SurveyAtlas: A Living Literature Review for Embodied Navigation

Xunyi Zhao[1,2] **Jian Zhou**[1] **Sihao Lin**[1,2] **Xinyu Yan**[1,2] **Jiajun Liu**[2,3] **Qi Wu**[1,2,†]

[1]Adelaide University [2]Responsible AI Research Centre, AIML [3]CSIRO Data61
[†]Corresponding author

Like many other fast-moving research areas, embodied navigation has seen rapid growth in methods and benchmark results, particularly in vision-and-language navigation (VLN). Reviews help track this progress, but static reviews soon become outdated and differing evaluation protocols make scores hard to compare. We present SurveyAtlas, an open-source system that maintains living literature reviews from field-specific configurations. Its embodied-navigation instance is NavAtlas. LLM agents find, screen and read papers, and protocol-aware result extraction links each result to its evidence and evaluation protocol. The review incorporates new papers, and leaderboards rank only standard-protocol results. We evaluate NavAtlas against human-written surveys, LLM judges and navigation researchers. It covers nearly every core paper the surveys cite, and its comparability judgments on top results agree with researchers at least as closely as LLM judges do, at one twentieth of the cost of judging every result. Updates cost cents per new paper, and annotators rated NavAtlas's answers to their questions above answers from full-text retrieval. The checked record shows that the top result on 8 of 12 major leaderboards is a main result outside the standard protocol, and training-free agents trail learned policies far more on VLN than on object and image goals.

Date: October 13, 2026
Project Page: https://billzhao1030.github.io/SurveyAtlas/
GitHub: https://github.com/billzhao1030/SurveyAtlas
Live Atlas: https://huggingface.co/spaces/SurveyAtlas/navatlas

1 Introduction

Keeping up with embodied navigation, especially vision-and-language navigation (VLN), requires following both new methods and changing benchmark results. VLN asks an agent to follow natural-language instructions through a 3D environment (Anderson et al., 2018b). It is a language grounding task with a long history at ACL venues (Ku et al., 2020; Gu et al., 2022), and surveys help researchers understand the methods developed for it. The surveys of Wu et al. (2024) and Gu et al. (2022) cite about half of the papers available when they first appeared, whereas the most recent survey (Wang et al., 2026) cites about a fifth (Figure 1a). These shares reflect survey selection rather than review quality, but a fixed reference list still leaves researchers to find and assess later papers themselves.

Keeping that record current is only part of the problem: a review must also establish which results can be compared. Even on the same VLN benchmark, papers may evaluate on different subsets of test episodes or use inputs unavailable under the standard protocol. On R2R-CE val-unseen, for example, a reported success rate of 93.3 uses one test scene, whereas the comparable full-split leader reaches 77.3 (Table 3). Authors usually disclose such choices, but when a result is copied into another paper's table, the caveat survives at best in an easily overlooked footnote or separate column. Language-model tools now generate surveys (Wang et al., 2024c; Liang et al., 2025) and answer questions about the literature (Skarlinski et al., 2024; Asai et al., 2024), but none of those compared in Table 1 checks whether the results it places side by side are comparable. A researcher asking which method is currently best therefore needs both an up-to-date record and the evaluation conditions behind its numbers.

These needs motivate a living review that incorporates new papers, traces its statements to their sources and ranks results only under comparable conditions. Living systematic reviews in medicine provide a precedent: they

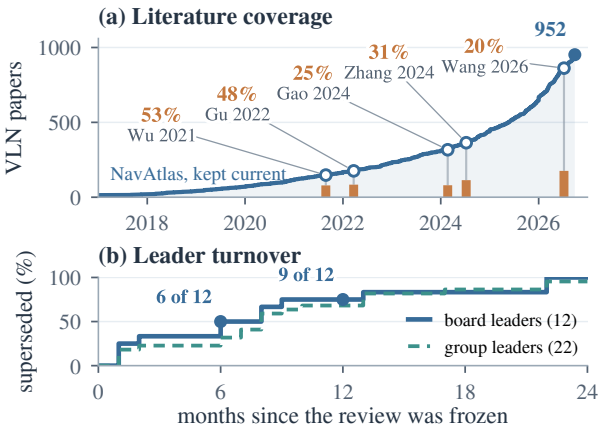


Figure 1 Static reviews fall behind. (a) Core VLN papers available (circles) and cited (bars) when each survey first appeared. (b) Share of leaders on 12 boards that changed after a review was frozen.

assess each trial's risk of bias and are updated as new trials appear (Elliott et al., 2014). For the navigation literature, language-model agents can make retrieval and reading affordable, but a correctly extracted score still does not tell us whether it belongs in the same ranking as another. We address this missing step with *protocol-aware result extraction*, which records a result's benchmark, evaluation set and departures from the standard protocol together with supporting evidence. The task extends leaderboard extraction (Hou et al., 2019; Kardas et al., 2020) from a result's task, dataset, metric and score to the conditions that make two scores comparable. These conditions also matter for NLP leaderboards that mix zero-shot and few-shot results or test subsets.

We present SurveyAtlas, an open-source system that maintains such a review (Figure 2). Maintainers first review a field configuration that specifies the literature's scope, taxonomy and benchmark protocols. Guided by this file, agents find, screen and read papers, recording notes and results with their source evidence. The system then checks each result's evaluation conditions before building leaderboards, and results outside the standard protocol remain visible with the reason they were set aside. Scheduled updates repeat these steps as new papers appear, and a website serves the resulting review, which we call an *atlas*. Because the atlas keeps evidence and protocol decisions with each result, readers can inspect or revise a ranking without searching for and rereading papers.

NavAtlas, the review that SurveyAtlas maintains for embodied navigation, holds 3,147 papers and 9,806 extracted results. We evaluate it with VLN as the central use case and object- and image-goal tasks as a comparison. The checked record changes the picture presented by headline numbers: on 8 of 12 major navigation leaderboards the highest reported result is a main result obtained on an episode subset or under a changed protocol, and training-free agents still trail learned policies on the full-split VLN benchmarks we study, although they come within 12 points or lead on object and image goals (§5). Our contributions are as follows.

- SurveyAtlas, an open-source system for maintaining evidence-linked literature reviews from field-specific configurations (§3).
- Protocol-aware result extraction, a new information extraction task, solved by a cheap check of every result and a targeted audit of likely leaders. On top-ranked results it agrees with researchers at least as closely as LLM judges do at one twentieth of the cost of judging every result, and its check transfers to robot manipulation (§3.4, §4). We release 520 judged rows from navigation and manipulation boards and the human labels.
- NavAtlas, the embodied-navigation atlas built with SurveyAtlas, its released snapshot and the findings supported by its checked record (§5).

2 Related Work: Keeping Up with a Field

Surveys and living reviews. Living systematic reviews in medicine use machine learning to screen studies (Marshall & Wallace, 2019; Wang et al., 2024b) and assess risk of bias (Marshall et al., 2016), and Trialstreamer maintains a database of trial reports with extracted details (Marshall et al., 2020). In particle physics, a curated list provides a living review of machine-learning papers (Feickert & Nachman, 2021). Computer science relies on GitHub paper lists, which record which papers exist but not how they were evaluated. SurveyAtlas fills this gap with protocol flags, which play the role that risk-of-bias assessment plays for clinical trials.

