FahMai-TeleBench: A Thai Grounded Tool-Use Benchmark for the Deployable Model Tier

Nutchanon Yongsatianchot[1] **Piyalitt Ittichaiwong**[1,2]
Jitkapat Sawatphol[2] **Tichanat Ruangsiri**[3]
[1]Faculty of Medicine Siriraj Hospital, Mahidol University
[2]CARIVA (Thailand) [3]Thai Oil Public Company Limited
nutchanon.yon@mahidol.ac.th, piyalitt.itt@mahidol.ac.th,
jitkapat.saw@preceptorai.tech, tichanat@thaioilgroup.com

Abstract

Multilingual agentic benchmarks remain rare, especially for under-resourced languages such as Thai, where nicknames, unspaced script, and fused honorifics change how an agent must query its tools. We present **FahMai-TeleBench**, a bilingual (Thai/English) benchmark for grounded tool use. A model answers employee-directory questions over a synthetic company directory by issuing tool calls, or refuses when the answer is absent. All 626 items are synthetic but modeled on a real Thai enterprise-directory deployment, and they are graded deterministically without an LLM judge. Our central design choice is to treat **tool availability as an experimental variable**: each item runs under four tool configurations (grep, structured search, both, and a Python REPL), so the benchmark measures not just *whether* a model answers but *how* it uses tools. The task is near-solved at the frontier (GPT-5.4 97–99%), yet sharply discriminating for the Thai-capable models an organization would host itself (Typhoon-2.5 58–74%, Typhoon-S-8B 25–37%). The Thai penalty is model-specific, and Typhoon-2.5 answers 90% of items correctly once the relevant records are placed in its context, so its failures are failures of tool use. The tool axis also exposes differences at the frontier: GPT-5.5 drops to 73% with a wide-schema structured search tool while scoring 97–98% with grep and the REPL.

1 Introduction

Most tool-use and function-calling benchmarks are English (Patil et al., 2024; Qin et al., 2024; Yao et al., 2025), so we know little about how agents behave when the language itself complicates the task. Thai is a useful stress test. People are addressed by nicknames that bear no systematic relation to their formal names, the same given name recurs across many employees, and romanizations are idiosyncratic. Because Thai is written without word spaces, an honorific title fuses with the name it precedes (คุณกมลา, "Khun Kamala"), so a literal search for the user's string returns nothing. These properties change how a model must query a tool, not merely how it reads text. We study a common enterprise setting in which they all appear at once: an employee-directory assistant.

The task is deliberately simple. A user asks for a colleague's phone extension, email, or department head, and the model answers by calling a tool over a company directory, or refuses when the directory does not contain the answer. We do not design the benchmark to defeat the strongest models; instead, we measure how existing models use combinations of common tools. This simplicity has two payoffs. First, each item has a unique gold answer computed from the directory, so grading is a deterministic string check with no LLM judge, and frontier models solving nearly every item confirms that the golds are reachable. Second, an easy task that the frontier solves still spreads models apart below it, in the *deployable tier*: open-weight models that an organization can host itself, keeping personnel records in-house, and can afford to run at assistant volume. There the spread is large: a Thai-specialized 8B model (Typhoon-S-8B) scores 25–37%, while larger open-weight models exceed 95%.

We run every item under four tool configurations: a raw-text grep, a structured search with 14 optional filter fields, both together, and a Python REPL. Holding the item fixed and varying only the toolset exposes process-level behavior that a single accuracy number erases, such as retry after a failed query and tool routing when alternatives compete. On a hard benchmark, "chose the wrong tool" is indistinguishable from "task too hard"; on an easy and well-posed one, the mechanism becomes legible.

The benchmark therefore separates two axes. The *language axis* compares Thai and English items within the same subtype, and the *tool axis* com-

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 185–196
November 6-10, 2026 ©2026 Association for Computational Linguistics

pares identical items across toolsets. This lets us ask whether multilingual ability, as measured by text-only evaluation, carries over to agentic use. For Typhoon-2.5, the Thai model we test in depth, it does not. When the relevant directory records are placed in its prompt, it answers 90% of a 149-item sample correctly. When it must find the same records with tools, it answers only 48–71% of those items.

We make four contributions. First, we release **FahMai-TeleBench**[1], a 626-item bilingual benchmark for grounded directory tool use, with deterministic grading and gold answers checked by human raters. Second, we design a question taxonomy that covers Thai-specific retrieval phenomena and spans both *depth* (multi-hop org bridges) and *breadth* (listing and aggregation), with a five-way refusal taxonomy. Third, we propose a tool-availability design, with four tool configurations per item, that measures the tool-use *process* rather than only the outcome. Finally, we evaluate twelve model variants and show that accuracy separates models mainly below the frontier, that the tool axis exposes differences at the frontier, and, with a per-language analysis and an oracle-context control, that the Thai-specialized models' failures are failures of tool use rather than of reading Thai.

2 Related Work

Tool-use benchmarks. Early work targets API and function calling: Gorilla (Patil et al., 2024) and the Berkeley Function-Calling Leaderboard (Patil et al., 2025), ToolLLM over 16,000+ REST APIs (Qin et al., 2024), and API-Bank (Li et al., 2023b). Later benchmarks add interaction and multi-hop structure: τ-bench and τ^2-bench emulate user–agent conversations under domain policies (Yao et al., 2025; Barres et al., 2026), ToolSandbox adds stateful conversational tool use (Lu et al., 2025), and ToolHop targets multi-hop tool chaining (Ye et al., 2025). Recent suites scale to 220 tools on real Model Context Protocol servers (Bandi et al., 2026) and grade autonomous-agent trajectories across nine task categories (Ye et al., 2026), and broader environments include AgentBench (Liu et al., 2024), WebArena (Zhou et al., 2024), and AppWorld (Trivedi et al., 2024). These benchmarks are almost all English and typically hold the toolset fixed. We make the toolset an experimental variable, and we target Thai.

