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

Anonymous ACL submission

Abstract

Multilingual agentic benchmarks remain rare, especially for low-resource languages such as Thai, where directory search differs from English in ways that change how an agent must use tools. We present **FahMai-TeleBench**, a bilingual (Thai/English) benchmark for grounded tool-use. A model answers employee-directory questions over a synthetic corporate database by issuing tool calls, or refuses when the answer is absent. All 626 items are fully synthetic and contamination-free by construction, yet the directory schema and the queries are inspired by a real Thai enterprise-directory deployment, and we grade them deterministically with no LLM judge. Our central design choice is to treat **tool availability as an experimental variable**: each item runs under four tool regimes (grep, structured search, both, and a Python REPL), which measures not just *whether* a model answers but *how* it uses tools. We find that the task is near-solved at the frontier (GPT-5.4 97–99%), yet sharply discriminating for the affordable, Thai-capable tier one would deploy (Typhoon-2.5 57–73%, Typhoon-S-8B 25–37%). The tool axis also surfaces a tool-schema sensitivity: GPT-5.5 drops to 74% on the wide-schema search tool while scoring 96–97% on grep and the REPL.

1 Introduction

Multilingual agentic benchmarks are rare, and they are rarest for low-resource languages. Most tool-use and function-calling benchmarks are English (Patil et al., 2024; Qin et al., 2024; Yao et al., 2024), so we know little about how agents behave when the language itself complicates the task. Thai is a useful stress test. It is written without word spaces, and people are addressed by nicknames that bear no systematic relation to their formal names. The same given name recurs across dozens of employees, romanizations are idiosyncratic, and honorific particles wrap names in ways that defeat naive matching. These properties change how a model must query a tool, not merely how it reads text. In this work, we study a concrete and widely deployed setting in which they all appear at once: an enterprise employee-directory assistant.

The task we study is deliberately simple. A user asks for a colleague's phone extension, email, or manager, and the model answers by calling a tool over a company directory, or refuses when the data does not contain the answer. We do not design the benchmark to defeat the strongest models. We instead explore the space of what existing models can do with combinations of common tools, and we find that the best current models essentially solve the task. This simplicity has two payoffs. First, because a capable frontier model can verify each item, we grade deterministically without a human or LLM judge. Second, an easy task that the frontier solves still spreads models apart below the frontier, which is exactly the tier one would deploy. A directory assistant runs on an affordable 8B–30B model, not a frontier reasoner. For instance, a small Thai-native model (Typhoon-S-8B) sits in the 25–37% range, while a mid-size general model clears 95%.

Our central design choice is to treat tool availability as an experimental variable. We run every item under four tool regimes: a raw-text grep, a structured search with many optional 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.

We make the following contributions. First, we release **FahMai-TeleBench**[1], a 626-item bilin-

[1]Code and data: https://github.com/BlindReviewYN/FahMaiTeleBench; released under CC BY 4.0 (data) and the MIT license (code). Evaluated models are used under their respective API terms or open-weight

gual benchmark for grounded directory tool-use, fully synthetic yet modeled on a real deployment's schema and queries, with deterministic grading and no LLM judge. 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), together with a five-way refusal taxonomy in which each refusal reason has its own canonical phrase. Third, we propose a tool-availability experimental design, with four regimes per item, that measures the tool-use *process* rather than only the outcome. Finally, we evaluate across the frontier and the deployable tier, and we locate the benchmark's discriminating power below the frontier.

