ARE SMALL LANGUAGE MODELS READY FOR ON-DEVICE AGENTS? BENCHMARKING SMALL LANGUAGE MODELS AS SMART-HOME AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Small language models are attractive for local smart-home assistants, but their ability to complete state-constrained tasks remains unclear. We present EDGEBENCH-HOME, a bilingual benchmark with 840 cases across seven task families: device grounding and control, multi-device selection, ambiguity resolution, automation, dialogue-context resolution, personalized memory, and environment queries. We evaluate a primary pool of 22 models—18 local models in the $\leq$1B, 1–3B, and 3–7B ranges and four API-accessed large language models—under State-Given and State-Seeking interfaces. An additional Qwen3.5-4B task-family result is available for Chinese only. We also test three local checkpoints on physical RK3588 devices using the same prompt-only cases and scoring protocol. Success is determined by direct checks of actions, answers, resulting state, prohibited effects, and budget compliance; only ambiguity-resolution judgments require an LLM judge. Across the observed local size bands, mean family-level success rises from 0.6% to 9.0% to 22.3% under State-Given and from 0.0% to 1.9% to 8.8% under State-Seeking. These descriptive results show higher capability in the larger observed ranges, while reliable performance under partial observation remains limited.

1 INTRODUCTION

On-device agents face a coupled capability and resource problem: they must turn underspecified language into correct environment changes while operating under constraints on context, tool calls, memory, and compute. Prior work on mobile and on-device language models has highlighted this capability–efficiency trade-off, but generally evaluates model-level performance or generic tasks rather than complete household outcomes (Laskaridis et al., 2024; Murthy et al., 2024; Abdin et al., 2024). In a smart home, the same request may require resolving an alias, inspecting device state, selecting a set of devices, preserving unrelated state, maintaining a scoped preference, answering without actuation, or recovering from an unavailable tool. Textual command accuracy alone does not test these requirements because a plausible response may conceal an invalid call or an incorrect final state.

We introduce EDGEBENCH-HOME, a bilingual benchmark for evaluating language models as smart-home agents. Its seven task families cover the interactions illustrated in Figure 1, from device control and environment queries to dialogue-context resolution and personalized memory use. Each episode combines a user request, an environment state, and typed tools; the evaluator checks task completion and constraint satisfaction from the resulting interaction. The primary bilingual comparison covers 22 models: 18 local models across three parameter ranges ($\leq$1B, 1–3B, and 3–7B), including three models deployed on RK3588, and four larger language models accessed through APIs. Qwen3.5-4B contributes an additional Chinese-only task-family result and is reported separately.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

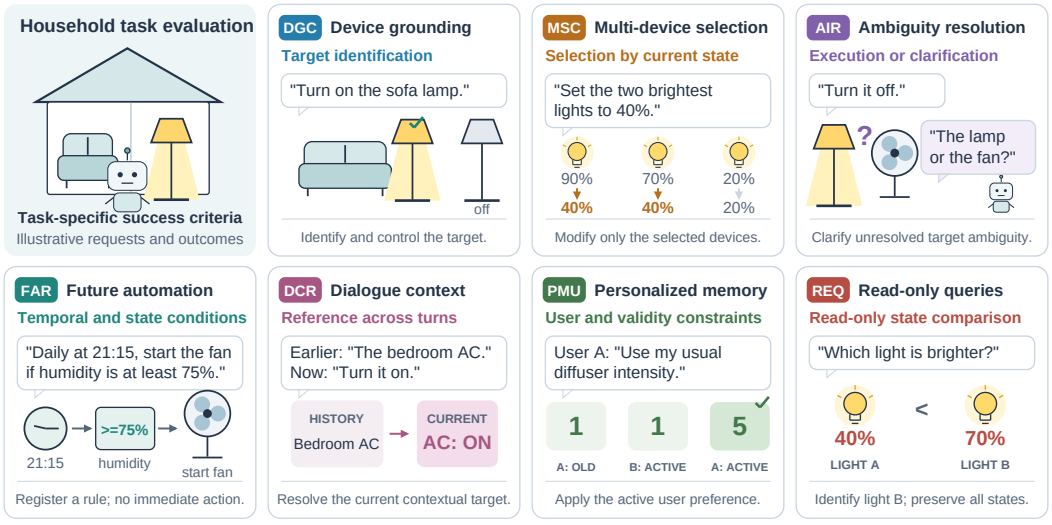

Figure 1: **Task taxonomy of the evaluated suite.** Schematic requests and required outcomes illustrate device grounding and control (DGC), multi-device selection and control (MSC), ambiguous intent resolution (AIR), future automation registration (FAR), dialogue-context resolution (DCR), personalized memory use (PMU), and read-only environment queries (REQ). DCR resolves references within the supplied dialogue; PMU selects a preference by user identity and validity. Examples illustrate task requirements, not observed model responses. Each language contains 60 cases per family; State-Given and State-Seeking are derived interfaces over the same language-specific cases.