Grounded querying and refusal. Text-to-SQL benchmarks such as Spider (Yu et al., 2018) and BIRD (Li et al., 2023a) grade an executed query against a gold result, but their queries are typically single-shot. Knowing when *not* to answer has been studied notably in SQuAD 2.0 (Rajpurkar et al., 2018), and tool-use safety in ToolEmu (Ruan et al., 2024). We keep exact, reference-based grading inside an agentic, multi-tool loop, and we add a five-way refusal taxonomy.

Thai and Southeast Asian evaluation. Evaluation for Thai and Southeast Asian languages has centered on language understanding and knowledge. Typhoon introduced a family of Thai LLMs and the ThaiExam benchmark (Pipatanakul et al., 2023), and SEA-HELM provides a holistic suite across SEA languages (Susanto et al., 2025); related efforts include SeaLLMs (Nguyen et al., 2024) and the SEACrowd data hub (Lovenia et al., 2024). None of these target grounded agentic tool use, which for Thai remains largely unmeasured.

3 FahMai-TeleBench

Data. The knowledge base is the telephone directory of one fictional company, **FahMai Co.** (ฟ้าใหม่), with **1,995 employee records** across 19 fields: Thai and romanized names, nicknames, extensions, emails, titles, and an org hierarchy of unit $\subset$ section $\subset$ department (16 departments, five of them house brands) over 11 branches. The schema and the query types follow a real Thai enterprise-directory deployment. Each first name and nickname is drawn at random from public Thai name lists, skewed toward uncommon names, and each surname is compounded from common Thai surname parts. No record is copied from a real person. Every record and item is newly generated, so no model trained before the release can have seen them. Every item has a unique gold, whether a single row, a complete set, or an exact count: given names and nicknames are deliberately shared across employees, but each item's constraints jointly identify one answer (in two items, either of two employees). Some fields are deliberately blank, as in real directories where many employees record no nickname; these support the field-blank refusals. We fix one schema on purpose: holding the knowledge base constant is what makes the tool configuration the only varied factor. The released generator can produce organizations

[1]Code (MIT) and data (CC BY 4.0): https://github.com/yongsa-nut/FahMaiTeleBench.

with other schemas, sizes, and blank profiles.

Construction. Deterministic scripts generate the directory and the items. A seed-pinned generator fills an org chart, with headcounts per department and position level, with generated employees; it keeps 95–98% of surnames unique apart from planted family clusters, shares nicknames to create homonyms, and leaves fields blank at controlled rates. Thirteen scripts build the items: a random generator (261 items, 21 subtypes) and twelve targeted scripts covering one to four subtypes each. (i) A script selects rows that fit the subtype, e.g., an employee whose given name 12 others share but whose title is unique. (ii) It phrases the question: 388 items fill one of 159 Thai or English templates, 161 use a question written for the item, and 77 are individually edited versions of either. (iii) It reads the gold from the selected rows: a field value, the full member set, or a computed count. The checks are queries over the directory, not keyword matches on the question: they count the rows that meet an item's constraints, compare a superlative's top value with the runner-up, or search every name field for a not-found name. A separate checker then re-derived each item's constraints from its question text, evaluated them on the directory, and graded an ideal answer for every matching row (Appendix B, Table 5). Phrasing follows a style guide drawn from real usage logs: questions read as a user would type them and never gloss a directory code, a field name, or the answer's role. An LLM coding assistant wrote the scripts, including every template and written question, from our taxonomy and style guide; every gold is computed from directory rows, not stated by a model, and we reviewed the items manually.

Validation. Two native Thai raters, one of them an author, independently checked a stratified 10% sample (58 items, all 29 subtypes) against the directory rows that define each gold. Raw agreement on gold correctness, well-formedness, and naturalness is 91%, 97%, and 95% (Gwet's AC1 0.91, 0.96, and 0.95). Every instance of the issue types the raters found has been corrected across all 626 items (Appendix D).

Question types. The 626 items span eight groups and 29 subtypes (Appendix A): A identity lookup, B name resolution, C counting and aggregation, D disambiguation, E multi-hop, F org and brand knowledge, G bilingual and code-switched fields, and H refusal. Items are 38% English and 62% Thai; 521 require an answer and 105 a refusal. Several subtypes target the Thai phenomena above. Noisy-name resolution (B2, B3) requires normalizing a nickname or honorific-wrapped name; B3's hard tier (10 of its 20 items) gives only a shared given name and a company-unique role, forcing disambiguation among 9–13 homonyms. Multi-hop bridges (E1, E2, E4) compose chains of up to four hops through org-unit nesting and a secretary↔executive bridge, since the directory has no manager field. Listing and aggregation items (group C, multi-entity turns D3, and implicit-hierarchy listings E3) require a complete set rather than a single row, and premise correction (F2) requires rejecting a false premise. The refusal group (H) covers five reasons: field not in the table, person not found, subjective question, out-of-company entity, and blank field.

config	tools	probes
T1 grep	`grep_csv` (substring search over all cells) + `read_csv_rows`	retry after an empty result
T2 search	`search_employees` (14 optional filter fields, combined with AND)	structured-query formulation
T3 both	T1 + T2 tools	tool routing
T4 REPL	`python_repl` (pandas over the directory)	aggregation in code

Table 1: The four tool configurations. Every item runs under each; *probes* names the behavior each configuration isolates.