Generated surveys and scholarly search. STORM writes Wikipedia-style articles from web sources (Shao et al., 2024), while AutoSurvey, SurveyX and SurveyForge generate surveys from a few hundred retrieved papers (Wang et al., 2024c; Liang et al., 2025; Yan et al., 2025). Related systems draft related-work sections (Agarwal et al., 2024), build comparison tables (Newman et al., 2024), search with an agent (He et al., 2025) or write cited reports from the web (OpenAI, 2025). These approaches return documents, and two recent proposals keep such documents current: the Dynamic Survey Framework routes new papers to sections of an existing survey (Mumcu et al., 2026), and a position paper proposes community-maintained live surveys (Lin et al., 2025). An atlas instead maintains a database of papers and checked results that

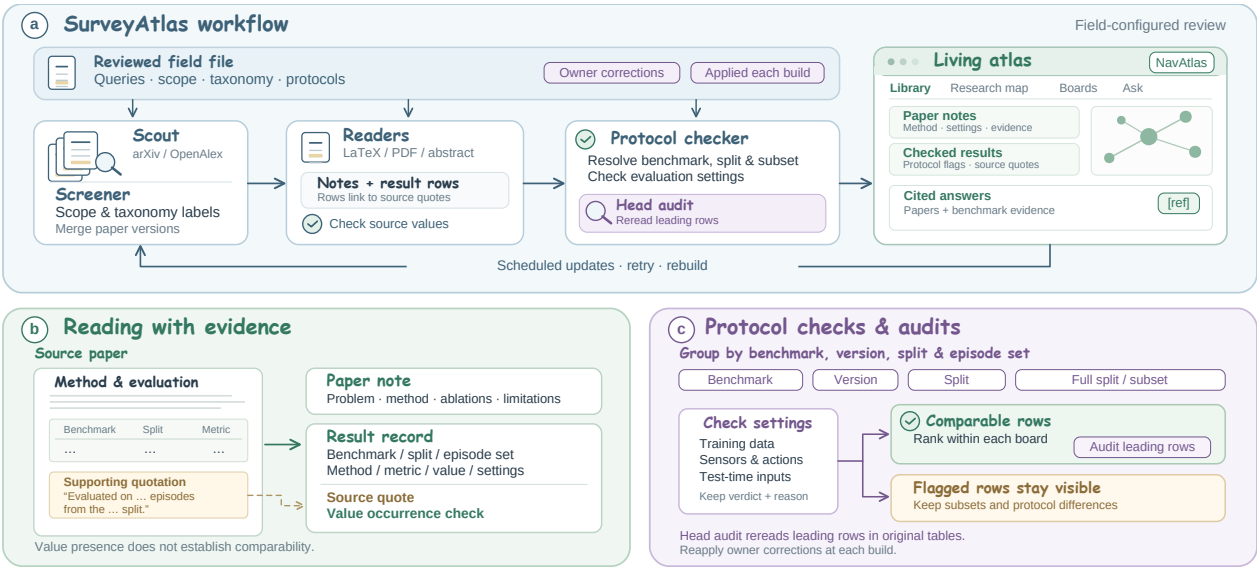


Figure 2 SurveyAtlas's workflow. (a) Field-configured agents maintain an evidence-linked atlas. (b) Readers extract notes and results with source quotations. (c) Checks and targeted audits rank comparable results, and flagged rows stay visible with their reasons.

readers can filter, rank and query. Semantic Scholar (Kinney et al., 2023), Elicit (Elicit, 2023), NLP-KG (Schopf & Matthes, 2024) and CHIME (Hsu et al., 2024) help readers search and organize the literature. PaperQA2, OpenScholar and Ai2 Scholar QA synthesize cited answers from large corpora (Lála et al., 2023; Skarlinski et al., 2024; Asai et al., 2024; Singh et al., 2025), but none maintains a field's structure over time or a checked record of its results.

Results and leaderboards. Extracting a result's task, dataset, metric and score is an established task (Hou et al., 2019; Kardas et al., 2020; Jain et al., 2020), recently addressed with large language models (LLMs) (Şahinuç et al., 2024; Singh et al., 2024). The Open Research Knowledge Graph also mines leaderboards from structured paper descriptions (Jaradeh et al., 2019; Kabongo et al., 2023). These efforts extract the numbers correctly but do not record whether they can be compared, and only challenge servers fix the protocol of their test splits (Yadav et al., 2019). Yet differences in experimental setup are a major source of misleading comparisons (Dehghani et al., 2021; Henderson et al., 2018; Dodge et al., 2019). In embodied navigation, beam search (Fried et al., 2018) and prior exploration of test environments (Wang et al., 2019) have been debated since the earliest benchmarks (Anderson et al., 2018a; Batra et al., 2020). Protocol-aware extraction therefore records these choices for every result.

3 SurveyAtlas

SurveyAtlas builds living reviews through retrieval, reading and protocol checks (§1). We describe its workflow using NavAtlas, its embodied-navigation instance. Scheduled updates keep the atlas current by adding and reading new papers. Each number is stored with its source passage, while notes and answers link to the papers they draw on. Each result also receives a comparability verdict that decides whether it can be ranked.

3.1 A field in one file

The field file specifies search queries, inexpensive prefilters, labeling rules in plain English, and landmark papers that every build must contain (Appendix A). It also defines a taxonomy of tasks, paradigms and traits, and lists benchmarks with their aliases, versions and standard protocols. The taxonomy organizes the library and research map, while the benchmark entries determine which results can share a leaderboard. The *owner*, the researcher or team maintaining the atlas, reviews this configuration before the first build. In NavAtlas, scope labels separate core navigation papers from adjacent and aerial work, and six paradigms range from task-specific learning to agentic systems in which a model controls its own loop. Although we focus on embodied navigation, SurveyAtlas is designed to support living literature reviews in other rapidly evolving research fields through field-specific configuration. The same engine built atlases of embodied AI and LLM agents and a pilot on perovskite

		Survey paper	Living review	Paper list	Leader-board	Knowledge graph	Scholarly QA	Generated survey	Updated survey	Deep research	**Survey-Atlas**
Living	Updates as papers appear	✗	∼	∼	∼	∼	✓	✗	✓	✗	✓
	Keeps a field's corpus	∼	✓	✓	✓	✓	✗	✗	∼	✗	✓
	Cost to keep current	high	high	medium	medium	medium	low	per run	low	per run	low
	New field without setup	✗	✗	✗	∼	✗	✓	✓	✗	✓	∼
Structure	Field taxonomy and scope	✓	✓	∼	∼	∼	✗	∼	✓	✗	✓
	Reads every paper	∼	✓	✗	∼	✗	✗	✗	∼	✗	∼
	Argues a view of the field	✓	✓	✗	✗	✗	✗	∼	✓	∼	✗
	Expert judgment per entry	✓	✓	∼	∼	∼	✗	✗	∼	✗	∼
Results	Extracts reported results	∼	✓	✗	✓	∼	∼	✗	∼	∼	✓
	Quotes evidence per number	✗	∼	✗	✗	✗	∼	✗	✗	✗	✓
	Records protocol departures	∼	✓	✗	∼	✗	✗	✗	✗	✗	✓
	Cited answers to questions	✗	✗	✗	✗	∼	✓	✗	✗	✓	✓

Table 1 Ways of keeping up with a field, in our assessment. ✓ yes, ✗ no, ∼ partly or by hand. Examples in column order (§2): Trialstreamer, Awesome lists, Papers with Code, Open Research Knowledge Graph, OpenScholar, AutoSurvey, Dynamic Survey Framework, OpenAI deep research.

tandem solar cells. Screening in the pilot is as precise as in our AI atlases, although its first search missed many older papers that describe the devices in other terms (Appendix J).

3.2 Finding the papers

Search prioritizes recall, since no classifier can recover a paper that was never retrieved. Broad queries are complemented by benchmark names and method families to cover papers using different terminology. We search arXiv and OpenAlex (Priem et al., 2022), including journals and proceedings, and merge preprints with their published versions. Screening narrows this broad collection to the field's scope. Records with no field-related term are excluded, while titles and abstracts provide the basis for scope, task, setting, paradigm, trait and benchmark labels and a one-sentence summary. Papers with low-confidence labels receive a second review. For NavAtlas, this funnel reduces about twenty thousand records to 3,147 papers, including 2,212 core papers, and the rule or label behind each exclusion is kept so that a missing paper can be traced (Appendix B).