Our contributions are:

1. **A focused benchmark and task taxonomy.** EDGEBENCH-HOME provides 840 bilingual smart-home cases organized into seven task families, covering device grounding and control, multi-device selection, ambiguity resolution, automation, dialogue context, personalized memory, and read-only environment queries.
2. **A strict outcome evaluation protocol.** The prompt-only State-Given and State-Seeking interfaces use the same cases under different information and interaction conditions. Evaluation checks actions, answers, resulting state, preservation of unrelated state, prohibited effects, and budget compliance, while using semantic assessment only for clarification quality.
3. **A bilingual comparison of local models.** We evaluate 22 models across three local size ranges and an API reference group, report language and task-family results, and isolate the Chinese-only Qwen3.5-4B task-family result from the complete bilingual aggregates.
4. **Physical-platform capability evidence.** We repeat the prompt-only evaluation for three local checkpoints on RK3588. These measurements characterize task capability for the tested board and models rather than general deployment readiness.

2 RELATED WORK

AgentBench (Liu et al., 2024), WebArena (Zhou et al., 2024), OSWorld (Xie et al., 2024), and SWE-bench (Jimenez et al., 2024) established a common lesson for interactive evaluation: agents should be judged by environment outcomes rather than by the surface form of their answers. ToolSandbox (Lu et al., 2024) and τ-bench (Yao et al., 2024) extend this view to stateful tools, intermediate dependencies, policies, user interaction, and repeated-trial reliability, while AgentDojo (Debenedetti et al., 2024) adds untrusted tool outputs and prompt-injection stress tests. Their limitation is that the environments and success conditions are largely generic: they do not model the combination of device availability, read-only inspection, identity-scoped memory, temporary preferences, and forbidden actuation that makes home control safety-critical. EDGEBENCH-HOME adapts the outcome-oriented

principle to these home-specific state transitions and treats generic function-call accuracy as only a diagnostic.

LLM smart-home assistants have also been studied in deployed or user-facing settings. King et al. (King et al., 2023b) use an LLM to interpret vague requests and control real smart-home devices, while Sasha (King et al., 2023a) explores goal-oriented requests such as making a room more comfortable and examines the resulting plans through a user study. SAGE (**?**) uses discrete actions, state checks, and a benchmark of challenging smart-home tasks. More recent work considers local execution: Birkmose et al. (Birkmose et al., 2025) study CPU-only LLMs for Home Assistant, combining intent and slot detection with response generation, and AdaHome (Lim et al., 2026) investigates adaptive assistance with local small language models, routing, and preference updates. These systems demonstrate the promise of natural-language home control, but each evaluates a narrower assistant pipeline or scenario. They do not provide a common suite for comparing task families, information conditions, strict state constraints, model size, and physical-device behavior. Our evaluation makes these dimensions explicit.

Smart-home and embodied benchmarks cover complementary parts of this problem. VirtualHome (Puig et al., 2018) and BEHAVIOR-1K (Li et al., 2024) provide executable household activities and rich simulation, while CASAS (Cook et al., 2013) provides longitudinal sensor streams without a typed action loop. HomeBench (Li et al., 2025) emphasizes valid and invalid single- and multi-device instructions; SimuHome (Seo et al., 2025) adds temporal and environmental dynamics; SMH-Bench (Li et al., 2026) evaluates smart-home reasoning from environment state; and PersonalHomeBench (Bharadwaj et al., 2026) focuses on personalized tasks and information gathering under partial observability. Their limitation is that these benchmarks do not provide the same focused protocol for small-model evaluation that jointly tests fuzzy device-set selection, evolving context, explicit non-mutation, memory selection, and physical-device capability under matched information conditions. EDGEBENCH-HOME targets this intersection while recognizing that the individual capabilities are grounded in prior work.

On-device model studies such as MELTing point (Laskaridis et al., 2024), MobileAIBench (Murthy et al., 2024), and the Phi-3 technical report (Abdin et al., 2024) characterize small-model behavior under local resource constraints. Model reports for Qwen2.5 (Team, 2024), Gemma (Team et al., 2024), Mistral (Jiang et al., 2023), and SmolLM (Allal et al., 2024) illustrate the available small-model design space. These studies provide useful model and deployment context, but they do not evaluate the strict environment outcomes considered here, including tool arguments, memory selection, recovery, and preservation of unrelated state. EDGEBENCH-HOME focuses on those task-level outcomes and reports platform conditions only to identify the tested comparison setting.

3 BENCHMARK AND EVALUATION PROTOCOL

3.1 TASK FORMULATION AND INTERFACES

An episode is $e = (s_0, o_0, u_{1:T}, \mathcal{T}, b, g)$: initial state, initial observation, user turns, typed tool contract, resource budget, and success predicate. At turn t, the agent emits a response or call $a_t = (n_t, x_t)$. The transition function returns (s_{t+1}, o_{t+1}, r_t), where r_t is a structured tool result. The model never receives hidden state, gold traces, expected final state, or evaluator fields.