Tools and grading. We expose three tool types and run every item under four tool configurations (Table 1). Because only the toolset changes across configurations, differences in accuracy or behavior are attributable to tool affordance, not the item. Answer items pass if the response contains every required gold entity (in Thai or romanized form) and, for counts, the gold number as a standalone number. Refusal items pass if the response contains one of the accepted phrasings for their reason (a canonical Thai phrase and close Thai or English paraphrases) and leaks no extension or employee ID. Every verdict is a deterministic string operation, so grading is exactly reproducible from the released model outputs.

4 Experimental Setup

We evaluate **twelve model variants** (eleven models; GPT-5.5 at two reasoning efforts) in three tiers. The **closed-source** tier comprises GPT-5.4

tier	model	T1 grep	T2 search	T3 both	T4 REPL
Closed-source	GPT-5.4 (medium)	97.4 [96.2–98.6]	**97.4** [96.2–98.6]	**98.2** [97.3–99.2]	**98.9** [97.9–99.7]
	GPT-5.5 (medium)	96.8 [95.4–98.1]	73.0† [69.5–76.5]	95.7 [94.1–97.3]	97.6 [96.3–98.7]
	GPT-5.5 (low)	96.3 [94.9–97.8]	65.5† [61.5–69.0]	93.8 [91.9–95.7]	98.1 [97.0–99.0]
	Claude Sonnet 4.6	93.9 [92.0–95.7]	93.5 [91.5–95.2]	95.0 [93.1–96.6]	92.0 [89.8–94.2]
	Gemini-3-Flash	88.2 [85.6–90.7]	87.2 [84.7–89.8]	91.2 [89.0–93.3]	81.2 [78.1–84.2]
Open-weight	GLM-5.1	**97.6** [96.3–98.7]	96.8 [95.4–98.1]	97.8 [96.5–98.9]	96.0 [94.4–97.4]
	DeepSeek-V4-Pro	97.1 [95.7–98.4]	96.3 [94.7–97.8]	95.4 [93.6–97.0]	98.2 [97.1–99.2]
	DeepSeek-V4-Flash	96.6 [95.2–97.9]	96.0 [94.4–97.4]	97.4 [96.2–98.6]	95.2 [93.5–96.8]
	Gemma-4-31B	88.5 [85.9–91.1]	90.3 [87.9–92.5]	92.2 [89.9–94.2]	88.5 [86.1–91.1]
	MiniMax-M2.7	88.5 [85.8–90.9]	90.9 [88.7–93.1]	92.5 [90.4–94.6]	88.3 [85.8–90.7]
Thai open-weight	Typhoon-2.5 (30B)	57.8 [53.8–61.5]	69.5 [66.0–73.2]	74.3 [70.8–77.6]	60.5 [56.7–64.2]
	Typhoon-S-8B	37.4 [33.5–41.2]	31.2 [27.8–34.8]	33.7 [30.0–37.7]	24.8 [21.6–28.3]

Table 2: Accuracy (%) of the twelve model variants under the four tool configurations, with item-level bootstrap 95% CIs [lo–hi] (2,000 resamples) over all 626 items. Best per column in **bold**. †GPT-5.5 search-tool loop (Section 5).

(medium), GPT-5.5 (medium and low), Claude Sonnet 4.6 (non-thinking), and Gemini-3-Flash. The **open-weight** tier comprises GLM-5.1, DeepSeek-V4-Pro, DeepSeek-V4-Flash, Gemma-4-31B, and MiniMax-M2.7. The **Thai open-weight** tier comprises Typhoon-2.5 (30B) and Typhoon-S-8B, the latter from Thailand's national ThaiLLM initiative[2], post-trained on a Qwen3-8B-based backbone (Yang et al., 2025; Pipatanakul and Taveekitworachai, 2026) and served with a 16k-token context window. We call each model through its provider's tool-calling API, directly or through OpenRouter, with a shared system prompt; evaluation runs took place between May and September 2026, and the DeepSeek models were run on the April 2026 V4 weights, through DeepSeek's API and through fp8 hosts on OpenRouter. We set the temperature to 0 where the provider supports it (0.2 for the two Typhoon models); reasoning models use their default sampling. An agentic loop is capped at **seven tool-call rounds**, and round-exhaustion counts as a failure. We retry transient provider errors (HTTP 5xx or empty responses), so the failures we count are wrong answers, round-exhaustion, context-window overflow on the largest tool results, and malformed tool calls. Following SEA-HELM's reported results (Susanto et al., 2025), we run a single pass per cell; we report 95% percentile intervals from a non-parametric bootstrap over items (2,000 resamples), which capture item-sampling uncertainty but not run-to-run variation.

[2]The other 8B models released under the ThaiLLM initiative do not produce tool calls, so a tool-use benchmark cannot evaluate them.

5 Results

Near-solved at the frontier, discriminating below it. Seven of the twelve model variants (six distinct models) reach 95% or more on their best tool configuration (Table 2), which shows that the items are answerable as specified. Below this group the spread is wide and ordered: Gemini-3-Flash, Gemma-4-31B, and MiniMax-M2.7 reach 81–93% across configurations, Typhoon-2.5 58–74%, and Typhoon-S-8B 25–37%. The two Thai-specialized models trail all others by far more than the bootstrap half-widths (at most about four points); the other gaps are smaller. Open weights are not the bottleneck: GLM-5.1 and DeepSeek-V4-Flash reach 97.8% and 97.4% on their best configuration, on par with the closed frontier. Accuracy therefore separates models mainly in the smaller and Thai-specialized part of the deployable tier; at the frontier, the tool axis separates them.