3.3 Reading with evidence

Reading agents turn each core and adjacent paper into structured notes and result rows that readers can inspect alongside the source evidence. They read arXiv LaTeX sources or PDF text, falling back to the abstract when the full text is not openly available. The notes summarize each paper's contribution, ablation-supported insights and limitations, while setting descriptions record training data, backbone, observation and action spaces, and privileged inputs. Each result row stores the benchmark, split, evaluation set (full split or subset), method and metric values, together with a supporting quotation. Sonnet reads most papers, and Opus reads the most cited and landmark papers as well as any paper whose numbers fail a deterministic check, which looks for every extracted value in the text, allowing for rounding and percentages. A row is *verified* when all its values are found. In NavAtlas, the check finds 99% of values and verifies 98% of rows. Only verified rows, or rows confirmed by a later audit, can enter a leaderboard. Finding a number in the source does not show that it is assigned to the right method or setting, which we evaluate separately (§4).

3.4 Protocol-aware result extraction

A correctly extracted number can still mislead when compared with results from different settings, so we define protocol-aware result extraction.

> *Input:* a paper and one of its reported results.
> *Output:* the canonical benchmark and its version, the split, whether the full split or a subset of episodes was evaluated, and whether the protocol departs from the benchmark's standard one, with the reason and a supporting passage.

Results can share a ranking only when they refer to the same benchmark version, split and evaluation set. Aliases in the field file map benchmark names to *boards*, each ranking results for one benchmark and split. A resolver separates incompatible versions with the same name, such as HM3D ObjectNav evaluated on two releases of its scene annotations (Ramakrishnan et al., 2021; Yadav et al., 2023). Subset results are grouped by their evaluation subset and never ranked against full-split results.

Within each board, Sonnet performs a *protocol check*, comparing the recorded setting with the benchmark's standard protocol and returning a comparability verdict

with a reason (Appendix D). Because readers are most likely to cite the highest-ranked results, Opus performs a *head audit* of the 15 highest comparable rows on each of the 13 boards that the atlas summarizes in its tables. It rereads the original tables and overrides verdicts it disagrees with. The owner can correct any decision in a file that every build applies. Each board then ranks the comparable rows, whose top row is its *leader*, and keeps every other row visible with the reason it was set aside. On R2R-CE, for example, the six highest results on the unseen validation split use an episode subset or a changed protocol, so the comparable ranking starts with the seventh result (Figure 3a).

Recorded facts stay separate from board policy: reading records the episode set, sensors, action space, step budget and test-time inputs, while the field file states what each benchmark's protocol allows. An owner who disagrees with a policy can therefore edit a protocol note and rerun the check for about $7 without rereading any paper. When a departure is unclear, the check accepts the result and leaves it to the head audit to catch a wrong acceptance among the highest-ranked rows. Comparability concerns a board's protocol, not a paper's quality: subset experiments and oracle analyses can answer useful questions outside that protocol, and their retained evidence shows readers why they are excluded from the standard ranking.

3.5 The living review

A website serves the atlas (Appendix K), providing a filterable library, a research map organized by task and paradigm, a timeline and individual paper pages with notes and checked results. Ask answers questions about the field (Figure 3b). It ranks papers with BM25 (Robertson & Zaragoza, 2009) over titles, abstracts and notes and adds the full-split rows and flags of any benchmark named in the question. Sonnet then answers with citations to atlas entries only, and for benchmark questions it can use a board's ranking directly, together with the protocol decisions behind it.

Daily or weekly updates apply the same reading and comparability checks to new arXiv papers, keeping the review current. New papers appear in a digest, and failed downloads or model calls are retried on the next run. Every build recomputes the boards and Ask's search index, although later arXiv versions of read papers are not yet reread.

4 Evaluating NavAtlas

We evaluate whether the October 2026 NavAtlas *snapshot* provides the maintained record of navigation research described in §1. Complete and Current assess coverage and freshness, while Traceable and Comparable assess source support and leaderboard comparability. We also test whether this record helps researchers answer questions (summary in Appendix Table 9). Three embodied-navigation researchers, including one author, provided human labels without seeing any system or judge label, and none of them wrote the owner corrections or the protocol notes. All three independently labeled the protocols of 64 navigation rows: 40 random top-ten rows and 24 rows supporting the findings in §5. Another 60 screening decisions, 40 extracted rows and 40 robot manipulation protocol rows were each labeled by one annotator (Appendix E). Two LLM judges, Claude Opus 5.5 and the open-weight Qwen3.8-27B (Qwen Team, 2025), label larger samples and see system outputs only when verifying extracted rows. Because Opus also takes part in the system and LLM judges can favor their own family (Panickssery et al., 2024; Zheng et al., 2023), we report both judges and use the human labels to assess their reliability, noting that annotators and judges read the same keyword-selected passages.

Complete. NavAtlas covers 97% of the core navigation papers cited by six human-written surveys: the five in Figure 1 and one on LLMs for navigation (Lin et al., 2023). Seven of the nine missing papers were never retrieved by any query. Because survey references measure coverage rather than field-wide recall, we also check screening directly. It is precise: Opus considers 148 of 150 randomly sampled core papers in scope, and human scope labels agree more often with the system than with either judge, although by Opus's estimate screening excludes a few hundred relevant papers. Within read papers, a full-text reading of 114 paper and board pairs finds no missed comparable result that would lead a board (Appendix F).

Current. Leaders change every few months, but keeping up with them is inexpensive. The comparable leaders of the 12 boards changed 32 times since January 2025, and a review frozen six months before the snapshot would name an outdated leader on half of them (Figure 1b). These counts rest on the system's decisions, which the annotators confirm for the current leaders. Reconstructing weekly updates shows that reading and checking a week's new papers costs a few dollars. Live updates after the snapshot cost about $0.2 per new paper (Appendix I). The replay assigns papers to weeks by their first posting, and the live cost also covers checks left over from earlier papers. Building the atlas from scratch costs about $498 at API list prices, and none of these figures includes the owner's effort or upkeep.

Traceable. Every leaderboard number is traceable to its source, and almost all checked values are correct.

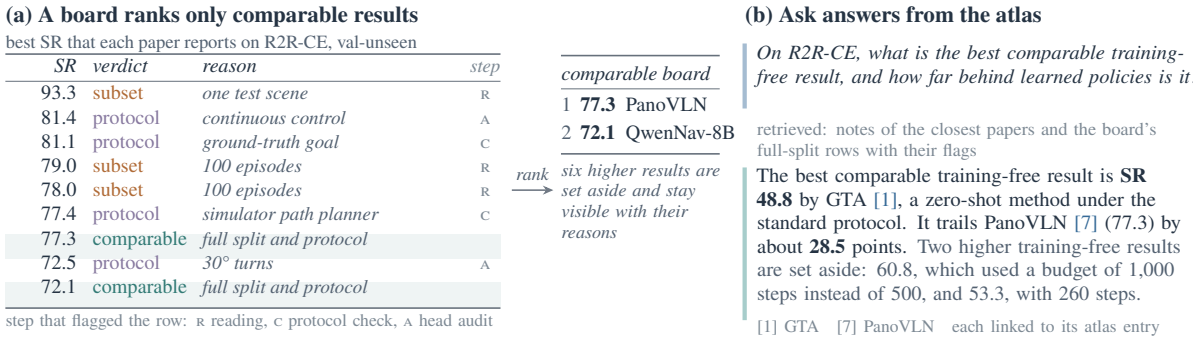


Figure 3 Two views of NavAtlas, October 2026. (a) A board ranks only comparable results. Excluded results retain the reason and the step that flagged them. (b) Ask cites atlas entries and here excludes results flagged by owners.

