistral AI. This shows how novel architectures might lead to significant gains in externalizing the knowledge embedded in models.

The clear winner in Fall 2025 was the Qwen series. The 14b versions of both Qwen2.5 and Qwen3 return the best results for that model size. With a top score of 80%, Qwen2.5:32b performs at a level that, in my view, most human experts would struggle to achieve in a closed-book exam. Both Qwen2.5 and Qwen3 show a clear increase from the default models (Qwen2.5:7b and Qwen3:8b) to the 14b versions. Qwen2.5:32b manages to improve this score slightly, whereas—somewhat

surprisingly—Qwen3:32b actually performs less well than Qwen3:14b. The success of Qwen in Fall 2025 was in line with its overall momentum in the field of open-weights models, where it was considered "the new standard" by independent observers.[13] Qwen is known for strong multilingualism, especially Chinese and English, and the dynamic switching of thinking and no-thinking modes—both traits that might be helpful in a MCQ test where the questions are in mixed English and Chinese.

Interestingly, the test also revealed an imbalance in the knowledge distribution between periods. The lowest averages are those for the earliest period (Han to end of Northern and Southern Dynasties) and the most recent (Qing and Republican) period. Assuming the questions across periods are of the same difficulty, it would follow that the models knew more about the time between c. 590 CE and 1650 CE than about Chinese Buddhism before or after. I have no explanation for this pattern.

Results Spring 2026

In spring 2026, we tested fourteen local and nine cloud-based models with 210 questions about the Buddhist history of China (150 questions) and Japan (60 questions). Ollama was again used as middleware. Below we included some of the better models of Fall 2025 to show how they compare with their successors and to show that the addition of new questions did not significantly alter the performance of the models.

13 See Nathan Lambert, https://www.interconnects.ai/p/qwen-3-the-new-open-standard (2010-10).

Model	Test 2025-10 (150 questions)	Test 2026-04 (210 questions)
Locally run		
deepseek-r1:14b		128/210 (61%)
gemma3:12b	93/150 (62%)	134/210 (64%)
gemma3:27b	96/150 (64%)	139/210 (66%)
gemma4:e4b		121/210 (58%)
gemma4:31b		**159/210 (76%)**
glm-4.7-flash:30b[3]		158/210 (75%)
mistral-nemo:12b		124/210 (59%)
mixtral:8x7b	98/150 (65%)	151/210 (72%)
phi4:14b	81/150 (54%)	132/210 (63%)
qwen2.5:14b	**114/150 (76%)**	158/210 (75%)
qwen2.5:32b	**120/150 (80%)**	**165/210 (79%)**
qwen3:14b	**114/150 (76%)**	158/210 (75%)
qwen3.5:9b		153/210 (73%)
qwen3.5:27b		**180/210 (86%)**
Cloud-based		
deepseek-v3.2:671b[37b]		**191/210 (91%)**
devstral-2:123b		171/210 (81%)
gpt-oss:120b		77/210 (37%)
glm-4.7:355b		188/210 (90%)
glm-5.1:754b		**192/210 (91%)**
kimi-k2-thinking:1t[32]		**197/210 (94%)**
kimi-k2.5:1t[32]		187/210 (89%)
mistral-large-3:675b		185/210 (88%)
qwen3.5:397b		187/210 (89%)

Table 2: Results of different language models Spring 2026

Discussion

Within a model series, not surprisingly, larger models again clearly performed better than smaller models. Larger models of an earlier generation can be better than smaller models of a later generation. There are clear differences in strength between different series, indicating that the experiments do indeed pick up a signal.

Qwen3.5:27b managed to improve over its Fall 2025 champion qwen2.5:32b, showing that the mid-sized models of the Qwen series are still gaining knowledge in this domain. In fact, with 86%, qwen3.5:27b is already performing beyond human expert level. However, among the locally run models in this size bracket (c. 30b parameters), the newer Gemma and GLM models (gemma4:31b, glm-4.7-flash:30b) are coming close. It will be interesting to see whether they will catch up.

Larger open-weight models were tested in the (Ollama) cloud. Cloud-based models achieved higher benchmarks than local models. The downside of larger models are higher costs and a lack of privacy where the models are not hosted on a private server. Response times of cloud-based models are comparable to local, but queues and connection problems can make them less predictable.

Significantly, seven out of the nine tested cloud-based models returned over 85% correct answers. Kimi-k2-thinking, which is said to run on a 1 trillion parameter MoE architecture (32b active), leads the field (94%)—but not by much. Kimi-k2-thinking in Spring 2026 approaches the power of frontier models. The advantage for the user is that we retain control over parametrization and agentic behavior. It appears that larger open-weights models already encode detailed domain knowledge and will soon be able to saturate this type of test.