The hidden state contains device attributes, room membership, availability, preferences, identities, and automation rules. The observation exposes only the information permitted by the interface and the results of previous tool calls. Budgets include context length, output length, interaction steps, and termination. We use a 16,384-token context limit, a 1,024-token maximum output for State-Given, a 512-token maximum output per State-Seeking turn, and at most 20 State-Seeking steps. Overflow, timeout, parse failure, and budget exhaustion are recorded as separate failure conditions.

Interaction interfaces. The evaluation uses State-Given (SG) and State-Seeking (SS) interfaces. SG provides the relevant state and device specifications in one prompt and requires the model to produce the complete response for the case. SS starts from a partial observation and allows the model to inspect the environment through permitted tools before producing its final response. Both interfaces use the same case set within each language, but differ in information visibility and interaction. Their

score difference is therefore descriptive; it should not be interpreted as an isolated causal estimate of tool use.

The unified scorer replays structured actions and checks task-specific outcomes. AIR clarification quality is assessed separately by a versioned LLM judge with labels `full`, `partial`, and `fail`; only `full` passes the strict metric, while `partial` is retained as a lenient diagnostic. The judge uses a frozen prompt, temperature 0, and response-hash binding, but its service identifier is a rolling alias and no completed human-agreement results are available for this campaign. All other task outcomes are determined by direct checks against the task requirements, the action trace, and the resulting state.

3.2 Task and dataset design

Task taxonomy. Figure 1 uses the task abbreviations below. The benchmark defines seven task families with shared evaluation contracts and task-specific success predicates.

DGC (Device Grounding and Control). Resolve explicit, contextual, colloquial, alias, noisy, self-corrected, and same-name device references, then execute one primary control intent with valid services and parameters while preserving unrelated state.

MSC (Multi-device Selection and Control). Select and control multiple targets through explicit or batch instructions, room-dependent and state-dependent rules, and top-k ranking, while preserving non-target state.

AIR (Ambiguity and Intent Resolution). Distinguish executable scene-level requests from under-specified slot or state ambiguity. Execute when the available evidence uniquely determines the action; otherwise request the missing information without acting.

FAR (Future Automation Registration). Create time-specific, state-driven, or combined time-and-state rules. Register the requested future behavior while leaving the current device state unchanged; the current release evaluates rule registration rather than a separate CRUD or conflict-detection suite.

DCR (Dialogue-context Resolution). Use the supplied dialogue history to resolve coreference, context continuation, topic shifts, and user corrections. Execute only the current request and do not replay completed historical actions.

PMU (Personalized Memory Use). Select relevant long- or short-term memory for the current user and task, account for validity and superseded preferences, and execute the corresponding device action without changing unrelated state.

REQ (Read-only Environment Queries). Answer structured questions about device status, counts, locations, capabilities, comparisons, extrema, abnormal device conditions, and filtered queries with an explicit no-mutation constraint. Specification lookup is permitted, but household state must remain unchanged.

Coverage and boundaries. We construct EdgeBench-Home as a bilingual benchmark. The formal evaluation reported here comprises 420 unique task cases per language, organized into seven families with 60 cases per family. Each case provides a model-visible input together with evaluator-only annotations containing ground truth and scoring metadata; these annotations are not additional cases. We evaluate each language-specific case set under two interfaces, State-Given and State-Seeking, reusing the same 420 cases for both interfaces. The two views therefore share case IDs and initial states, but they differ in visible information and permitted interaction.

Construction and quality control. The task definitions and dataset instances were designed and validated by domain experts, with LLM assistance used for natural-language realization and candidate construction. The released benchmark contains 420 Chinese and 420 English cases, and the formal model campaign evaluates a 420-case track for each language. Chinese and English are released as language-specific case sets with matched family counts; the paper does not treat them as independent random samples from all possible user language. Each episode specifies a structured task, an initial state, a user request, and a success predicate. Dataset validation is performed before model evaluation: deterministic checks verify the task instance, its expected state transition or answer, preservation requirements, and task-specific negative constraints. Audit metadata records the dataset revision, provenance, validation outcome, and corrections. These construction checks are separate from the model traces scored in the experiments.

3.3 SCORING AND REPORTING

Episode outcome. Each episode receives a binary strict outcome after protocol checks and task verification. A control case passes only when the requested goal is satisfied, all required actions and arguments are valid, unrelated state is preserved, no prohibited effect occurs, and the trace stays within the stated budget. Automation cases additionally require the requested rule to be registered without performing the future action immediately. Query cases require a correct answer and no state mutation. Clarification cases require an appropriate clarification response; a rollback cannot erase a prohibited action. Semantically equivalent safe traces are accepted.

The scorer also stores component-level diagnostics and failure labels, such as mode errors, parse failures, invalid actions, incorrect final state, incorrect answers, clarification failures, and incomplete trajectories. These diagnostics support task-specific analysis, but the primary reported metric is strict micro success rather than an average of component scores.