Tool availability is the variable. Holding the item fixed and varying only the toolset moves scores substantially, and the direction differs by model. The clearest case is Sonnet 4.6 on noisy names (B3, $n = 20$): 90% with grep alone and 50% with structured search alone, because substring grep tolerates the noise and exact-field filters do not. Offered both, Sonnet 4.6 opens with search on 167 of the 169 items in the subtypes where grep is the better tool, and reaches 70% on B3. Retry behavior, read off the grep-only traces, also separates models. On the 90 name- and shorthand-resolution items (nickname variants B2, noisy names B3, enterprise shorthand B4, and brand and branch priors F1), GPT-5.4 issues a second, different query on 49 and solves 95% of B3, whereas Typhoon-2.5 never does (0 of 90) and solves no B3 item. None of this is

model	EN	TH	Δ
GPT-5.4 (medium)	98.8	97.4	−0.8
GPT-5.5 (medium)	91.1	90.1	+0.2
GPT-5.5 (low)	89.8	87.0	−1.6
Claude Sonnet 4.6	96.1	91.7	−4.3
Gemini-3-Flash	89.3	85.1	−4.7
GLM-5.1	97.8	96.4	−1.0
DeepSeek-V4-Pro	97.2	96.3	+0.1
DeepSeek-V4-Flash	96.9	95.8	−0.1
Gemma-4-31B	93.7	86.8	−7.3
MiniMax-M2.7	92.9	87.7	−4.4
Typhoon-2.5 (30B)	62.4	66.9	+3.0
Typhoon-S-8B	30.3	32.5	−1.5

Table 3: Accuracy (%) on English and Thai items, pooled over the four tool configurations. Δ is the mean over subtypes of (TH − EN), restricted to the 26 subtypes with at least five items in each language; negative values mean a Thai penalty. Code-switched items are excluded.

visible in a single-tool, outcome-only score.

A tool-schema sensitivity: GPT-5.5's search-tool loop. GPT-5.5 is frontier-class with grep (96.8/96.3%, medium/low effort) and the REPL (97.6/98.1%) but drops to 73.0% and 65.5% with the structured search tool. The mechanism is a degenerate loop: it fills the search tool's optional fields with hallucinated values, gets empty results, and repeats the call until the seven-round cap. Of GPT-5.5-medium's 169 search-tool failures, 152 (90%) are round-exhaustion, against two exhaustions in all its other configurations combined, and no other model exceeds 18 round-exhaustions across all configurations (Table 6). Offering grep alongside (T3) lets GPT-5.5 route around the search tool and largely recover (95.7%). Both efforts fail by looping on the same number of search items (152 each), but these sets overlap only partly (Jaccard similarity 0.25), which suggests that the loop is a property of the model–interface pairing rather than of particular items. The drop is language-independent (Table 3).

Language axis versus tool axis. Because the English and Thai halves are not parallel translations, we compare the two languages within subtypes (Table 3). The Thai penalty is model-specific rather than uniform. GPT-5.4, GPT-5.5, GLM-5.1, and both DeepSeek models show little or no Thai penalty (Δ between -1.6 and $+0.2$), while Gemma-4-31B, Gemini-3-Flash, Sonnet 4.6, and MiniMax-M2.7 pay about four to seven points. The Thai-specialized Typhoon-2.5 shows the reverse pattern and scores higher on Thai. Language can also shift behavior rather than accuracy: nearly all of Sonnet 4.6's failures on bilingual field lookups are Thai-phrased employee-ID requests that it refuses as personal data. The tool axis, in contrast, moves every model regardless of language, and it explains most of Typhoon-2.5's deficit. On the refusal group, which demands Thai comprehension and a phrasing contract but little tool work, Typhoon-2.5 scores 95.5%, against 59.5% on the tool-requiring remainder. An oracle-context control makes the comparison direct (Appendix E): with tools disabled and the gold rows placed among distractors in the prompt, Typhoon-2.5 answers 89.9% of 149 answer items, against 48.3–71.1% agentically on the same items, while Gemma-4-31B stays within its agentic range (91.3% against 86.6–91.3%). Typhoon-2.5 reads the Thai records well; operating the tools is what it cannot reliably do, and a text-only multilingual evaluation would miss this.

Error analysis. Every one of the 626 items is solved by at least one model variant, a necessary (though not sufficient) condition for gold validity, and only one item is failed by GPT-5.4 under all four configurations. Failure types differ by model (Appendix C): GPT-5.4's failures are all wrong answers, more than half of GPT-5.5's are round-exhaustions from the search loop, and Gemma-4-31B's are almost all wrong answers.

6 Discussion and Conclusion

An easy task examined carefully can be more informative than a hard task examined coarsely. The frontier is near its ceiling by design: on a simple but real task, the small local models one may actually deploy still struggle, and the reason is tool use rather than language. Two practical implications follow from the tool axis. First, the tool *interface* can matter as much as the choice of model: GPT-5.5 (medium) swings 24 points between grep and search, and Sonnet 4.6 loses 40 points on noisy names when grep is replaced by search, and 20 when it must choose between them. Second, offering both a structured and a raw-text tool is cheap insurance: for eight of the twelve variants, T3 is within one point of the best configuration, and it is never more than 4.3 points below it. The GPT-5.5 result further suggests that easy, well-posed benchmarks can serve as regression detectors. Finally, a model's multilingual ability measured on text does not predict its agentic behavior in that language, which argues for evaluating agents in the languages they will serve.