		F1 against			
	Flags from	Opus	Qwen	Annotators	$
rules	Subset label	.58	.54	.50	0
	Label + rules	.75	.75	.83	0
Survey-Atlas	Label + check	.84	.79	.86	7
	Check + audit	**.87**	.78	**1.00**	27
judges	Opus judge	–	.84	.91	
	Qwen judge	.84	–	.77	
ref.	Judge prompt, Sonnet	.91	.84	.86	546

Table 2 Protocol-flag F1 against two LLM judges on the rows each decides (220-row sample) and against three annotators' majority labels (40 top-ten rows). Unclear judge predictions count as no flag. $: cost for all 9,806 results. Full results: Table 12.

Each row quotes a supporting passage and enters a board only after its values are checked against the paper (§3.3). Opus confirms all values in 145 of 146 sampled verified rows, and annotators confirm all values in their 40 rows. Benchmark and split are harder to verify, mainly because some papers omit the split.

Comparable. Comparability is where copied leaderboards fail and the system's design matters most. We sample 220 rows from the 12 boards in §5: the top ten of each and 100 random lower rows. A leaderboard that simply copies reported numbers agrees with the judges on only about half of these rows (Appendix Table 12). Subset labels catch fewer than half of the flagged rows because most departures concern the protocol rather than the episode set. These departures appear in the recorded settings, so keyword rules already help, and the protocol check and the head audit add the largest gains (Table 2).

Three annotators independently labeled 40 top rows under blind conditions, with high agreement (Fleiss' κ 0.89). The complete system matches their majority label on every row, more closely than either judge. These 40 rows do not establish that the system is error-free or represent its performance on lower rows. The head audit rereads accepted top results on all but the two ImageNav boards, and other rows rely on the check alone. The check agrees with the researchers more closely than Qwen but less closely than Opus. Human labels agree much more often with Opus than with Qwen, so we use Opus as the reference for larger samples, on which the check's gain over keyword rules is clear only against Opus (Appendix F). Applying the judges' own prompt to every result yields higher agreement with both judges than the check and audit do. It does not improve agreement with the annotators and costs twenty times as much. The system avoids this cost by reserving full-text scrutiny for the likely leaders. The audit raises agreement with Opus but lowers it with Qwen, perhaps because Opus judges rows it also audited.

On these well-known benchmarks, the check performs similarly without protocol notes or with open-weight Qwen3.8-27B replacing Sonnet (Appendix Table 12). A secondary study tests whether the protocol check transfers beyond navigation. On 300 rows from 12 robot manipulation boards, where departures concern the task set, the number of trials, the robot or privileged inputs, it clearly improves on subset labels, whereas rules written for manipulation add almost nothing. Human labels for 40 of these rows support this pattern, although both judges agree with them more closely than the check does (Appendix F).

Useful answers. On questions written by our annotators, the atlas's answers were rated above those based on full-text retrieval. The annotators wrote 15 questions before seeing any answer (Appendix H). The same Sonnet model answered each question in three conditions: through Ask, using BM25 passages from the same papers' full texts with an equal context budget, and without retrieval. Under blind grading for correctness, coverage, citation support and usefulness, Ask ranked best on 12 of the 15 questions and never worst, scoring 0.47 points above full-text retrieval on a five-point scale (paired bootstrap interval 0.30 to 0.62).

		Highest reported		Comparable leader	
	Board	SR	Flag	Learned	Training-free
instructions	R2R	91.1	P oracle*	**86.0**	49.5†
	R2R-CE	93.3	S one scene	**77.3**	48.8
	RxR-CE	78.0	*none*	**78.0**	40.7†
	REVERIE‡	68.1	P human help	**60.4**	33.2†
	SOON	50.9	*none*	**50.9**	n/a
object and image goals	HM3D ObjNav v1	86.6	P new actions	**79.1**†	72.3
	HM3D ObjNav v2	97.5	P pre-explored	83.7	**84.7**†
	MP3D ObjNav	85.6	P prebuilt map	**64.0**†	52.3
	HM3D-OVON	85.5	S re-curated	**66.3**	54.9
	GOAT-Bench	90.6	S subset	62.7	**83.7**
	Gibson ImageNav	97.5	S easy cases	**94.0**	n/a
	Inst. ImageNav	85.0	P ground truth*	63.0†	**82.6**

Table 3 Highest reported success rate (SR, %) and comparable group leaders on 12 major boards. Flags: episode subset (S) or changed protocol (P). Bold: higher group leader. *Labeled oracle or ground truth by its paper. ‡Excludes a flagged 76.9 from a custom evaluation. †Confirmed by annotators, not by both judges (Appendices E, G).

Opus also preferred Ask when grading the same answers (Appendix H). The comparison evaluates the atlas as a whole and does not separate the notes from the protocol information. Since only one of the questions names a benchmark, we test the protocol information separately with 36 template questions whose gold answers come from the atlas itself, and Ask answers 31 of them correctly, against 10 for full-text retrieval.

5 Findings on Embodied Navigation

We use the checked record to revisit how VLN methods are changing, which results can be compared and how training-free agents fare against learned policies. Related navigation tasks provide a comparison across goals. For two of these questions, we group results by the *decision model*, which chooses where the agent goes. A result is *learned* if its paper uses task-specific learning, large-scale pretraining or a fine-tuned foundation model, and its decision model is trained on navigation data. It is *training-free* if its paper presents a zero-shot modular pipeline or an agentic system, and its decision model is not trained on navigation data. Each group's highest comparable result on a board is its *group leader*. Three annotators independently labeled every group leader and highest main result on the 12 boards under blind conditions, using the same passages as the judges, and their majority labels agree with the system on every one of them (Appendix E). All findings retain their direction under every tested way of resolving disputed results, although some counts change (Appendix G).

Foundation models and training-free methods are gaining ground. Across the navigation corpus, task-specific policies trained from scratch accounted for nearly all core method papers in 2018 but for less than a third in 2026 (Figure 4a), giving way to fine-tuned foundation models and above all to training-free agents, whose share of method papers rose from 7% in 2022 to 40%. The shift varies by task (Appendix Figure 7): training-free methods account for two thirds of the method papers on instance navigation and close to half on object navigation, but only a fifth in instruction-following VLN, and they are rare when the goal is a point, an image view or a sound. Training-free adoption is uneven even for visual goals: it accounts for 45% of ObjectNav method papers but only 11% of ImageNav method papers.

The highest reported result is usually not comparable. Table 3 compares the highest reported success rate on each of 12 major boards with the comparable group leaders, and the highest result is flagged on 10 of these boards. Two are oracle variants that their papers label as such, while the other eight are main results obtained on episode subsets or under changed protocols, and the annotators confirm all eight departures. The gap can be large: on MP3D ObjectNav, the highest reported result relies on a map of the test scenes built in advance and exceeds the comparable leader by 21.6 points. Most top-ten rows on these boards are also flagged. Later papers that copy such a result mark the departure in 9 of the 11 cases we found, but only in a separate column, row group or footnote. Other information needed for comparison is often missing: the resolver finds no stated version in almost a third of HM3D ObjectNav papers, and few papers report runtime, token usage or monetary cost (Appendix C).

Training-free agents trail most on instruction following. On the full splits of the four VLN boards with comparable results from both groups, training-free agents trail learned policies by 27 to 37 points (Figure 4b). Without owner corrections, which determine three of the four training-free leaders, the gaps range from 14 to 37 points, and neither judge's verdicts close any of them (Appendix G). The distinction between full splits and subsets matters here. Across 24 common navigation benchmarks, training-free agents report subset results far more often than learned policies: 41% of their results against 6%. Many use the 100-episode subset released by Open-Nav (Qiao et al., 2025), where a success rate near 50% has a standard error of 5 points. On every object- and image-goal board with a training-free result, by contrast, these agents are within 12 points of learned policies or ahead (Figure 4c).