Aggregation. For a set E of scored cases, strict micro success is $|E|^{-1} \sum_{e \in E} \mathbb{1}[e \text{ passes}]$. We report this 420-case overall rate as the primary comparison. The family mean is the unweighted mean of the seven task-family success rates and is used to display task-specific patterns. It is not assumed to equal the overall rate because the workbook preserves the two aggregations as separate source fields. Every language/interface cell contains 420 cases with 60 cases per family. The same cases are used for SG and SS within a language, and each evaluation starts from the case's specified initial state.

3.4 MODELS AND COMPARISON PROTOCOL

The main paper reports the frozen prompt-only campaign. For each run, we retain the exact model and tokenizer revisions, platform, decoding parameters, protocol version, input hashes, per-episode output and trace, execution results, final state, and failure labels. All models use temperature 0 under the common context and output budgets above, and no evaluation case is used for model-specific tuning. The primary comparison contains 22 models with overall results in both languages: 18 local models in three nominal parameter ranges and four API-accessed large language models in a separate reference group. Qwen3.5-4B has Chinese task-family rows only and is reported as a Chinese-only supplementary result. The size-band and interface comparisons are descriptive rather than causal.

The measured local pool spans nominal sizes from 0.5B to 7B, including MiniCPM, Qwen3, Qwen3.5, and LFM2.5 checkpoints. The reported size labels are comparison labels rather than claims about exact total tensor counts. Three local checkpoints are measured on RK3588, while the remaining primary local checkpoints are evaluated on the server platform. The four API-accessed large language models are retained as a separate reference group and do not enter local size-band aggregates. Qwen3.5-4B is not part of the 22-model bilingual pool because its English task-family rows are unavailable.

Physical edge setting. We select the Rockchip RK3588 as a concrete physical platform because its compact system-on-chip combines general-purpose CPU resources with an integrated GPU and NPU (**?**). We use it as a reproducible deployment setting for task capability, not as a proxy for all edge hardware. We deploy three local checkpoints on RK3588—MiniCPM4-0.5B-16K, Qwen3.5-0.8B-16K, and Qwen3-1.7B-16K—covering the smallest two size ranges. These models use the same bilingual cases, prompt-only protocol, and SG/SS interfaces as the server evaluation. The platform comparison is therefore limited to the tested models and does not estimate a general hardware effect.

4 EXPERIMENTAL RESULTS

4.1 FORMAL TASK EVALUATION

The primary bilingual evaluation covers 22 models: three RK3588 deployments, fifteen server-hosted open models, and four large language models accessed through APIs. Each overall cell is a strict micro success rate over 420 records in the corresponding language track, under State-Given (SG) or State-Seeking (SS). The task-family tables contain the same 22 models plus a separate Chinese-only Qwen3.5-4B row. The latter is excluded from all bilingual and size-band summaries.

State-Given presents the relevant environment context in a single generation, whereas State-Seeking starts from partial observation and permits tool-mediated interaction. Overall rates are used for the

main comparison, and family means are used for task-specific analysis. The two values answer different questions and are therefore reported separately. AIR clarification labels use the versioned LLM judge described in the protocol; human-agreement results are not yet available.

Table 1: Strict task-family success (%) by model under State-Given (SG). F is the unweighted mean across the seven task families. The three local size bands and the API group contain the 22 models with complete bilingual overall and task-family coverage. Scores are averaged over Chinese and English; bold and underlined values mark the best and second-best local models within each size range.

Model	Yr	Size	Plat.	State-Given (SG)							
				F	DGC (%)	MSC (%)	AIR (%)	FAR (%)	DCR (%)	PMU (%)	REQ (%)
≤1B											
MiniCPM4-0.5B-16K	2025	0.5B	RK3588	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Qwen2.5-0.5B	2024	0.5B	Server	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Qwen3-0.6B	2025	0.6B	Server	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Qwen3.5-0.8B-16K	2026	0.8B	RK3588	1.9	8.3	0.0	0.8	0.0	**0.8**	3.3	0.0
Qwen3.5-0.8B	2026	0.8B	Server	**2.4**	**9.2**	**0.8**	**1.7**	0.0	**0.8**	**4.2**	0.0
MiniCPM-1B-SFT	2024	1B	Server	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MiniCPM-S-1B-SFT	2024	1B	Server	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1–3B											
LFM2.5-1.2B-Instruct	2026	1.2B	Server	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
DeepSeek-R1-Distill-Qwen-1.5B	2025	1.5B	Server	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Qwen2.5-1.5B-Instruct	2024	1.5B	Server	3.9	10.0	0.8	0.0	0.0	3.3	13.3	0.0
Qwen3-1.7B-16K	2025	1.7B	RK3588	15.1	36.7	**5.0**	5.0	0.0	23.3	35.8	0.0
Qwen3-1.7B	2025	1.7B	Server	15.8	38.3	**5.0**	5.0	0.0	25.0	37.5	0.0
MiniCPM-2B-DPO	2024	2B	Server	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MiniCPM5-2B	2026	2B	Server	8.8	10.8	**5.0**	**26.7**	0.0	6.7	12.5	0.0
Qwen3.5-2B	2026	2B	Server	**20.5**	**47.5**	**5.0**	0.8	0.0	**39.2**	**50.8**	0.0
Qwen2.5-3B-Instruct	2024	3B	Server	16.8	44.2	0.8	8.3	0.0	33.3	30.8	0.0
3–7B											
MiniCPM3-4B	2024	4B	Server	14.5	40.0	0.0	13.3	0.0	23.3	25.0	0.0
Qwen2.5-7B-Instruct	2024	7B	Server	**30.0**	**65.0**	**10.0**	**25.8**	0.0	**51.7**	**57.5**	0.0
Large language models											
GLM-5.2	2026	–	API	77.0	100.0	77.5	70.0	45.8	99.2	100.0	46.7
MiMo-V2.5-Pro	2026	–	API	71.3	98.3	61.7	55.0	43.3	95.8	100.0	45.0
MiniMax-M3	2025	–	API	62.3	86.7	50.8	67.5	26.7	74.2	86.7	43.3
Qwen3.8-Flash	2026	–	API	69.5	99.2	75.0	51.7	19.2	94.2	100.0	47.5

