

Secret Dates in System Prompts Undermine Language Model Evaluation

Originally published Friday 8 October 2026 at <https://www.unite.ai/secret-dates-in-system-prompts-undermine-language-model-evaluation/>



Without the ability to [benchmark](#) Large Language Models (LLMs), it is difficult for consumers and businesses to understand what progress a model has made over recent versions, and how it stands up to its competitors:

Text Arena **Overall**

View overall rankings across various AI models in text-to-text tasks across math, coding, creative writing, and other open-ended domains.

Oct 8, 2026 8,734,330 votes 414 models

Rank	Rank Spread	Model	Score	Votes	Price \$/M	Context
1	1 - 1	gemin-4-argon-high Google - Proprietary	1525 +3	4,882	\$2 / \$10	1M
2	2 - 10	claude-opus-5.5-high Anthropic - Proprietary	1507 +8	6,272	\$4 / \$20	1M
3	2 - 8	claude-opus-4.6-high Anthropic - Proprietary	1504 +3	79,415	\$5 / \$25	1M
4	2 - 10	claude-fable-5-high Anthropic - Proprietary	1504 +4	40,036	\$10 / \$50	1M
5	2 - 11	claude-opus-4.7-high Anthropic - Proprietary	1501 +4	65,192	\$5 / \$25	1M
6	2 - 13	claude-fable-5.1-max Anthropic - Proprietary	1501 +8	13,063	\$10 / \$50	1M
7	2 - 13	claude-opus-4-6 Anthropic - Proprietary	1498 +3	84,095	\$5 / \$25	1M
8	2 - 17	gemin-3.6-flash-high Google - Proprietary	1497 +3	31,437	\$0.75 / \$3.75	1M

The influential LLM leaderboard at [arena.ai](#). [Source](#)

Since LLMs are non-deterministic (i.e., they will *not* always produce consistent outputs given the same inputs), evaluating them is tricky. Even when researchers use identical prompts, model settings and benchmark datasets, seemingly minor differences in the execution environment can produce different answers and significantly alter performance scores.

Factors such as hardware configuration, [numerical precision](#), [inference batch size](#), and even [the ordering of multiple-choice answers](#) can influence results. This can make it hard to see whether a reported improvement reflects genuine progress, or merely some semi-random variation in the conditions under which the model was tested.

To a certain extent, one can account for some of these variables, or at least establish what margin-of-error they generate, so that comparative evaluation becomes meaningful. It's important to try, since a lot of money, and a lot of reputation depends on being able to benchmark AI systems of this kind with some degree of accuracy.

However, according to new research, one particular variable can not only be destructive to benchmarks, but is also very difficult to eliminate from the prompts that define them – *today's date*.

Times Change

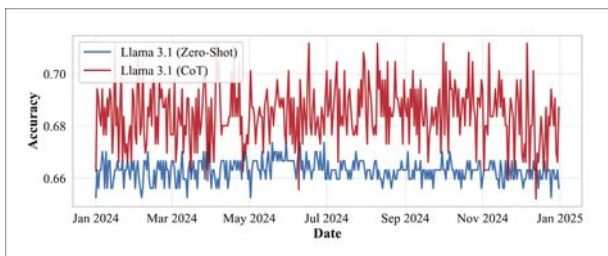
The [new paper](#)^{*}, an academic collaboration between Germany, Mexico and the USA, asserts that the fact that the current date is automatically and secretly included in the system prompt of all frontier and many deployments of [open-weight models](#) means that reproducibility could be nigh-on impossible:

'We identify a critical, often overlooked source of non-determinism in LLM evaluation: the hidden injection of the current date into system prompts.'

'Across 9 models, 6 datasets, and four tasks, this dynamic metadata alters model performance and reshuffles leaderboard rankings, surpassing the variance introduced by other system-level factors such as batch size or numerical precision.'

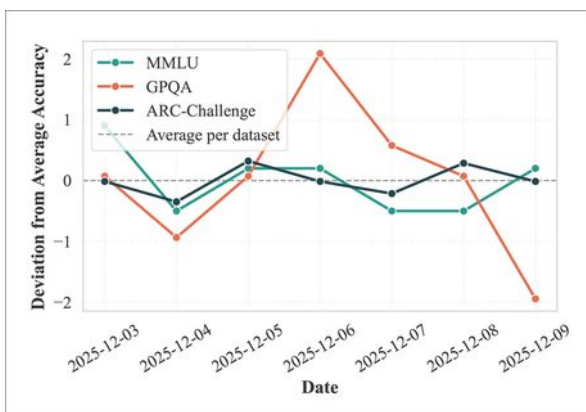
The researchers also note that the identified effect is larger for tasks requiring generated answers, with performance varying by up to 6% on multiple-choice questions, 14% on mathematical reasoning, and 7% on code generation – and with machine translation scores also varying significantly.