Limitations

We deliberately fix one schema and one domain, so that the tool configuration is the only factor that varies; results may not transfer to directories with other schemas (for example, with explicit manager fields) or to other domains. Gold validity rests on checks at construction, on an automated re-check of every item against the directory, on every item being solved by at least one model, and on a two-rater validation of a 58-item stratified sample; the remaining items were reviewed by the authors but not independently validated. The question phrasings were written by an LLM coding assistant; they follow a style guide drawn from real usage logs but may still differ from how real users phrase requests. Grading is string-based: golds list Thai and romanized forms and refusals accept a fixed set of paraphrases, so a correct answer in an unanticipated surface form is scored wrong. Each cell comes from a single run. The bootstrap intervals reflect which items were sampled, not how much a score would change if the same items were run again; that change may be larger for the reasoning models, which ran at their default sampling temperature rather than at 0. Open-weight models were queried through third-party providers whose serving precision we do not control, and Typhoon-S-8B's 16k-token context limits how much tool output it can read. The seven-round cap bounds the deepest items, and genuinely interactive refinement, such as clarification turns, is out of scope. Finally, difficulty is calibrated to current models and will likely erode as they improve, though the GPT-5.5 result suggests this erosion need not be monotonic.

Acknowledgments

We thank Pachaya Sailamul for independently rating the human-validation sample.

References

Chaithanya Bandi, Razvan-Gabriel Dumitru, Ben Hertzberg, Divyansh Agarwal, Geobio Boo, Tejas Polakam, Sami Hassaan, Jeff Da, HiJae Kim, Vipul Gupta, Manasi Sharma, Andrew Park, Martin Dimakis, Ernesto Gabriel Hernandez Montoya, Dan Rambado, Ivan Salazar, Rafael Cruz, Mohammad-Hossein Rezaei, Chetan Rane, and 4 others. 2026. MCP-Atlas: A large-scale benchmark for tool-use competency with real MCP servers. *arXiv preprint arXiv:2602.00933*.

Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. 2026. τ^2-Bench: Evaluating conversational agents in a dual-control environment. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*.

Jinyang Li, Binyuan Hui, Ge Qu, Jiaxi Yang, Binhua Li, Bowen Li, Bailin Wang, Bowen Qin, Ruiying Geng, Nan Huo, Xuanhe Zhou, Chenhao Ma, Guoliang Li, Kevin C. C. Chang, Fei Huang, Reynold Cheng, and Yongbin Li. 2023a. Can LLM already serve as A database interface? A BIg bench for large-scale database grounded text-to-SQLs. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023), Datasets and Benchmarks Track*, volume 36, pages 42330–42357.

Minghao Li, Yingxiu Zhao, Bowen Yu, Feifan Song, Hangyu Li, Haiyang Yu, Zhoujun Li, Fei Huang, and Yongbin Li. 2023b. API-Bank: A comprehensive benchmark for tool-augmented LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3102–3116, Singapore. Association for Computational Linguistics.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, and 3 others. 2024. AgentBench: Evaluating LLMs as agents. In *International Conference on Learning Representations (ICLR)*.

Holy Lovenia, Rahmad Mahendra, Salsabil Maulana Akbar, Lester James V. Miranda, Jennifer Santoso, Elyanah Aco, Akhdan Fadhilah, Jonibek Mansurov, Joseph Marvin Imperial, Onno P. Kampman, Joel Ruben Antony Moniz, Muhammad Ravi Shulthan Habibi, Frederikus Hudi, Railey Montalan, Ryan Ignatius, Joanito Agili Lopo, William Nixon, Börje F. Karlsson, James Jaya, and 42 others. 2024. SEACrowd: A multilingual multimodal data hub and benchmark suite for Southeast Asian languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5155–5203, Miami, Florida, USA. Association for Computational Linguistics.

Jiarui Lu, Thomas Holleis, Yizhe Zhang, Bernhard Aumayer, Feng Nan, Haoping Bai, Shuang Ma, Shen Ma, Mengyu Li, Guoli Yin, Zirui Wang, and Ruoming Pang. 2025. ToolSandbox: A stateful, conversational, interactive evaluation benchmark for LLM tool use capabilities. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1160–1183, Albuquerque, New Mexico. Association for Computational Linguistics.

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Zhiqiang Hu, Chenhui Shen, Yew Ken Chia, Xingxuan Li, Jianyu Wang, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, Hang Zhang, and Lidong Bing. 2024. SeaLLMs - large language models for Southeast Asia. In *Proceedings of the 62nd Annual Meeting of the*

Association for Computational Linguistics (Volume 3: System Demonstrations), pages 294–304, Bangkok, Thailand. Association for Computational Linguistics.

Shishir G. Patil, Huanzhi Mao, Fanjia Yan, Charlie Cheng-Jie Ji, Vishnu Suresh, Ion Stoica, and Joseph E. Gonzalez. 2025. The Berkeley function calling leaderboard (BFCL): From tool use to agentic evaluation of large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 48371–48392.

Shishir G. Patil, Tianjun Zhang, Xin Wang, and Joseph E. Gonzalez. 2024. Gorilla: Large language model connected with massive APIs. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, volume 37, pages 126544–126565.

Kunat Pipatanakul, Phatrasek Jirabovonvisut, Potsawee Manakul, Sittipong Sripaisarnmongkol, Ruangsak Patomwong, Pathomporn Chokchainant, and Kasima Tharnpipitchai. 2023. Typhoon: Thai large language models. *arXiv preprint arXiv:2312.13951*.

Kunat Pipatanakul and Pittawat Taveekitworachai. 2026. Typhoon-S: Minimal open post-training for sovereign large language models. *arXiv preprint arXiv:2601.18129*.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. ToolLLM: Facilitating large language models to master 16000+ real-world APIs. In *International Conference on Learning Representations (ICLR)*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Yangjun Ruan, Honghua Dong, Andrew Wang, Silviu Pitis, Yongchao Zhou, Jimmy Ba, Yann Dubois, Chris J. Maddison, and Tatsunori Hashimoto. 2024. Identifying the risks of LM agents with an LM-emulated sandbox. In *International Conference on Learning Representations (ICLR)*.