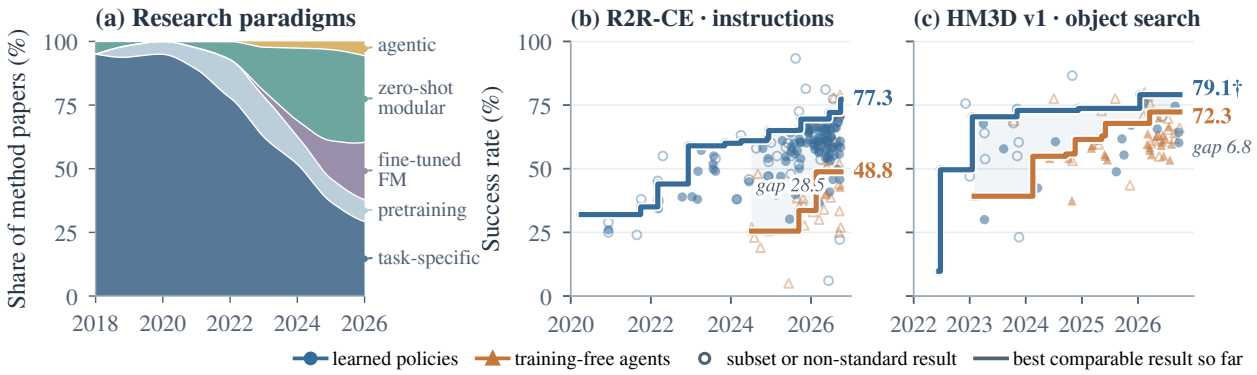


Figure 4 NavAtlas, October 2026. (a) Paradigm shares of core method papers, interpolated between years. (b, c) Best SR per paper, comparable group leaders (lines), and their gap (shading). †See Table 3.

The gap stays within 13 points even when every HM3D result of unknown version is assigned against them. They lead on GOAT-Bench, Instance ImageNav and HM3D ObjectNav v2, although one result of unknown version could reverse the last of these leads. The training-free agents that close the gap are mostly not LLM agents: nearly all training-free leaders are zero-shot modular pipelines, and two of the three that match or beat learned policies choose actions with frozen detectors and a semantic map, using a language model only offline, if at all. These patterns suggest that training-free agents gain most when perception recognizes the goal and least when a long instruction must be grounded. Training-free progress has nevertheless been substantial. Between 2024 and 2026, the R2R-CE gap narrowed from 39.5 to 28.5 points as its training-free leader rose from 25.5% to 48.8%. The HM3D ObjectNav v1 gap fell from 12.2 to 6.8 points over the same period (Figure 4b,c).

6 Conclusion

SurveyAtlas maintains evidence-linked reviews from field-specific configurations. In NavAtlas, protocol checks and targeted audits expose misleading headline comparisons and substantial task differences between learned and training-free methods. Top-ranked decisions match LLM judges' agreement with researchers at one twentieth of exhaustive judging's cost. Recording evaluation conditions alongside results makes the living review inspectable and its rankings interpretable.

Acknowledgments

Use of AI. The system uses Claude models for screening, reading, protocol checks, the head audit and Ask (§3). The evaluation uses Claude Opus and Qwen as judges. We used Claude Code to help write the system's code and draft and edit this paper, and Codex to edit its language and design Figure 2. The authors checked all claims, numbers and citations.

References

Sanyam Agarwal, Devi Parikh, Dhruv Batra, Peter Anderson, and Stefan Lee. Visual landmark selection for generating grounded and interpretable navigation instructions. In *CVPR Workshop on Deep Learning for Semantic Visual Navigation*, 2019.

Shubham Agarwal, Gaurav Sahu, Abhay Puri, Issam H. Laradji, Krishnamurthy DJ Dvijotham, Jason Stanley, Laurent Charlin, and Christopher Pal. LitLLM: A toolkit for scientific literature review. *arXiv preprint arXiv:2402.01788*, 2024.

Dong An, Hanqing Wang, Wenguan Wang, Zun Wang, Yan Huang, Keji He, and Liang Wang. ETPNav: Evolving topological planning for vision-language navigation in continuous environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(7):5130–5145, 2025.

Peter Anderson, Angel Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, and Amir R. Zamir. On evaluation of embodied navigation agents. *arXiv preprint arXiv:1807.06757*, 2018a.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3674–3683, 2018b.

Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'Arcy, David Wadden, Matt Latzke, Minyang Tian, Pan Ji, Shengyan Liu, Hao Tong, Bohao Wu, Yanyu Xiong, Luke S. Zettlemoyer, Graham Neubig, Daniel S. Weld, Doug Downey, Wen-Tau Yih, Pang Wei Koh, and Hannaneh Hajishirzi. OpenScholar: Synthesizing scientific

literature with retrieval-augmented LMs. *arXiv preprint arXiv:2411.14199*, 2024.

Erkan Aydin, Thomas G. Allen, Michele De Bastiani, Arsalan Razzaq, Lujia Xu, Esma Ugur, Jiang Liu, and Stefaan De Wolf. Pathways toward commercial perovskite/silicon tandem photovoltaics. *Science*, 383(6679):eadh3849, 2024.

Dhruv Batra, Aaron Gokaslan, Aniruddha Kembhavi, Oleksandr Maksymets, Roozbeh Mottaghi, Manolis Savva, Alexander Toshev, and Erik Wijmans. ObjectNav revisited: On evaluation of embodied agents navigating to objects. *arXiv preprint arXiv:2006.13171*, 2020.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D data in indoor environments. In *International Conference on 3D Vision (3DV)*, 2017.

Volkan Cirik, Yuan Zhang, and Jason Baldridge. Following formulaic map instructions in a street simulation environment. In *NeurIPS Workshop on Visually Grounded Interaction and Language (ViGIL)*, 2018.

Mostafa Dehghani, Yi Tay, Alexey Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. The benchmark lottery. *arXiv preprint arXiv:2107.07002*, 2021.

Matt Deitke, Winson Han, Alvaro Herrasti, Aniruddha Kembhavi, Eric Kolve, Roozbeh Mottaghi, Jordi Salvador, Dustin Schwenk, Eli VanderBilt, Matthew Wallingford, Luca Weihs, Mark Yatskar, and Ali Farhadi. RoboTHOR: An open simulation-to-real embodied AI platform. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. Show your work: Improved reporting of experimental results. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

Elicit. Elicit: The AI research assistant. Web application, 2023.

Julian H. Elliott, Tari Turner, Ornella Clavisi, James Thomas, Julian P. T. Higgins, Chris Mavergames, and Russell L. Gruen. Living systematic reviews: An emerging opportunity to narrow the evidence-practice gap. *PLoS Medicine*, 11(2): e1001603, 2014.

Matthew Feickert and Benjamin Nachman. A living review of machine learning for particle physics. *arXiv preprint arXiv:2102.02770*, 2021.

Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. Speaker-follower models for vision-and-language navigation. In *Advances in Neural Information Processing Systems*, 2018.

Peng Gao, Peng Wang, Feng Gao, Fei Wang, and Ruyue Yuan. Vision-language navigation with embodied intelligence: A survey. *arXiv preprint arXiv:2402.14304*, 2024.

Jing Gu, Eliana Stefani, Qi Wu, Jesse Thomason, and Xin Wang. Vision-and-language navigation: A survey of tasks, methods, and future directions. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 7606–7623, 2022.

Yichen He, Guanhua Huang, Peiyuan Feng, Yuan Lin, Yuchen Zhang, Hang Li, and Weinan E. PaSa: An LLM agent for comprehensive academic paper search. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

Yufang Hou, Charles Jochim, Martin Gleize, Francesca Bonin, and Debasis Ganguly. Identification of tasks, datasets, evaluation metrics, and numeric scores for scientific leaderboards construction. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

Chao-Chun Hsu, Erin Bransom, Jenna Sparks, Bailey Kuehl, Chenhao Tan, David Wadden, Lucy Lu Wang, and Aakanksha Naik. CHIME: LLM-assisted hierarchical organization of scientific studies for literature review support. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.