2 Related Work

Tool-use benchmarks. Many benchmarks evaluate whether LLMs can invoke external tools correctly. Early work targets API and function calling: Gorilla and the Berkeley Function-Calling Leaderboard (Patil et al., 2024), ToolLLM over 16,000+ REST APIs (Qin et al., 2024), and API-Bank (Li et al., 2023b). More recent benchmarks add interaction and multi-hop structure. τ-bench and τ^2-bench emulate user–agent conversations under domain policies (Yao et al., 2024; Barres et al., 2025), ToolSandbox adds stateful conversational tool use (Lu et al., 2024), and ToolHop targets multi-hop tool chaining (Ye et al., 2025). Recent large-scale suites push breadth and trustworthiness further: MCP-Atlas evaluates tool use against 220 tools on real Model Context Protocol servers (Bandi et al., 2026), and Claw-Eval grades autonomous-agent trajectories across nine task categories (Ye et al., 2026). Broader agent environments embed models in richer settings, including AgentBench (Liu et al., 2024), WebArena (Zhou et al., 2024), and AppWorld (Trivedi et al., 2024). These benchmarks are almost all English, and they fix a single toolset. We instead make the toolset an experimental variable, and we target Thai.

Grounded querying and refusal. Text-to-SQL grades an executed query against a gold result, as in Spider (Yu et al., 2018) and BIRD (Li et al., 2023a), but it is not agentic. Knowing when *not* to answer has been studied since SQuAD 2.0 (Rajpurkar et al., 2018), and tool-use safety since ToolEmu (Ruan et al., 2024). We keep the exact, execution-style grading of this tradition 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 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., 2023) and the SEACrowd data hub (Lovenia et al., 2024). None of these target grounded agentic tool-use. Grounded tool-use for Thai therefore remains largely unmeasured, and this work addresses that gap.

3 FahMai-TeleBench

Design principles. FahMai-TeleBench rests on three commitments. First, *airtight grading without an LLM judge*: every gold answer is unique by construction and graded by substring or exact-count match, and every refusal is graded against a fixed canonical phrase. Second, *synthetic data, real task*: the records are fictional, but the schema and the query types are inspired by a real Thai enterprise-directory deployment. Third, *tool availability as a variable*, described below.

Data. The knowledge base is a single synthetic directory, **FahMai Co.**, a fictional Thai conglomerate of **1,995 generated rows** (Thai and romanized names, nicknames, extensions, emails, titles, and a four-level org hierarchy of unit $\subset$ section $\subset$ department plus brand subsidiaries). No record corresponds to a real person, and we verified zero overlap with common Thai corpora, VISTEC (Limkonchotiwat et al., 2021) and Wisesight (Suriyawongkul et al., 2019), so the benchmark is contamination-free by construction. The directory is engineered so every item is uniquely answerable: lookups resolve to exactly one row, nicknames are deliberately shared to create homonym items, and org units have single well-defined heads. A controlled number of fields are left blank to reflect reality, since many employees do not record a nickname, and to support field-present-but-empty refusals.

Question types. The 626 items span **8 groups / 29 subtypes**, 38% English and 62% Thai, with 521 answer items and 105 refusal items (full taxonomy in Appendix A). Several subtypes carry the benchmark's intent. Thai noisy-name resolution (B2,

licenses.

tool	definition
`grep_csv` `read_csv_rows`	Case-insensitive substring search over all CSV cells; read a contiguous row slice.
`search_employees`	Structured search, ∼15 optional AND-combined filters (name, nickname, extension, email, dept, section, position, level, id).
`python_repl`	Run Python against a persistent pandas `df` over the CSV.

Table 1: The three tool types.

config	tools	probes
T1 grep-only	`grep_csv` + `read_csv_rows`	retry after empty; raw-text robustness
T2 search	`search_employees`	structured-query formulation
T3 both	search + grep	tool routing
T4 repl	`python_repl`	aggregation; REPL over-use

Table 2: The four tool configurations.

B3) requires normalizing a nickname or honorific-wrapped name; the hard tier gives only a shared given name and a company-unique role, forcing disambiguation among 9–13 homonyms. The taxonomy spans both *depth* and *breadth*: multi-hop org bridges (E1, E2, E5) compose up to a 4-hop chain through org-unit nesting and a secretary↔executive bridge, since no explicit manager edge exists, while listing and aggregation items (C, D4, E3) require enumerating a complete set rather than a single row. Counterfactual premise-correction (F2) forces the model to reject a false premise. Finally, a **five-way refusal taxonomy** (H) covers field-not-in-table, person-not-found, subjective, out-of-company, and field-blank, each bound to its own canonical phrase.