Yosephine Susanto, Adithya Venkatadri Hulagadri, Jann Railey Montalan, Jian Gang Ngui, Xianbin Yong, Wei Qi Leong, Hamsawardhini Rengarajan, Peerat Limkonchotiwat, Yifan Mai, and William Chandra Tjhi. 2025. SEA-HELM: Southeast Asian holistic evaluation of language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12308–12336, Vienna, Austria. Association for Computational Linguistics.

Harsh Trivedi, Tushar Khot, Mareike Hartmann, Ruskin Manku, Vinty Dong, Edward Li, Shashank Gupta, Ashish Sabharwal, and Niranjan Balasubramanian. 2024. AppWorld: A controllable world of apps and people for benchmarking interactive coding agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16022–16076, Bangkok, Thailand. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. 2025. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. In *The Thirteenth International Conference on Learning Representations (ICLR)*.

Bowen Ye, Rang Li, Qibin Yang, Yuanxin Liu, Linli Yao, Hanglong Lv, Zhihui Xie, Chenxin An, Lei Li, Lingpeng Kong, Qi Liu, Zhifang Sui, and Tong Yang. 2026. Claw-Eval: Towards trustworthy evaluation of autonomous agents. *arXiv preprint arXiv:2604.06132*.

Junjie Ye, Zhengyin Du, Xuesong Yao, Weijian Lin, Yufei Xu, Zehui Chen, Zaiyuan Wang, Sining Zhu, Zhiheng Xi, Siyu Yuan, Tao Gui, Qi Zhang, Xuanjing Huang, and Jiecao Chen. 2025. ToolHop: A query-driven benchmark for evaluating large language models in multi-hop tool use. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2995–3021, Vienna, Austria. Association for Computational Linguistics.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir Radev. 2018. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3911–3921, Brussels, Belgium. Association for Computational Linguistics.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. 2024. WebArena: A realistic web environment for building autonomous agents. In *International Conference on Learning Representations (ICLR)*.

A Full Subtype Taxonomy

Table 4 lists the 29 subtypes. "Gold" gives the grading shape: *answer* = required gold entities (Thai or romanized forms), *+neg* = also a forbidden entity, *count* = the gold number as a standalone number, *listing* = at least k correct entities, *refuse* = an accepted phrasing for the refusal reason.

B Item Construction

Table 5 gives, for each subtype, how many items come from the random generator and how many from a targeted script, how many distinct phrasing templates produced them, how each question was phrased, and what the checks compute on the directory.

Re-check against the directory. A separate checker tests whether each question determines its answer. For each item it re-derives the constraints stated in the final question text and evaluates them on the directory. In 427 items the constraint values are parsed with the template that generated the item and matched by exact field equality; in 193, one phrase is first mapped to a field value by a recorded annotation (for example, "VP of Logistics" to the position VICE PRESIDENT LOGISTICS, or a fixed list of the fruit and colour words among the 188 distinct nicknames). For every row that satisfies the constraints, the checker grades an ideal answer with the benchmark's grader. An item that fails is corrected: a question that more than one employee satisfies names the target's level or rank, or its gold accepts both readings (16 items); a count must equal the directory's (2 items); and a listing's gold credits every valid member and requires nothing the question does not ask for (46 items). Of the 626 items, 469 single-answer and refusal items match exactly one row or have a refusal reason that holds, and 148 count, set, or two-reading items have golds consistent with the directory. Three items identify a department's head by its unit code (the department code plus VP), since that department has three VPs, and six depend on facts outside the directory's fields: three reporting-line items, which we checked against the organization chart behind the directory, and three regional-geography items. The checker flagged all eleven items in a deliberately corrupted control set.

Worked example. One hard-tier item of subtype B3 asks "What's the extension for Thavan, the GM of Saifah?" Its build script makes it in the three steps of Section 3.

(i) Selection. The target is Thawan Boonnamphong, General Manager of the Saifah brand. Thirteen employees share the given name Thawan (ธาวร), but no one else holds this title, so the title alone identifies the target.

(ii) Question. The question spells the name "Thavan", a w→v romanization variant that matches no name field in the directory. A literal search for the user's string returns nothing, so the model must normalize the name, and only then can the title narrow the 13 candidates to one.

(iii) Gold. The gold is the extension on the target's row (79517). The extensions of the other twelve Thawans are forbidden, so a response that names the wrong Thawan fails even if it also contains 79517.

Checks. Before writing the item, the script confirms that "Thavan" has no hits in the four name fields, that exactly one employee holds the title, and that the gold differs from every forbidden value. If any check fails, the script aborts. Items from the random generator follow the same steps, except that step (i) samples rows at random from those that meet the subtype's condition and step (ii) draws one of several templates; for a reverse lookup (A3), for instance, it fills "ext {ext} belongs to?" or "ต่อ {ext} เบอร์ใคร" with a sampled extension, and the gold is that row's first and last name.

C Failure Types

We bucket each failure from its trace and sum over the four tool configurations (Table 6). *Answered-wrong* is a wrong final answer; *round-exhaust* means the agent hit the seven-round cap (mostly the GPT-5.5 search loop); *error* is a request that failed after retries, mostly a context-window overflow on a large tool result and in four cases malformed tool-call arguments. For each subtype, taking the mean over the twelve model variants of each variant's best configuration, the hardest subtypes are the multi-hop bridges (E1 76%, E4 83%), noisy-name resolution (B3 83%), brand unit heads (F3 85%), and premise correction (F2 85%). The refusal subtypes average 92–100%, but two of them discriminate sharply: Typhoon-S-8B scores 5% on subjective questions (H3) and 15% on out-of-company questions (H4).

D Human Validation of the Gold Answers

Sample. The 58-item sample stratifies 10% of each of the 29 subtypes and includes the single item that GPT-5.4 fails under every configuration. Both raters labeled the same items.