Yue Hu, Budhitama Subagdja, Ah-Hwee Tan, and Quanjun Yin. Vision-based topological mapping and navigation with self-organizing neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

InternVLA-N1 Team. InternVLA-N1: An open dual-system vision-language navigation foundation model with learned latent plans, 2025. Technical report, Shanghai AI Laboratory.

Sarthak Jain, Madeleine van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy. SciREX: A challenge dataset for document-level information extraction. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

Mohamad Yaser Jaradeh, Sören Auer, Manuel Prinz, Viktor Kovtun, Gábor Kismihók, and Markus Stocker. Open Research Knowledge Graph: Towards machine actionability in scholarly communication. In *International Conference on Knowledge Capture (K-CAP)*, 2019.

Salomon Kabongo, Jennifer D'Souza, and Sören Auer. ORKG-Leaderboards: A systematic workflow for mining leaderboards as a knowledge graph. *International Journal on Digital Libraries*, 2023.

Marcin Kardas, Piotr Czapla, Pontus Stenetorp, Sebastian Ruder, Sebastian Riedel, Ross Taylor, and Robert Stojnić. AxCell: Automatic extraction of results from machine learning papers. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

Mukul Khanna, Ram Ramrakhya, Gunjan Chhablani, Sriram Yenamandra, Theophile Gervet, Matthew Chang, Zsolt Kira, Devendra Singh Chaplot, Dhruv Batra, and Roozbeh Mottaghi. GOAT-Bench: A benchmark for multi-modal lifelong navigation. In *IEEE/CVF Conference on Computer*

Vision and Pattern Recognition (CVPR), pp. 16373–16383, 2024.

Rodney Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, Miles Crawford, Doug Downey, Jason Dunkelberger, Oren Etzioni, Rob Evans, Sergey Feldman, Joseph Gorney, David W. Graham, F.Q. Hu, Regan Huff, Daniel King, Sebastian Kohlmeier, Bailey Kuehl, Michael Langan, Daniel Lin, Haokun Liu, Kyle Lo, Jaron Lochner, Kelsey MacMillan, Tyler C. Murray, Christopher Newell, Smita Rao, Shaurya Rohatgi, Paul Sayre, Shannon Zejiang Shen, Amanpreet Singh, Luca Soldaini, Shivashankar Subramanian, Amber Tanaka, Alex D Wade, Linda M. Wagner, Lucy Lu Wang, Christopher Wilhelm, Caroline Wu, Jiangjiang Yang, Angele Zamarron, Madeleine van Zuylen, and Daniel S. Weld. The Semantic Scholar Open Data Platform. *arXiv preprint arXiv:2301.10140*, 2023.

Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *European Conference on Computer Vision (ECCV)*, pp. 104–120, 2020. ISBN 9783030586041.

Jacob Krantz, Stefan Lee, Jitendra Malik, Dhruv Batra, and Devendra Singh Chaplot. Instance-specific image goal navigation: Training embodied agents to find object instances. *arXiv preprint arXiv:2211.15876*, 2022.

Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-Across-Room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4392–4412, 2020.

Jakub Lála, Odhran O'Donoghue, Aleksandar Shtedritski, Sam Cox, Samuel G. Rodriques, and Andrew D. White. PaperQA: Retrieval-augmented generative agent for scientific research. *arXiv preprint arXiv:2312.07559*, 2023.

Hui Li and Wei Zhang. Perovskite tandem solar cells: From fundamentals to commercial deployment. *Chemical Reviews*, 120(18):9835–9950, 2020.

Xun Liang, Jiawei Yang, Yezhaohui Wang, Chen Tang, Zifan Zheng, Shichao Song, Zehao Lin, Yebin Yang, Simin Niu, Hanyu Wang, Bo Tang, Feiyu Xiong, Keming Mao, and Zhiyu Li. SurveyX: Academic survey automation via large language models. *arXiv preprint arXiv:2502.14776*, 2025.

Jianghao Lin, Rong Shan, Jiachen Zhu, Yunjia Xi, Yong Yu, and Weinan Zhang. Stop DDoS attacking the research community with AI-generated survey papers. In *Advances in Neural Information Processing Systems (Position Paper Track)*, 2025.

Jinzhou Lin, Han Gao, Xuxiang Feng, Rongtao Xu, Changwei Wang, Man Zhang, Li Guo, and Shibiao Xu. Advances in embodied navigation using large language models: A survey. *arXiv preprint arXiv:2311.00530*, 2023.

Iain J. Marshall and Byron C. Wallace. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. *Systematic Reviews*, 8:163, 2019.

Iain J. Marshall, Joël Kuiper, and Byron C. Wallace. RobotReviewer: Evaluation of a system for automatically assessing bias in clinical trials. *Journal of the American Medical Informatics Association*, 2016.

Iain J. Marshall, Benjamin E. Nye, Joël Kuiper, Anna Noel-Storr, Rachel Marshall, Rory Maclean, Frank Soboczenski, Ani Nenkova, James Thomas, and Byron C. Wallace. Trialstreamer: A living, automatically updated database of clinical trial reports. *Journal of the American Medical Informatics Association*, 2020.

Lina Mezghani, Sainbayar Sukhbaatar, Thibaut Lavril, Oleksandr Maksymets, Dhruv Batra, Piotr Bojanowski, and Karteek Alahari. Memory-augmented reinforcement learning for image-goal navigation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022.

Furkan Mumcu, Lokman Bekit, Michael J. Jones, Anoop Cherian, and Yasin Yilmaz. Agentic AI-empowered dynamic survey framework. *arXiv preprint arXiv:2602.04071*, 2026.

Benjamin Newman, Yoonjoo Lee, Aakanksha Naik, Pao Siangliulue, Raymond Fok, Juho Kim, Daniel S. Weld, Joseph Chee Chang, and Kyle Lo. ArxivDIGESTables: Synthesizing scientific literature into tables using language models. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

Tai-Long Nguyen, Do-Van Nguyen, and Thanh-Ha Le. Reinforcement learning based navigation with semantic knowledge of indoor environments. In *International Conference on Knowledge and Systems Engineering (KSE)*, 2019.

OpenAI. Introducing deep research. OpenAI blog, 2025.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems*, 2024.

Jason Priem, Heather A. Piwowar, and Richard Orr. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833*, 2022.

Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. REVERIE: Remote embodied visual referring expression in real indoor environments. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9979–9988, 2020.

Yanyuan Qiao, Wenqi Lyu, Hui Wang, Zixu Wang, Zerui Li, Yuan Zhang, Mingkui Tan, and Qi Wu. Open-Nav: Exploring zero-shot vision-and-language navigation in continuous environment with open-source LLMs. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6710–6717, 2025.

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

Santhosh K. Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X. Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-Matterport 3D dataset (HM3D): 1000 large-scale 3D environments for embodied AI. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2021.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009.

Furkan Şahinuç, Thy Thy Tran, Yulia Grishina, Yufang Hou, Bei Chen, and Iryna Gurevych. Efficient performance tracking: Leveraging large language models for automated construction of scientific leaderboards. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

Tim Schopf and Florian Matthes. NLP-KG: A system for exploratory search of scientific literature in natural language processing. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

Yijia Shao, Yucheng Jiang, Theodore A. Kanell, Peter Xu, Omar Khattab, and Monica S. Lam. Assisting in writing Wikipedia-like articles from scratch with large language models. In *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2024.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. ALFRED: A benchmark for interpreting grounded instructions for everyday tasks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Amanpreet Singh, Joseph Chee Chang, Chloe Anastasiades, Dany Haddad, Aakanksha Naik, Amber Tanaka, Angele Zamarron, Cecile Nguyen, Jena D. Hwang, Jason Dunkelberger, Matt Latzke, Smita Rao, Jaron Lochner, Rob Evans, Rodney Kinney, Daniel S. Weld, Doug Downey, and Sergey Feldman. Ai2 Scholar QA: Organized literature synthesis with attribution. *arXiv preprint arXiv:2504.10861*, 2025.