Tools and the tool-availability axis. We expose three tool *types* (Table 1) and run every item under four configurations (Table 2). Because only the toolset changes across configs, differences in accuracy or behavior are attributable to tool affordance, not the item. This is the design's backbone and the source of the process-level findings in Section 5.

Grading. Answer items are graded by `must_contain_any_of` over a gold set with both Thai and romanized forms, or by exact count; refusal items against the canonical phrase for their reason, accepted in either language. Every verdict is a deterministic string operation, so scores are exactly reproducible.

4 Experimental Setup

We evaluate **twelve model configurations** (eleven models; GPT-5.5 at two efforts) across three tiers. The **closed-source** tier comprises GPT-5.4 (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; Typhoon-S-8B comes from Thailand's national ThaiLLM initiative[2] and is fine-tuned on the previous-generation Qwen3-8B backbone (Qwen Team, 2025; Pipatanakul et al., 2024), whose agentic capabilities are limited. We reach each model through its native tool-calling interface. We set the temperature to 0 where the provider supports it, and reasoning-native models use their default settings. We run an agentic loop capped at **seven tool-call rounds**, and round-exhaustion counts as a failure. Every model runs all four configurations. We retry transient provider errors (HTTP 5xx or empty responses), so the failures we count are genuine: wrong answers, round-exhaustion, or context-window overflow on the largest-result items. Because each model's per-item output is effectively deterministic, we run a single pass per cell and quantify uncertainty by a nonparametric bootstrap over items (2,000 resamples, 95% percentile intervals), following SEA-HELM (Susanto et al., 2025).

5 Results

Table 3 gives per-config accuracy over all 626 items with item-level bootstrap 95% CIs, for all twelve model configurations across all four configurations.

Near-solved at the frontier, sharply discriminating below. Seven models exceed 95% on their best config, and this frontier saturation certifies that the items are well-posed. Below it the spread is wide and orderly. A capable middle band reaches 78–92% (Gemini-3-Flash, Gemma-4, and MiniMax), and the Thai open-weight tier then drops steeply: Typhoon-2.5 reaches 57–73% and Typhoon-S-8B 25–37%. The bootstrap intervals are tight (half-widths under 4 points), so the roughly 20-point tier gaps are not sampling artifacts. The discriminating power lives entirely below the fron-

[2] Among the ThaiLLM-initiative 8B models, only the Typhoon line produces tool-use output; the others do not emit tool calls, so they cannot be evaluated on a tool-use task.

tier	model	T1 grep	T2 search	T3 both	T4 repl
Closed source	GPT-5.4 (medium)	97.1 [95.7–98.2]	**97.4** [96.0–98.6]	**98.2** [97.1–99.2]	**98.9** [97.9–99.7]
	GPT-5.5 (medium)	96.3 [94.9–97.8]	74.1† [70.8–77.6]	95.8 [94.1–97.3]	97.1 [95.7–98.4]
	GPT-5.5 (low)	96.2 [94.7–97.6]	67.1† [63.4–70.8]	93.8 [91.9–95.7]	97.0 [95.7–98.2]
	Claude Sonnet 4.6	94.2 [92.5–96.0]	94.6 [92.8–96.3]	95.5 [93.8–97.1]	91.7 [89.5–93.9]
	Gemini-3-Flash	89.5 [86.9–91.9]	87.5 [85.0–89.9]	90.9 [88.7–93.1]	78.1 [74.8–81.2]
Open weight	GLM-5.1	**98.2** [97.1–99.2]	95.7 [94.1–97.3]	97.8 [96.5–98.9]	95.5 [93.9–97.1]
	DeepSeek-V4-Pro	96.5 [94.9–97.8]	96.2 [94.6–97.6]	96.8 [95.4–98.1]	97.1 [95.5–98.4]
	DeepSeek-V4-Flash	95.2 [93.5–96.8]	95.8 [94.2–97.3]	96.8 [95.4–98.1]	95.5 [93.8–97.1]
	Gemma-4-31B	86.6 [83.9–89.1]	90.1 [87.7–92.3]	92.0 [89.9–94.1]	85.0 [82.3–87.9]
	MiniMax-M2.7	88.0 [85.3–90.6]	90.4 [88.0–92.7]	92.2 [90.1–94.2]	88.5 [85.9–90.9]
Thai open weight	Typhoon-2.5 (30B)	56.5 [52.4–60.5]	67.3 [63.6–70.8]	72.5 [69.0–76.0]	59.9 [56.2–63.4]
	Typhoon-S-8B	36.9 [33.2–40.7]	31.3 [28.0–35.0]	35.9 [32.1–39.9]	24.9 [21.9–28.3]