Conclusions

All models—even smaller ones—performed significantly better than random (25%) at least since Fall 2025. Obviously, some knowledge is present in all models. As one would expect, within each model series, larger models generally outperform smaller models.

For the time being, the Qwen series emerged as the winner. It appears to have encoded knowledge regarding Chinese and Japanese Buddhist history comprehensively and (or?) is able to best externalize this knowledge in a MCQ test. While there was no significant difference between qwen2.5 and qwen3 in the c.30b bracket, qwen3.5:32b improves the score by approximately 10%.

Cloud-based models greater than 120b, which cannot be run locally on consumer hardware, are surpassing local models and are poised to saturate the benchmark soon. Next to expanding the test bank to different regions, we will have to consider changing the format of the questions to make the task more challenging for the models. MCQ-based tests are rather uni-dimensional; they can measure knowledge to a degree, but how does this ability translate to other aspects of research? Qwen might be the most knowledgeable, but can it also suggest the best secondary sources or write the best essays in this domain? What must a true "Buddhist Studies" benchmark for LLMs measure and how?

Bibliography

Dong, Yi, Ronghui Mu, Yanghao Zhang, et al. "Safeguarding Large Language Models: A Survey." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20, no. 10 (May 2024): 1–21. https://doi.org/10.48550/arXiv.2406.02622

Gekhman, Zorik, Eyal Ben David, Hadas Orgad, et al. "Inside-Out: Hidden Factual Knowledge in LLMs." *Proceedings of the Conference on Language Modeling (COLM) 2025*. https://doi.org/10.48550/arXiv.2503.15299

Gu, Zhouhong, Xiaoxuan Zhu, Haoning Ye, et al. "Xiezhi: An Ever-Updating Benchmark for Holistic Domain Knowledge Evaluation." arXiv:2306.05783v3 (March 2024): https://arxiv.org/html/2306.05783v3

Hashiloni, Kai Golan, Shay Cohen, Asaf Shina, et al. "DharmaBench: Evaluating Language Models on Buddhist Texts in Sanskrit and Tibetan." 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, pages 2088–110. Mumbai: The Asian Federation of Natural Language Processing and The Association for Computational Linguistics, 2025.

Hendrycks, Dan, Collin Burns, Steven Basart, et al. "Measuring Massive Multitask Language Understanding." *Proceedings of the International Conference on Learning Representations (ICLR), 2021*. https://doi.org/10.48550/arXiv.2009.03300

Li, Haonan, Yixuan Zhang, Fajri Koto, et al. "CMMLU: Measuring massive multitask language understanding in Chinese." arXiv:2306.09212 (2024): https://arxiv.org/abs/2306.09212

Liu, Chuang, Renren Jin, Yuqi Ren, et al. "LHMKE: A Large-scale Holistic Multi-subject Knowledge Evaluation Benchmark for Chinese Large Language Models." *LREC-COLING 2024*, 10 (May 2024): 10476–10487. https://aclanthology.org/2024.lrec-main.916.pdf

Nehrdich, Sebastian, Avery Chen, Marcus Bingenheimer et al. "MITRA-zh-eval: Using a Buddhist Chinese Language Evaluation Dataset to Assess Machine Translation and Evaluation Metrics" In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities (NLP4DH 2025)*, pages 129–37. Albuquerque: Association for Computational Linguistics, 2025.

Orgad, Hadas, Michael Toker, Zorik Gekhman, et al. "LLMs Know More Than They Show: On the Intrinsic Representation of LLM Hallucinations." In *Proceedings of the International Conference on Learning Representations (ICLR)*, arXiv:2410.02707 (2025): https://doi.org/10.48550/arXiv.2410.02707

Sacks, Oliver. *Musicophilia: Tales of Music and the Brain*. London: Knopf, 2007.

Simhi, Adi, Itay Itzhak, Fazl Barez, et al. "Trust Me, I'm Wrong: LLMs Hallucinate with Certainty Despite Knowing the Answer." *Proceedings of the International Conference on Learning Representations (ICLR)*, arXiv:2502.12964 (2025): https://doi.org/10.48550/arXiv.2502.12964

Yang, An, Anfeng Li, Baosong Yang, et al. “Qwen3 Technical Report.” arXiv:2505.09388 (2025): https://arxiv.org/abs/2505.09388

Zeng, Hui, Jingyuan Xue, Meng Hao, et al. “Evaluating the Generation Capabilities of Large Chinese Language Models.” arXiv:2308.04823 (2023): https://arxiv.org/abs/2308.04823