Shruti Singh, Shoaib Alam, Husain Malwat, and Mayank Singh. LEGOBench: Scientific leaderboard generation benchmark. *arXiv preprint arXiv:2401.06233*, 2024.

Michael Skarlinski, Sam Cox, Jon M. Laurent, James D. Braza, Michaela M. Hinks, Michael J. Hammerling, Manvitha Ponnapati, Samuel G. Rodriques, and Andrew D. White. Language agents achieve superhuman synthesis of scientific knowledge. *arXiv preprint arXiv:2409.13740*, 2024.

Henry J. Snaith, Antonio Abate, James M. Ball, Giles E. Eperon, Tomas Leijtens, Nakita K. Noel, Samuel D. Stranks, Jacob Tse-Wei Wang, Konrad Wojciechowski, and Wei Zhang. Anomalous hysteresis in perovskite solar cells. *The Journal of Physical Chemistry Letters*, 5(9):1511–1515, 2014.

Liuyi Wang, Jiagui Tang, Zongtao He, Ronghao Dang, Chengju Liu, and Qijun Chen. Enhanced language-guided robot navigation with panoramic semantic depth perception and cross-modal fusion. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024a.

Liuyi Wang, Kai Sheng, Zongtao He, Jinlong Li, Yongrui Qin, Haojie Dai, Xiangyi Wang, Jingwei Yang, Qingqing Yan, Chengju Liu, and Qijun Chen. A comprehensive survey and systematic real-world evaluation of embodied vision-and-language navigation. *IEEE Transactions on Automation Science and Engineering*, 2026. To appear. arXiv:2607.09792.

Shuai Wang, Harrisen Scells, Shengyao Zhuang, Martin Potthast, Bevan Koopman, and Guido Zuccon. Zero-shot generative large language models for systematic review screening automation. In *European Conference on Information Retrieval (ECIR)*, 2024b.

Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6622–6631, 2019.

Yidong Wang, Qi Guo, Wenjin Yao, Hongbo Zhang, Xin Zhang, Zhen Wu, Meishan Zhang, Xinyu Dai, Min Zhang, Qingsong Wen, Wei Ye, Shikun Zhang, and Yue Zhang. AutoSurvey: Large language models can automatically write surveys. In *Advances in Neural Information Processing Systems*, 2024c.

Wansen Wu, Tao Chang, Xinmeng Li, Quanjun Yin, and Yue Hu. Vision-language navigation: A survey and taxonomy. *Neural Computing and Applications*, 36(7):3291–3316, 2024.

Fei Xia, Amir R. Zamir, Zhiyang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson Env: Real-world perception for embodied agents. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Deshraj Yadav, Rishabh Jain, Harsh Agrawal, Prithvijit Chattopadhyay, Taranjeet Singh, Akash Jain, Shivkaran Singh, Stefan Lee, and Dhruv Batra. EvalAI: Towards better evaluation systems for AI agents. *arXiv preprint arXiv:1902.03570*, 2019.

Karmesh Yadav, Ram Ramrakhya, Santhosh Kumar Ramakrishnan, Theo Gervet, John Turner, Aaron Gokaslan, Noah Maestre, Angel Xuan Chang, Dhruv Batra, Manolis Savva, Alexander William Clegg, and Devendra Singh Chaplot. Habitat-Matterport 3D semantics dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4927–4936, 2023.

Xiangchao Yan, Shiyang Feng, Jiakang Yuan, Renqiu Xia, Bin Wang, Bo Zhang, and Lei Bai. SurveyForge: On the outline heuristics, memory-driven generation, and multi-dimensional evaluation for automated survey writing. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025.

Yan Yin, Zhiyu Chen, Gang Liu, Jiasong Yin, and Jianwei Guo. Autonomous navigation of mobile robots in unknown environments using off-policy reinforcement learning with curriculum learning. *Expert Systems with Applications*, 2024.

Naoki Yokoyama, Ram Ramrakhya, Abhishek Das, Dhruv Batra, and Sehoon Ha. HM3D-OVON: A dataset and benchmark

for open-vocabulary object goal navigation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5543–5550, 2024.

Wei Zhang, Yunfeng Zhang, Ning Liu, Kai Ren, and Pengfei Wang. IPAPRec: A promising tool for learning high-performance mapless navigation skills with deep reinforcement learning. *IEEE/ASME Transactions on Mechatronics*, 2021.

Yue Zhang, Ziqiao Ma, Jialu Li, Yanyuan Qiao, Zun Wang, Joyce Chai, Qi Wu, Mohit Bansal, and Parisa Kordjamshidi. Vision-and-language navigation today and tomorrow: A survey in the era of foundation models. *Transactions on Machine Learning Research*, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2023.

Fengda Zhu, Xiwen Liang, Yi Zhu, Qizhi Yu, Xiaojun Chang, and Xiaodan Liang. SOON: Scenario oriented object navigation with graph-based exploration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12684–12694, 2021a.

Yi Zhu, Yue Weng, Fengda Zhu, Xiaodan Liang, Qixiang Ye, Yutong Lu, and Jian-Bin Jiao. Self-motivated communication agent for real-world vision-dialog navigation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021b.

Appendix

Contents of the appendix. Appendices A to D describe the field file, search and screening, survey-reference coverage, and prompts. Appendices E to H cover human annotation, further evaluation details, robustness checks and Ask. Appendices I to K report costs and updates, describe other atlases and fields beyond AI, and walk through the website. A live copy of NavAtlas is available at https://huggingface.co/spaces/SurveyAtlas/navatlas. It is updated daily, so its counts exceed those of the snapshot reported here.

A A Field in One File

The NavAtlas field file has six parts (Figure 5). Search queries for arXiv and OpenAlex are paired with prefilter rules, including a pattern that relevant records must match and a list of allowed arXiv categories. The taxonomy defines scopes, tasks, settings, paradigms, contribution types and traits, each with a one-line definition supplied to the classifier. Plain-English labeling rules resolve borderline cases: drone navigation, for example, is labeled aerial even when it follows language instructions. Benchmark entries specify canonical names and aliases, incompatible versions sharing a name, commonly reused subsets, and a short standard protocol for each benchmark family. Landmark papers must appear in every build, guarding against silent retrieval failures. Website labels, including the display order and colors of paradigms, complete the file.

The file supplies the engine and website with all field-specific choices. The result-processing prompts and website metric columns nevertheless assume agent benchmarks with success metrics. To start an atlas, a user describes the field in a paragraph. Claude drafts the file, and the user reviews it before the first build. We used this process to build atlases of embodied AI (17,452 papers) and LLM agents (21,245 papers) without changing the engine.

B Finding the Papers

Tables 4 and 5 show how NavAtlas found its papers. Queries are deliberately broad: increasing recall is inexpensive at this stage, and later steps cannot recover papers that search misses. A fallback query such as "vision AND language AND navigation" retrieves many irrelevant records, but also finds papers that use unusual task names. Queries for specific benchmarks and method families find additional papers missed by broad searches.

Step	Records
Merged records (arXiv, OpenAlex)	**20,480**
dropped: no navigation term	3,676
dropped: weak OpenAlex match	5,237
dropped: before 2010	173
dropped: other arXiv categories	172
dropped: other types, no abstract	325
Labeled by the classifier	**10,897**
relabeled by Opus	1,810
out of scope	7,717
duplicates	33
Kept papers	**3,147**
core	2,212
adjacent	481
aerial	454
Read (core and adjacent)	**2,693**

Table 4 The NavAtlas funnel at the snapshot.