Table 3: Accuracy (%) by model and tool configuration, with item-level bootstrap 95% CIs [lo–hi] (2,000 resamples). Best per column in **bold**. † marks the GPT-5.5 search-tool failure (§5).

tier, where a directory assistant would actually be deployed.

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 datum is Sonnet on noisy-name resolution (B3): 90% with grep, 50% with search, and 70% with both. This is a 40-point swing on identical items, caused by an always-search default that picks the wrong tool when both are offered. Retry behavior, read off the grep-only traces, similarly separates models. GPT-5.4 re-queries after an empty grep result and recovers on B3 (90%), whereas Typhoon greps the full noisy string once and never retries (0%). None of this is 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 on grep (96.3/96.2%) and the REPL (97.1/97.0%) but **collapses on the structured search tool to 74.1% (medium) and 67.1% (low)**. The mechanism is a degenerate loop. It over-specifies the search tool's optional fields with hallucinated values, gets empty results, and repeats the call to the round cap. The failures are overwhelmingly round-exhaustion rather than wrong answers: 146 of GPT-5.5-medium's 163 search-tool failures (90%) are the loop hitting the seven-round cap, and the same model triggers the loop essentially never under grep or the REPL. GPT-5.4, Sonnet, GLM-5.1, and DeepSeek do not exhibit this on the identical tool, so the failure is specific to the (model, tool-interface) pair rather than a general capability regression. Offering grep alongside (T3) lets GPT-5.5 route around the search tool and largely recover (95.8%). Lowering the reasoning effort shifts *which* items loop, not how many, which confirms the loop is structural.

Error analysis. A per-item × model × config view (Appendix B) shows that **all 626 items are solved by at least one model**, a necessary check on the gold, and that GPT-5.4 fails just one item across all four configs. Failure *modes* are tiered (Table 5). GPT-5.4's failures are all wrong answers, with no crashes or exhaustions. GPT-5.5's are dominated by round-exhaustion, the search loop above. Gemma-4's are almost all wrong answers. The deployable tier is not the accuracy bottleneck: capable open-weight models such as DeepSeek-V4-Flash and GLM-5.1 match the frontier (95–98% on their best config), so a strong open model already suffices.

6 Discussion and Conclusion

Our results suggest that an easy task examined carefully can be more informative than a hard task examined coarsely. The tool-availability design turns a saturated accuracy benchmark into a behavioral one, and the deployable-tier focus turns a "solved" result into a deployment-relevant one. Two practical riders follow from the tool axis. First, the tool *interface* can matter more than the choice of model: GPT-5.5 swings 22 points on which tool it is given, and an always-search policy costs Sonnet 40 points on noisy names. Second, giving a capable model both a structured and a raw-text tool is a cheap insurance policy. The GPT-5.5 result further suggests that easy, unambiguous benchmarks may serve as regression detectors, since an unexpected failure on an easy task is a relatively unambiguous signal. We release the synthetic directory, all 626 items, the four-config harness, and the rule-based grader.