Protocol. Each rating page shows the question, the evidence that defines the gold, and the expected answer; it never shows model outputs or grader

verdicts. The evidence is the ground-truth directory row or rows, tailored to each refusal reason: the full schema for field-not-in-table items, near-miss rows and a whole-table scan for person-not-found items, and the person's row with the blank field highlighted for field-blank items. Superlative items show the computed ranking, and multi-hop items show the bridge person's row and the section roster. The rater answers three questions: does the gold match the evidence, is the question well-formed, and does it read naturally.

Agreement. Raw agreement is 91% (gold correctness), 97% (well-formedness), and 95% (naturalness), and Gwet's AC1 is 0.91, 0.96, and 0.95. We report AC1 because both raters answer "yes" on more than 90% of items, a prevalence regime in which Krippendorff's α is degenerate. Both raters confirmed the gold on 53 of the 58 items. Checking the other five against the directory found one wrong gold, which omitted the extension its question asked for; the other four golds were correct.

Full-set check. The raters' flags point to two issue types that can recur across items: a gold that does not accept the phone number its question asks for, and a question that glosses a directory code in parentheses. A check of all 626 items for both, and for golds whose token also appears in the question, ensures that every gold accepts the number its question asks for, no gold can be satisfied by a token from its own question, and no question glosses a code, a field name, or the answer's role; 73 items needed a correction. With the re-check in Appendix B, 14 of the 58 sampled items differ from the versions the raters labeled (11 in wording, 3 in gold); the agreement figures describe the labeled versions.

E Oracle-Context Control

The control asks whether a model that fails agentically can answer the same item when retrieval is done for it. We sample 149 answer items, stratified by group, excluding refusals, exact-count aggregation, superlatives, items with more than 12 gold rows, and the bridge subtypes E1 and E4, whose gold rows alone do not contain the chain. For each item, the prompt contains the question and a 12-row CSV excerpt: the gold rows plus random distractor rows, shuffled. Tools are disabled, and the grader is unchanged. Table 7 compares the result with the same model's agentic accuracy on the same items.

Typhoon-2.5 recovers most of the items it fails under every tool configuration, while Gemma-4-31B, whose agentic accuracy is already high, scores within its agentic range, so the oracle format is not simply easier. For Typhoon-2.5 the bottleneck is the tool loop, not reading Thai records.

grp	subtype	n	EN/TH	gold	example (abbrev.)
A	A1 executive by code or title	25	11/14	answer/+neg	*who is the RETVP* → (Wiriya / วิริยะ) ∧ (Chanchai / จันทชัย)
A	A2 name lookup	20	5/15	answer	*phone for Taksa-Orn Narawat* → (73987 / TAKSA-ORN.NA)
A	A3 reverse lookup	20	8/12	answer	*ext 71215 belongs to?* → (Tanet / ธเนศ) ∧ (Buathongprasert)
B	B1 casual name	25	4/21	answer	*Hook from SF, the number* → (73096 / YADTHIP.AN)
B	B2 nickname variant	30	10/20	answer	ขอเบอร์ต่อของคุณ *PANPAN* → (75477)
B	B3 noisy name form	20	11/9	answer/+neg	*email of Khun Kamala Chais…, our CFO* → (KAMALA.CH@FAHMAI.CO.TH)
B	B4 enterprise shorthand	20	8/12	answer/count	*staff at the Rama IX HQ branch* → count = 1,255
C	C1 unit listing	25	9/16	listing	*who's in CEO-SEC* → list ≥1
C	C2 member count	20	4/16	count	มีคนชื่อเล่นโอ๊ตกี่คน (how many are nicknamed Oat) → count = 6
C	C3 filtered count	20	10/10	count/listing	กี่คนในแผนก *B2B* ระดับ *IC* ที่เริ่ม… → count = 5
C	C4 surname family	24	11/13	count/listing	นามสกุล อภิกอบสุข มีกี่คน → count = 4
C	C5 superlative	10	5/5	answer	อายุงานยาวนานที่สุด (longest tenure) → (Kanok / กนก) ∧ (Khaengkadchai)
D	D1 nickname grid	25	7/18	listing	มิ้น คือใคร (who is Mint) → list ≥3
D	D2 code disambiguation	25	8/17	answer/+neg	*SFDR* ใครนะ ไม่ใช่ *SFVP* → (Saengdao) ∧ (Awutphat), ¬Wirat
D	D3 multi-entity turn	20	9/11	listing	*ext for CFO, CTO, COO* → list ≥3
E	E1 two-hop	40	18/22	answer/listing	เลขา *CFO* ชื่อเล่นอะไร (the CFO's secretary's nickname) → (Mint / มิ้น)
E	E2 bridge lookup	10	5/5	answer	*GM* แบรนด์สายฟ้า ใครนะ (GM of the Saifah brand) → (Thawan / ถาวร) ∧ (Boonnamphong)
E	E3 implicit hierarchy	25	11/14	answer/listing	ขอรายชื่อ *VP* ทั้งหมด (list all VPs) → list ≥10
E	E4 deep multi-hop	12	6/6	answer	*ext of the secretary of the VP of X's dept* → (78417)
F	F1 brand and branch priors	20	6/14	answer/listing	สาขาภาคใต้ (southern branches) → (HKT / Phuket) ∧ (HDY / Hat Yai)
F	F2 premise correction	25	12/13	answer	*I heard Kamala Ch. is the CEO, her ext?* → CFO (premise corrected) or Vachir (the CEO)
F	F3 brand unit head	20	9/11	answer	ใครเป็นหัวหน้าฝ่ายวิศวกรรมของแบรนด์ดาวเหนือ (head of engineering, Daonuea brand) → (Phonphai / พรไพร)
G	G1 cross-script field	20	10/10	answer	*What is the CEO's Thai nickname?* → (เบอร์รี่ / Berry)
G	G2 code-switch	20	0/20	answer	ขอ *email* ของ *Chief Executive Officer* → (VACHIR.CH@…)
H	H1 field not in table	25	12/13	refuse	เงินเดือน *CFO* (CFO's salary) → "ไม่สามารถให้ข้อมูลนี้ได้"
H	H2 person not found	25	6/19	refuse	ขอเบอร์ สมชายใจดี (made-up name) → "ไม่พบข้อมูล"
H	H3 subjective	20	11/9	refuse	ใครเก่งที่สุดในทีม *tech* (best in tech) → "ไม่สามารถให้ความเห็นได้"
H	H4 out of company	20	10/10	refuse	*CTO* ของ *Samsung* → "ไม่ใช่ข้อมูลของฟ้าใหม่"
H	H5 blank field	15	5/10	refuse	ชื่อเล่น *COO* (none on record) → "ไม่มีชื่อเล่นในระบบ"