Table 2: Strict task-family success (%) by model under State-Seeking (SS). Rows follow the State-Given table; year, size, and platform are given there. F is the unweighted mean across the seven task families. The three local size bands and the API group contain the 22 models with complete bilingual overall and task-family coverage. Scores are averaged over Chinese and English; bold and underlined values mark the best and second-best local models within each size range.

Model	F	DGC (%)	MSC (%)	AIR (%)	FAR (%)	DCR (%)	PMU (%)	REQ (%)
≤1B								
MiniCPM4-0.5B-16K	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Qwen2.5-0.5B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Qwen3-0.6B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Qwen3.5-0.8B-16K	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Qwen3.5-0.8B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MiniCPM-1B-SFT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MiniCPM-S-1B-SFT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1–3B								
LFM2.5-1.2B-Instruct	0.1	0.0	0.0	0.8	0.0	0.0	0.0	0.0
DeepSeek-R1-Distill-Qwen-1.5B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Qwen2.5-1.5B-Instruct	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Qwen3-1.7B-16K	3.2	9.2	0.0	5.8	0.0	7.5	0.0	0.0
Qwen3-1.7B	3.7	**10.0**	0.8	**6.7**	0.0	7.5	0.8	0.0
MiniCPM-2B-DPO	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MiniCPM5-2B	3.7	9.2	**5.0**	4.2	0.0	0.0	7.5	0.0
Qwen3.5-2B	**3.9**	3.3	0.0	5.0	0.0	**10.0**	**9.2**	0.0
Qwen2.5-3B-Instruct	2.7	6.7	0.0	0.0	0.0	5.0	7.5	0.0
3–7B								
MiniCPM3-4B	**9.6**	**20.0**	**2.5**	**5.8**	0.0	18.3	**20.8**	0.0
Qwen2.5-7B-Instruct	8.0	10.0	0.0	0.0	0.0	**38.3**	7.5	0.0
Large language models								
GLM-5.2	68.9	99.2	50.8	43.3	50.8	99.2	98.3	40.8
MiMo-V2.5-Pro	67.1	98.3	46.7	43.3	45.0	100.0	99.2	37.5
MiniMax-M3	60.4	91.7	47.5	42.5	34.2	84.2	85.8	36.7
Qwen3.8-Flash	69.2	97.5	56.7	42.5	45.0	98.3	98.3	45.8

The two interface tables show a separation between the API reference group and the complete local comparison. Across the four API-accessed large language models, family mean success ranges from

Table 3: Language-specific overall strict success (%) for the complete bilingual model pool. Each cell is the mean across models in the group; n is the number of models.

Group	n	Chinese SG	Chinese SS	English SG	English SS
≤1B	7	0.0	0.0	1.2	0.0
1–3B	9	9.5	2.4	8.5	1.3
3–7B	2	22.5	6.7	22.0	11.0
API models	4	80.0	78.5	60.1	54.3

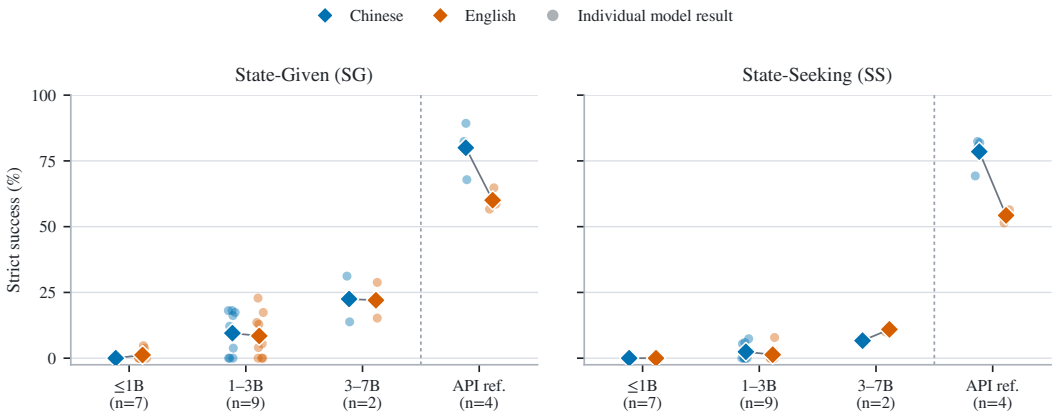

Figure 2: **Overall strict success by model group and interface.** Small points show per-model, per-language scores; diamonds show unweighted group means. Each underlying model-language score covers 420 cases. Size-band summaries include models with complete Chinese and English task-family results; Qwen3.5-4B is excluded because its English task-family rows are unavailable. The API group is a reference, not a size band.