Limitations

The directory is a single synthetic organization, so generalization to other schemas remains untested. The grader relies on substring matching, which requires care with short or lone-digit golds; we handle these with exact-count items. Gold validity rests on construction-time checks plus the observation that every item is solvable by some model. We have not yet run independent human verification, such as inter-annotator agreement on a sample. We model tool-use as a single-turn conversation with a seven-round cap, so genuinely interactive refinement 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.

References

Chaithanya Bandi, Razvan-Gabriel Dumitru, Ben Hertzberg, Divyansh Agarwal, Geobio Boo, Tejas Polakam, and 1 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. 2025. τ^2-Bench: Evaluating conversational agents in a dual-control environment. *arXiv preprint arXiv:2506.07982*.

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 (NeurIPS)*.

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 (EMNLP)*.

Peerat Limkonchotiwat, Wannaphong Phatthiyaphaibun, Raheem Sarwar, Ekapol Chuangsuwanich, and Sarana Nutanong. 2021. Handling cross- and out-of-domain samples in Thai word segmentation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1003–1016.

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, and 1 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 (EMNLP)*.

Jiarui Lu, Thomas Holleis, Yizhe Zhang, Bernhard Aumayer, Feng Nan, Felix Bai, Shuang Ma, Shen Ma, Mengyu Li, Guoli Yin, Zirui Wang, and Ruoming Pang. 2024. ToolSandbox: A stateful, conversational, interactive evaluation benchmark for LLM tool use capabilities. *arXiv preprint arXiv:2408.04682*.

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. 2023. SeaLLMs – large language models for southeast asia. *arXiv preprint arXiv:2312.00738*.

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 (NeurIPS)*.

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, Potsawee Manakul, Natapong Nitarach, Sittipong Sripaisarnmongkol, Pathomporn Chokchainant, Kasima Tharnpipitchai, and 1 others. 2024. Typhoon 2: A family of open text and multimodal Thai large language models. *arXiv preprint arXiv:2412.13702*.

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)*.

Qwen Team. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

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 (ACL)*.

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)*.

Arthit Suriyawongkul, Ekapol Chuangsuwanich, Pattarawat Chormai, and Charin Polpanumas. 2019. Wisesight sentiment corpus. Zenodo.

Yosephine Susanto, Adithya Venkatadri Hulagadri, Jann Railey Montalan, Jian Gang Ngui, Wei Qi Leong, Hamsawardhini Rengarajan, Peerat Limkonchotiwat, William Chandra Tjhi, and Ai Ti Aw. 2025. SEA-HELM: Southeast Asian holistic evaluation of language models. *arXiv preprint arXiv:2502.14301*.

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 (ACL)*.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. 2024. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*.

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 (ACL)*.

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 (EMNLP)*.

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

The 626 items span 8 groups / 29 subtypes. "Gold" gives the grading shape: *answer* = `must_contain_any_of` (Thai and romanized forms), +*neg* = also a negative constraint, *count* = exact integer, *listing* = $\geq k$ entities, *refuse* = canonical refusal phrase.

B Cross-model failure-mode signature

Bucketed from traces, summed over all four configs. *answered-wrong* = wrong final answer; *round-exhaust* = hit the seven-round cap (the GPT-5.5 search loop). All 626 items are solved by at least one model, and GPT-5.4 fails exactly one item across all four configs. Difficulty concentrates on multi-hop and noisy Thai names: by best-of-config mean accuracy across the twelve model configurations, the hardest cells are multi-hop bridges (E1 83%, E5 90%), noisy or shorthand name resolution (B3 87%, B1 90%), and counterfactual premise-correction (F2 88%); the refusal group stays at the high end (H1–H4 92–100%).