Table 4: Full subtype taxonomy (eight groups, 29 subtypes). **Groups:** A (65) identity lookup; B (95) name resolution; C (99) counting and aggregation; D (70) disambiguation; E (87) multi-hop, bridge, and hierarchy; F (65) org and brand knowledge; G (40) bilingual and code-switched fields; H (105) refusal (five reasons).

subtype	n	gen./script	templates	T/W/R	what the checks compute
A1 executive by code or title	25	25/0	12	14/11/0	the code or title matches one employee (one item accepts either of two VPs)
A2 name lookup	20	20/0	6	20/0/0	the full name matches one employee
A3 reverse lookup	20	20/0	13	20/0/0	the extension, mobile, or email matches one employee
B1 casual name	25	25/0	13	23/0/2	nickname plus department, branch, name, or level matches one employee
B2 nickname variant	30	16/14	9	18/10/2	the nickname, without the variant's affix, matches its holders, or no one
B3 noisy name form	20	0/20	0	0/20/0	the spelling has no name-field hit; the title matches one employee
B4 enterprise shorthand	20	0/20	4	0/2/18	the expanded name gives one count or one head
C1 unit listing	25	25/0	13	25/0/0	the gold credits every member of the unit
C2 member count	20	20/0	10	20/0/0	the count equals the matching rows
C3 filtered count	20	0/20	4	20/0/0	the count or list equals the rows meeting all filters
C4 surname family	24	4/20	6	24/0/0	the count or list equals the holders of the surname
C5 superlative	10	0/10	2	0/3/7	the top value beats the runner-up
D1 nickname grid	25	10/15	7	25/0/0	the gold credits every holder of the name
D2 code disambiguation	25	12/13	9	25/0/0	the code matches one employee; the look-alike code's holder is rejected
D3 multi-entity turn	20	7/13	5	19/0/1	each code matches one employee
E1 two-hop	40	8/32	6	25/5/10	each hop resolves to one employee (one item asks for a three-person team)
E2 bridge lookup	10	6/4	5	9/1/0	the brand has one GM (one item also accepts the brand's VP)
E3 implicit hierarchy	25	9/16	4	6/9/10	one employee tops the level ladder, or the gold credits every member of the level; three reporting-line items follow the org chart
E4 deep multi-hop	12	0/12	4	7/0/5	each hop resolves to one employee; the largest or smallest department is unique; three items take the department-code VP as its head
F1 brand and branch priors	20	10/10	4	10/10/0	the brand has one GM; the branch's city or region; the nickname's category
F2 premise correction	25	0/25	2	25/0/0	the premise is false; the real role holder differs
F3 brand unit head	20	8/12	8	8/0/12	the brand's GM, VP, or section head is unique
G1 cross-script field	20	0/20	6	10/0/10	the title matches one employee
G2 code-switch	20	0/20	4	20/0/0	the title matches one employee
H1 field not in table	25	16/9	0	0/25/0	no column holds the requested field
H2 person not found	25	5/20	0	0/25/0	the name has no hit in any name field
H3 subjective	20	6/14	0	0/20/0	the question asks for an opinion (no directory check)
H4 out of company	20	5/15	0	0/20/0	the organization appears nowhere in the directory
H5 blank field	15	4/11	3	15/0/0	the employee's nickname field is blank
total	626	261/365	159	388/161/77	

Table 5: How each subtype is built. *gen./script*: items from the random generator and from targeted scripts. *templates*: distinct phrasing templates behind the subtype's items. *T/W/R*: questions that fill a template as generated, questions written for the item, and individually edited versions of either. The last column states what the checks compute on the directory rows to confirm the gold.

model	total	wrong	round-exh.	error
GPT-5.4 (medium)	50	50	0	0
GPT-5.5 (medium)	231	77	154	0
GPT-5.5 (low)	290	134	156	0
Claude Sonnet 4.6	160	156	4	0
Gemini-3-Flash	327	323	4	0
GLM-5.1	74	67	6	1
DeepSeek-V4-Pro	81	74	5	2
DeepSeek-V4-Flash	92	83	8	1
Gemma-4-31B	254	252	1	1
MiniMax-M2.7	249	226	18	5
Typhoon-2.5 (30B)	863	857	6	0
Typhoon-S-8B	1,709	1,683	13	13

Table 6: Failure types per model variant, summed over the four tool configurations.

model	oracle	agentic	recov.
Typhoon-2.5 (30B)	89.9	48.3–71.1	23/26
Gemma-4-31B	91.3	86.6–91.3	7/7

Table 7: Oracle-context control on 149 answer items. *Recovered* counts the items that the model fails under all four tool configurations but answers correctly in the oracle setting.