62.3% to 77.0% under SG and from 60.4% to 69.2% under SS. The Chinese-only Qwen3.5-4B row reaches 46.4% and 45.5%, but it is not directly compared with bilingual family means. Among complete bilingual local rows, the highest SG family mean is 30.0% for Qwen2.5-7B-Instruct, while the highest SS family mean is 9.6% for MiniCPM3-4B. The language table and Figure 2 report overall micro-success: the API group averages 80.0/78.5% in Chinese and 60.1/54.3% in English under SG/SS, respectively. This descriptive contrast supports retaining language-specific results because a bilingual average can hide substantial differences. API/local differences remain descriptive because model family, release, serving environment, and platform are not matched.

4.2 Model Size and Interface Effects

The local results show higher capability in the larger observed size ranges, with substantial variation within each range. Averaging the model-language family means gives 0.6/0.0%, 9.0/1.9%, and 22.3/8.8% for SG/SS in the ≤1B, 1–3B, and 3–7B ranges. These summaries contain 7, 9, and 2 complete bilingual local models, respectively, and are descriptive rather than a causal scaling law. The 3–7B summary excludes Chinese-only Qwen3.5-4B. Within the Qwen3.5 family, the available family means rise from 2.4/0.0% at 0.8B to 20.5/3.9% at 2B; the 4B result is Chinese-only and is not part of that bilingual trend.

SG rewards models that can map a complete description directly to a valid outcome, whereas SS additionally requires selecting inspections, interpreting tool results, and continuing the interaction. The observed SS difference is 0.6 points in the smallest band, 7.1 points in the middle band, and 13.5 points in the largest band. These differences combine information visibility with the interaction policy and step budget, so they should not be read as the isolated effect of tool use. Removing AIR from the family mean changes each complete local size-band summary by at most 1.0 percentage point in the current data, although the AIR scores themselves still depend on the LLM judge.

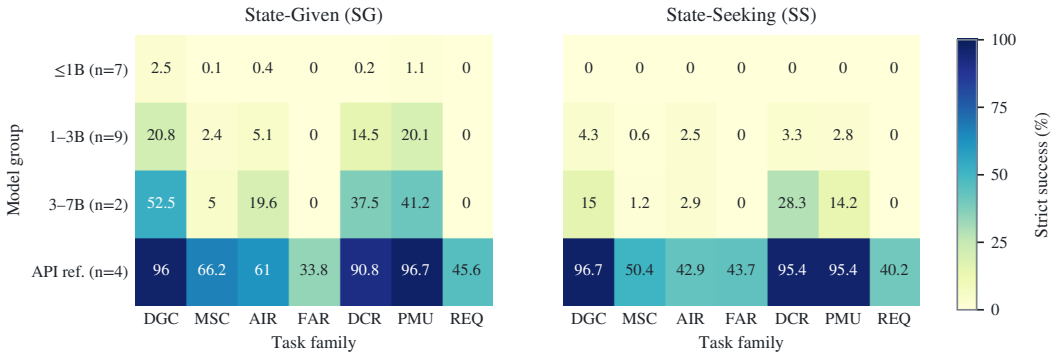

Figure 3: **Strict success by task family and model group.** Cells report the unweighted mean of available per-model, per-language rates for prompt-only runs; each underlying family rate is based on 60 cases. DGC: device grounding and control; MSC: multi-device selection and control; AIR: ambiguity and intent resolution; FAR: future automation registration; DCR: dialogue-context resolution; PMU: personalized memory use; REQ: read-only environment queries. Qwen3.5-4B is excluded from size-band means because its English family results are unavailable.

The task-family breakdown explains the modest means. Figure 3 shows that the complete bilingual local pool remains at 0.0% on FAR and REQ under both interfaces, while device grounding, dialogue-context resolution, and memory use are comparatively stronger in the largest local band. The Chinese-only Qwen3.5-4B row is strong on DGC and PMU, reaching 90.0% and 96.7% under SG and 86.7% and 95.0% under SS, but its coverage prevents a bilingual comparison. API models reach 19.2–50.8% on FAR and 36.7–47.5% on REQ. MSC and AIR remain bottlenecks; AIR should be interpreted with the stated judge limitation. Since each family contains 60 cases, a single changed case moves a family rate by about 1.7 percentage points.