Providing example answers and encouraging [step-by-step reasoning](#) did not resolve the problem – in fact, the latter made it worse:



Accuracy fluctuations across different dates in 2024 for Llama 3.1 (8B) on the MMLU benchmark. Step-by-step reasoning (red) produces substantially greater variability than direct answers (blue), demonstrating how chain-of-thought prompting amplifies sensitivity to the date. [Source](#)

The effect was also identified in proprietary models, with [GPT-5.1](#) showing accuracy fluctuations of up to 4% across three multiple-choice benchmarks during a week of testing in December 2025. The researchers used an empty system prompt, disabled reasoning and set [randomness](#) to zero – but the date was still inserted automatically by the provider:



GPT-5.1's accuracy fluctuated across seven consecutive days in December 2025, despite identical user prompts and model settings. The greatest variation occurred on GPQA (orange), reaching four percentage points between the best and worst days. The dashed line represents each benchmark's average accuracy over the week.

The researchers recommend removing the date from system prompts where possible, or fixing and documenting it, to ensure fair comparisons. However, they note that date-centric influence may be more deeply-ingrained:

'One possible reason for the date sensitivity is that the system prompt might be fixed during supervised fine-tuning (SFT) and reinforcement learning from human feedback (RLHF), making the model brittle to any slight modification.'

To investigate the issue, the researchers tried six different wordings for the system prompt, and found that these changes affected accuracy just as much as changing the date (0.78%). Therefore the date appears to have as much influence as deliberate [prompt engineering](#) – except that it changes automatically, without the user even knowing

Other approaches were tried to mitigate the ‘current date’ problem, including [few-shot learning](#) (where the model got five example answers before responding). This helped a little, bringing the average variation down from 2.52% to 2.27%, but didn’t fix the problem.

Changes in GPU hardware, batch size, numerical precision, answer order and system-prompt wording were also tested. The biggest effects were seen with answer-order and prompt wording, which came closest to the impact of changing the date.

Last Days

It’s reasonable to expect that an LLM/VLM will understand the date from the very beginning of a chat; however, there seems to be no explicit reason to build it into the system prompt (the unseen rubric that imposes [guardrails](#), and conditions the LLM’s behaviors in ways inaccessible to the user) when the LLM could routinely make a sub-Kb RAG call to ingest the latest date, as a minor housekeeping routine prior to engagement with the user.

No doubt other possibilities exist to resolve the matter; however, since the imposition of the current date into the system prompt has not hitherto been seen as a problem, there has presumably been little or no investigation in regard to this.

Online Dating

For the tests, identical prompts were used for every model, with only the date in the system prompt being changed. Every day of 2024 was tested, from January 1st through to December 31st, with all other settings kept the same.

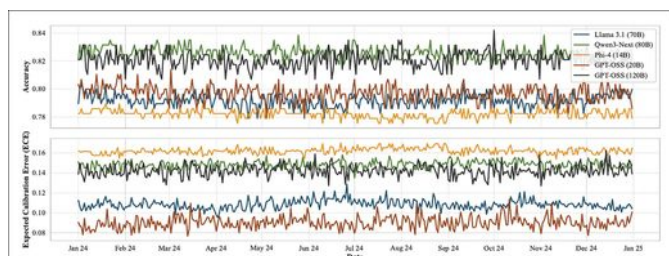
Six benchmarks were used: [MMLU](#); [GPQA](#); and [ARC-Challenge](#), for multiple-choice questions (scored by answer-token probability). [GSM8K](#) for step-by-step math (final answer checked); [HumanEval](#) for Python code generation (unit-tested); and [WMT](#) for English-to-German; English-to-Finnish; and English-to-Czech translation (full output evaluated). Time-dependent questions were excluded.

Nine models were tested: Llama 3.1 Instruct ([8B](#) and [70B](#)); Gemma 3 Instruct ([4B](#) and [27B](#)); Qwen3 ([4B](#)); Qwen3-Next ([80B](#)); Phi-4 ([14B](#)); and GPT-OSS ([20B](#) and [120B](#)).

Accuracy (the percentage of questions answered correctly) was used to score the multiple-choice and math tests, while [Expected Calibration Error \(ECE\)](#) measured how well the models’ confidence matched their actual performance.

Code was checked using `pass@1` (the percentage of generated code solutions that pass all tests on the first attempt); translations were scored using [BLEU](#) and [chrF](#).

The results confirmed the researchers’ hypothesis: simply changing the date in the system prompt changed how accurately the models answered the same questions.



Test results showing how five leading models performed on MMLU as the system-prompt date was changed throughout 2024. Accuracy is shown at the top, with expected calibration error (ECE) below. All five models showed fluctuations in both measures, despite no other changes to the test conditions. Lower ECE scores indicate better alignment between confidence and accuracy.

Along with other results shown earlier in the article, this is potentially bad news for LLM benchmarking, because a model could score better or worse depending on which day it was tested, possibly changing its position on a leaderboard, without any actual improvement or decline in its capabilities.

Conclusion

This issue highlights the divide between the deterministic computing systems we have been used to prior to around 2023, and the very different nature of diffusion-based and similar AI systems that have evolved, and continue to evolve, since then.

Date resolution was essentially solved [on January 1st 1970](#), but has, apparently, returned to haunt the world of computing in the form of [cut-off dates](#), among other [temporal concerns](#).

** Titled 'Dating the Model: Hidden Dates in System Prompts Affect LLM Evaluation'*

First published Friday, October 9, 2026

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.

Website: martinanderson.ai

Contact: martin@martinanderson.ai