grp	subtype	n	EN/TH	gold	example (abbrev.)
A	A1 ceo/president	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 email/identity	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 hard nickname variant	30	10/20	answer	นัตตี้คือใครนะ → (นัต)
B	B3 noisy name form	20	11/9	answer/+neg	*email of Khun Kamala Chais…* → (KAMALA.CH@FAHMAI.CO.TH)
B	B5 enterprise shorthand	20	8/12	answer/count	*staff at the Rama IX (R9) HQ* → count = 1255
C	C1 dept listing	25	9/16	listing	*who's in CEO-SEC* → list ≥1
C	C3 dept member count	20	4/16	count	มีคนชื่อเล่นโอ๊ตกี่คน → count = 6
C	C4 filtered count	20	10/10	count/listing	กี่คนในแผนก *B2B* ระดับ *IC* ที่เริ่ม… → count = 5
C	C5 surname family	24	11/13	count/listing	นามสกุล อภิกอบสุข มีกี่คน → count = 4
C	C6 superlative	10	5/5	answer	อายุงานยาวนานที่สุด → (Kanok / กนก) ∧ (Khaengkadchai)
D	D1 nickname grid	25	7/18	listing	มิ้น คือใคร → list ≥3
D	D2 EVP-vs-VP disambig	25	8/17	answer/+neg	*SFDR* ใครนะ ไม่ใช่ *SFVP* → (Saengdao) ∧ (Awutphat), ¬Wirat
D	D4 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 → (Mint / มิ้น)
E	E2 bridge lookup	10	5/5	answer	*GM* แบรนด์สายฟ้า ใครนะ → (Thawan / ถาวร) ∧ (Boonnamphong)
E	E3 implicit hierarchy	25	11/14	answer/listing	ขอรายชื่อ *VP* ทั้งหมด → list ≥10
E	E5 deep multi-hop	12	6/6	answer	*ext of the secretary of the VP of X's dept* → (78417)
F	F1 brand prior	20	6/14	answer/listing	สาขาภาคใต้ → (HKT / Phuket) ∧ (HDY / Hat Yai)
F	F2 counterfactual	25	12/13	answer	*I heard Kamala is the CTO…* → corrected to CFO (Rittichai…)
F	F3 subsidiary GM	20	9/11	answer	ใครเป็น *GM* สายฟ้า → (Thawan / ถาวร) ∧ (Boonnamphong)
G	G1 bilingual	20	10/10	answer	ขออีเมลของ *CEO* → (VACHIR.CH@FAHMAI.CO.TH)
G	G3 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* → "ไม่สามารถให้ข้อมูลนี้ได้"
H	H2 person-not-found	25	6/19	refuse	เบอร์ สมชายใจดี (made-up) → "ไม่พบข้อมูล"
H	H3 subjective	20	11/9	refuse	ใครเก่งที่สุดในทีม *tech* → "ไม่สามารถให้ความเห็นได้"
H	H4 out-of-company	20	10/10	refuse	*CTO* ของ *Samsung* → "ไม่ใช่ข้อมูลของฟ้าใหม่"
H	H7 field-blank	15	5/10	refuse	ชื่อเล่น *COO* (none on record) → "ไม่มีชื่อเล่นในระบบ"

Table 4: Full subtype taxonomy (8 groups, 29 subtypes). **Groups:** A (65) direct identity; B (95) nickname & noisy-name resolution; C (99) counting & aggregation; D (70) disambiguation; E (87) multi-hop, bridge & hierarchy; F (65) org & brand knowledge; G (40) bilingual / code-switch; H (105) refusal & safety (five reasons, each with its own canonical phrase).

tier	model	total fails	answered-wrong	round-exhaust
Closed source	GPT-5.4 (med)	52	52	0
	GPT-5.5 (med)	229	82	147
	GPT-5.5 (low)	288	137	151
	Claude Sonnet 4.6	150	149	1
	Gemini-3-Flash	338	333	5
Open weight	GLM-5.1	80	72	7
	DeepSeek-V4-Pro	84	78	6
	DeepSeek-V4-Flash	104	91	13
	Gemma-4-31B	290	289	1
	MiniMax-M2.7	256	234	16
Thai open weight	Typhoon-2.5 (30B)	900	895	5
	Typhoon-S-8B	1,696	1,674	15

Table 5: Per-model failure-mode signature, summed over all four configs.