4.3 Physical-platform capability comparison

The evaluation includes three physical Rockchip RK3588 deployments: MiniCPM4-0.5B-16K, Qwen3.5-0.8B-16K, and Qwen3-1.7B-16K. Qwen3-1.7B-16K is the strongest tested board model, reaching 17.4% Chinese and 12.9% English State-Given success; its State-Seeking success drops to 5.5% and 0.1%, respectively. Figure 4 compares it with the nominally matched server Qwen3-1.7B result: bilingual State-Given/State-Seeking success is 15.8%/3.7% on Server and 15.1%/2.8% on RK3588, a difference of 0.7/0.9 percentage points. For this single matched pair, the rates are similar; the result does not establish a general platform effect because model, backend, and hardware pairs are limited. The smaller Qwen3.5-0.8B deployment remains near the floor, with State-Seeking at 0.0%. These are task-level strict-success rates over 420 records per cell. The tested RK3588 results therefore support capability characterization and gated assistance, rather than a claim of autonomous control readiness.

5 Discussion, Limitations, and Broader Impact

Main implications. EdgeBench-Home provides a concrete, auditable protocol for evaluating small language models on smart-home tasks. Its suite brings grounding, multi-device control, dialogue context, personalized memory, queries, and automation into a common structured environment with explicit task constraints. In the complete bilingual comparison, the API reference models score higher than the tested local models, while local family means are higher in the larger observed size ranges. State-Seeking remains difficult for local models, especially in the 1–3B and 3–7B groups. The Qwen3-1.7B RK3588 result is useful as a physical-platform capability point, but its low strict success supports gated assistance rather than autonomous control.

Scope and reproducibility. The simulator and evaluator define deterministic replay and scoring contracts, while the RK3588 study tests the same task capability on one physical platform. Hardware, user, acoustic, vendor, latency, and privacy studies remain distinct reporting dimensions. The

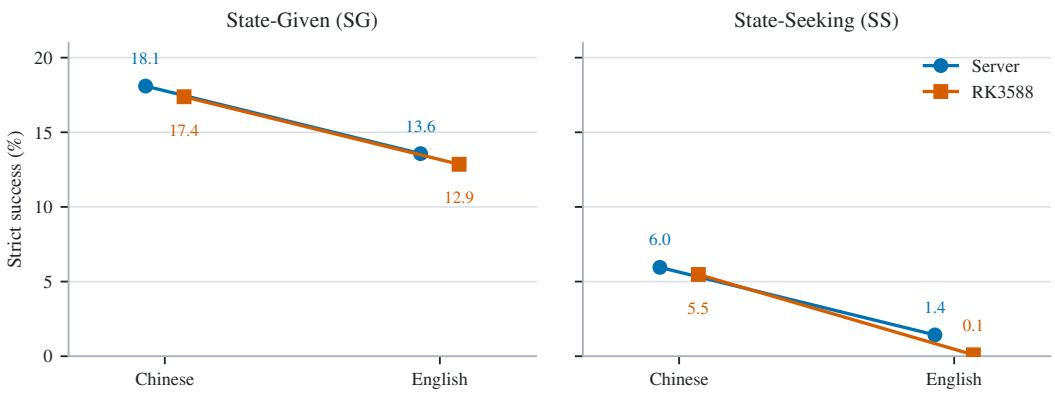

Figure 4: **Matched Qwen3-1.7B comparison across Server and RK3588.** Points show strict success under the prompt-only protocol for Chinese and English; each language-interface result covers 420 cases. The State-Given and State-Seeking panels use the same vertical scale.

benchmark contains seven families with 60 cases per family and two language-specific tracks; it does not cover all languages, devices, failure modes, or household practices. The primary model comparison has 22 complete bilingual models, while Qwen3.5-4B has Chinese task-family coverage only. Any released traces should remove secrets and personal information, and any integration that can trigger consequential actions should document its confirmation policy.

The evaluation protocol specifies typed tool contracts, deterministic state transitions, and executable scoring. A model result requires raw per-episode traces and evaluator metadata, not aggregates alone. Overall micro success and family means are preserved as separate quantities because they use different aggregation levels. The bilingual release contains 420 case IDs per language; within each language, State-Given and State-Seeking are derived views over the same IDs. The size-band results are descriptive because the ranges contain different model families and only two complete bilingual models in the largest local band. The current campaign also lacks completed human agreement results for AIR, so semantic clarification scores should be treated as provisional.

Broader impact. Smart-home agents can make local interaction more convenient, but incorrect actions can affect privacy, comfort, access, and physical safety. Local execution may reduce the need to send some inputs to a remote service, but it does not by itself guarantee privacy or prevent unsafe behavior. The task families in EDGEBENCH-HOME include read-only requests, ambiguity handling, personalized preferences, future rules, and preservation of unrelated state, exposing risks that answer-level accuracy can miss. The results do not establish deployment safety: even the strongest tested local models remain unreliable on several task families, especially under State-Seeking. Practical systems should require confirmation for consequential actions, restrict autonomous actuation, and provide fallback behavior when the model is uncertain. The controlled bilingual benchmark also covers a limited set of languages, devices, and user situations; its results should not be interpreted as universal evidence of fairness or reliability.

REFERENCES

Marah Abdin et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

Loubna Ben Allal et al. Smollm: Towards small language models. *arXiv preprint arXiv:2409.15790*, 2024.

Manasa Bharadwaj, Yolanda Liu, InJung Yang, Sungil Kim, Nikhil Verma, KoKeun Kim, Kevin Ferreira, and YoungJoon Kim. PersonalHomeBench: Evaluating agents in personalized smart homes. *arXiv preprint arXiv:2604.16813*, 2026. URL https://arxiv.org/abs/2604.16813v3.

Rune Birkmose, Nathan Mørkeberg Reece, Esben Hofstedt Norvin, Johannes Bjerva, and Mike Zhang. On-device LLMs for home assistant: Dual role in intent detection and response generation. In *Proceedings of the Tenth Workshop on Noisy and User-generated Text*, pp. 57–67. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.wnut-1.7/.

Diane J. Cook et al. Casas: A smart home in a box. *Computer*, 46(7):62–69, 2013.

Edoardo Debenedetti et al. Agentdojo: A dynamic environment to evaluate prompt injection attacks and defenses for llm agents. In *Advances in Neural Information Processing Systems*, 2024.

Albert Q. Jiang et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Carlos E. Jimenez et al. Swe-bench: Can language models resolve real-world github issues? In *International Conference on Learning Representations*, 2024.

Evan King, Haoxiang Yu, Sangsu Lee, and Christine Julien. Sasha: Creative goal-oriented reasoning in smart homes with large language models. *arXiv preprint arXiv:2305.09802*, 2023a. URL https://arxiv.org/abs/2305.09802.

Evan King, Haoxiang Yu, Sangsu Lee, and Christine Julien. "get ready for a party": Exploring smarter smart spaces with help from large language models. *arXiv preprint arXiv:2303.14143*, 2023b. URL https://arxiv.org/abs/2303.14143.

Stefanos Laskaridis, Kleomenis Katevas, Lorenzo Minto, and Hamed Haddadi. MELTing point: Mobile evaluation of language transformers. *arXiv preprint arXiv:2403.12844*, 2024. URL https://arxiv.org/abs/2403.12844v4.

Cheng Li et al. Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation. In *Conference on Computer Vision and Pattern Recognition*, 2024.

Kuan Li, Shuo Zhang, Huacan Wang, Fangzhou Yu, Zecheng Sheng, Yi Gu, Weipeng Ming, Lei Xue, Chen Liu, Sen Hu, Ronghao Chen, Siyue Lin, Yuqing Hou, Xiaofeng Mou, and Yi Xu. SMH-Bench: Benchmarking LLM agents for environment-grounded reasoning and action in smart homes. *arXiv preprint arXiv:2606.01912*, 2026. URL https://arxiv.org/abs/2606.01912v1.

Silin Li, Yuhang Guo, Jiashu Yao, Zeming Liu, and Haifeng Wang. HomeBench: Evaluating LLMs in smart homes with valid and invalid instructions across single and multiple devices. *arXiv preprint arXiv:2505.19628*, 2025. URL https://arxiv.org/abs/2505.19628v2.

Eu Jin Lim, Zhaoxing Li, and Sebastian Stein. AdaHome: An adaptive smart home assistant using local small language models. *arXiv preprint arXiv:2607.18034*, 2026. URL https://arxiv.org/abs/2607.18034.

Xiao Liu et al. Agentbench: Evaluating llms as agents. In *International Conference on Learning Representations*, 2024.

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. ToolSandbox: A stateful, conversational, interactive evaluation benchmark for LLM tool use capabilities. *arXiv preprint arXiv:2408.04682*, 2024. URL https://arxiv.org/abs/2408.04682v2.

Rithesh Murthy, Liangwei Yang, Juntao Tan, Tulika Manoj Awalgaonkar, Yilun Zhou, Shelby Heinecke, Sachin Desai, Jason Wu, Ran Xu, Sarah Tan, Jianguo Zhang, Zhiwei Liu, Shirley Kokane, Zuxin Liu, Ming Zhu, Huan Wang, Caiming Xiong, and Silvio Savarese. MobileAIBench: Benchmarking LLMs and LMMs for on-device use cases. *arXiv preprint arXiv:2406.10290*, 2024. URL https://arxiv.org/abs/2406.10290v1.

Xavier Puig et al. Virtualhome: Simulating household activities via programs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Gyuhyeon Seo, Jungwoo Yang, Junseong Pyo, Nalim Kim, Jonggeun Lee, and Yohan Jo. SimuHome: A temporal- and environment-aware benchmark for smart home LLM agents. *arXiv preprint arXiv:2509.24282*, 2025. URL https://arxiv.org/abs/2509.24282v3.

Gemma Team et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Tianbao Xie et al. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. In *Advances in Neural Information Processing Systems*, 2024.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. Tau-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*, 2024. URL https://arxiv.org/abs/2406.12045v1.

Shuyan Zhou et al. Webarena: A realistic web environment for building autonomous agents. In *International Conference on Learning Representations*, 2024.