Records are merged across sources using arXiv identifiers, DOIs and normalized titles. A preprint and its proceedings version thus become one entry with publication metadata. Inexpensive rules remove records judged outside the field, and the classifier labels the remainder from titles and abstracts. Low-confidence labels go to Opus, which resolves most borderline cases. Venues are identified using Semantic Scholar, OpenAlex and arXiv comments. Publisher BibTeX is obtained through CrossRef when available.

The whole process is visible on the pipeline page (Figure 6). A researcher looking for a missing paper can inspect the queries, their yields and the reasons for exclusions, then add the paper or a query to the field file. Scheduled updates retrieve recent arXiv records and apply merging, prefiltering, screening and venue resolution. New in-scope papers are then read, and their rows pass through the result-processing steps.

C Coverage Details

Table 6 reports recall by survey. Coverage is weaker for adjacent references: household instruction-following papers, such as work on ALFRED (Shridhar et al., 2020), enter the atlas only if a navigation query retrieves them. Semantic Scholar stores some references as parsing fragments, such as "2019b. The regretful agent". We first matched references by arXiv identifier, DOI and normalized title, allowing fuzzy matches above

```
META = {"id": "navatlas", "title": "NavAtlas", "field": "Embodied Navigation", ...}
ARXIV_QUERIES = [                                  # 54 queries, written for recall
    ("vln-vl",   'abs:"vision-and-language navigation"'),
    ("vln-broad", "all:vision AND all:language AND all:navigation"),     # broad fallback
    ("objnav",   'abs:ObjectNav OR abs:"object navigation" OR abs:"object-goal navigation" ...'), ...]
OPENALEX_QUERIES = [("vln-vl", '"vision-and-language navigation"'), ...]          # 16 queries
RELEVANCE_RE = re.compile("navigat|objectnav|pointnav|imagenav|\\bvln\\b|habitat|object search|...")
SCOPES = [("core", "Core", "Embodied navigation is the paper's main subject (method, benchmark, ...)."),
          ("adjacent", "Adjacent", "Navigation is a substantial but secondary part: EQA, mobile manipulation, ..."),
          ("aerial", "Aerial", "UAV / drone / aerial vision-language navigation or tracking."),
          ("out", "Out of scope", "Web/GUI navigation, autonomous driving, geometric planning with no goal semantics, ...")]
PARADIGMS = [("specialist", "Task-specific learning", "learned", "Policy trained for the task from scratch ..."),
             ("fm-tuned", "Foundation model fine-tuned", "learned", "An LLM / VLM / VLA fine-tuned into the policy ..."),
             ("zs-modular", "Zero-shot modular", "training-free", "Frozen CLIP / detectors / LLM in a fixed pipeline ..."),
             ("agentic", "Agentic", "training-free", "The LLM/VLM owns the control flow ..."), ...]
BENCHMARKS = {"Instruction VLN": ["R2R", "R2R-CE", "RxR", "RxR-CE", ...], "Goal VLN": ["REVERIE", "SOON", ...], ...}
BENCH_PROTOCOLS = {"R2R-CE / RxR-CE / VLN-CE": "Habitat continuous environments: egocentric RGB-D, low-level actions "
    "(FORWARD 0.25 m, TURN 15 deg, STOP) or a learned waypoint predictor ... success within 3 m. Deviations: "
    "ground-truth waypoints, teleporting between nodes, pre-built maps of the test scenes, ...", ...}
BENCH_VARIANTS = {"HM3D-ObjNav": {"HM3D-ObjNav-v1": "HM3D-Semantics v0.1, Challenge 2022: 2,000 episodes, 20 scenes",
                                  "HM3D-ObjNav-v2": "HM3D-Semantics v0.2, Challenge 2023: 1,000 episodes, 36 scenes"}}
SUBSET_PROTOCOLS = {"R2R-CE": {"100-ep (Open-Nav)": "Open-Nav's 100 randomly sampled R2R-CE val-unseen episodes"}, ...}
CLASSIFY_RULES = """- scope: decide first. ... UAV / drone / aerial anything -> aerial (even if it is VLN). ..."""
LANDMARKS = [("1711.07280", "R2R / VLN"), ("1806.02724", "Speaker-Follower"), ("1811.10092", "RCM"), ...]
```

Figure 5 Abridged excerpt from the 519-line NavAtlas field file. The engine and website read this file as data to obtain the configuration for embodied navigation.

	Source	Query (abridged)	Hits	Core
most core papers	arXiv	navigation in title, CS	4,664	**1,431**
	arXiv	agentic and navigation*	4,273	946
	arXiv	vision, language and navigation	1,750	769
	OpenAlex	VLN and navigation	684	555
	arXiv	embodied and navigation	1,032	490
	OpenAlex	"vision-and-language nav."	517	432
	arXiv	VLN	502	415
	OpenAlex	ObjectNav phrases	553	384
fewest	arXiv	mobile manipulation phrases	38	3
	arXiv	pixel navigation phrases	2	1

Table 5 The eight highest-yielding and two lowest-yielding NavAtlas queries, out of 70, by number of core papers. Hits counts retrieved records, and Core counts those retained as core papers. Queries may overlap. *arXiv search appears to stem "agentic" to "agent", also returning work on agents in general.

	Survey	Core refs	In atlas	Recall	Cites
per survey	Wu et al. (2024)	94	90	95.7%	79 of 148
	Gu et al. (2022)	102	101	99.0%	84 of 176
	Lin et al. (2023)	39	37	94.9%	n/a
	Gao et al. (2024)	87	87	100.0%	80 of 317
	Zhang et al. (2024)	125	124	99.2%	114 of 364
	Wang et al. (2026)	202	200	99.0%	176 of 862
all	All, Opus	329	320	**97.3%**	
	All, Qwen	333	314	94.3%	
	Core for both	315	308	97.8%	

Table 6 Atlas recall on each survey's core references, using Opus labels unless otherwise stated. Core for both: references both judges label core. Cites: VLN papers cited by the survey out of those in the atlas predating its first arXiv version (Figure 1). Wu et al. (2024) first appeared on arXiv in 2021. Lin et al. (2023) covers navigation more broadly and is not compared with the VLN count.

0.92 similarity (Python's `difflib` ratio on normalized titles). We then manually searched for the 65 unmatched references that Opus labeled in scope, without blinding. Fifteen were already in the atlas. Most mismatches arose from fragments or title changes between versions, such as "CLIP on Wheels" becoming "CoWs on Pasture". Three references were not separate papers: two fragments of the citing survey's text and the 2026 survey's own preprint. We excluded these before computing coverage. Of the nine missing core references, seven were never retrieved: three papers from 2018 and 2019 (Cirik et al., 2018; Agarwal et al., 2019; Nguyen et al., 2019), two journal papers and one conference paper from 2021 and 2024 (Hu et al., 2021; Yin et al., 2024; Wang et al., 2024a), and the InternVLA-N1 technical report (InternVLA-N1 Team, 2025). The prefilter removed a vision-dialog navigation agent (Zhu et al., 2021b). The classifier excluded IPAPRec (Zhang et al., 2021), a LiDAR-based method, as navigation without a semantic goal to reach.

The major venues for Figure 1 are CVPR, ICCV, ECCV, NeurIPS, ICLR, ICML, ACL, EMNLP, NAACL, AAAI, IJCAI, CoRL, ICRA, IROS, RSS, TPAMI, IJCV, TMLR, RA-L and T-RO. Among VLN papers at these venues, the five surveys cite 73%, 67%, 36%, 49% and 36% of those available when they appeared.

Table 7 lists the 24 benchmark entries used in the subset analysis in §5. HM3D ObjectNav v1, v2 and results with no stated version are counted separately.

Without OpenEQA, ALFRED and unversioned HM3D ObjectNav, the subset shares are 42.7% for training-free

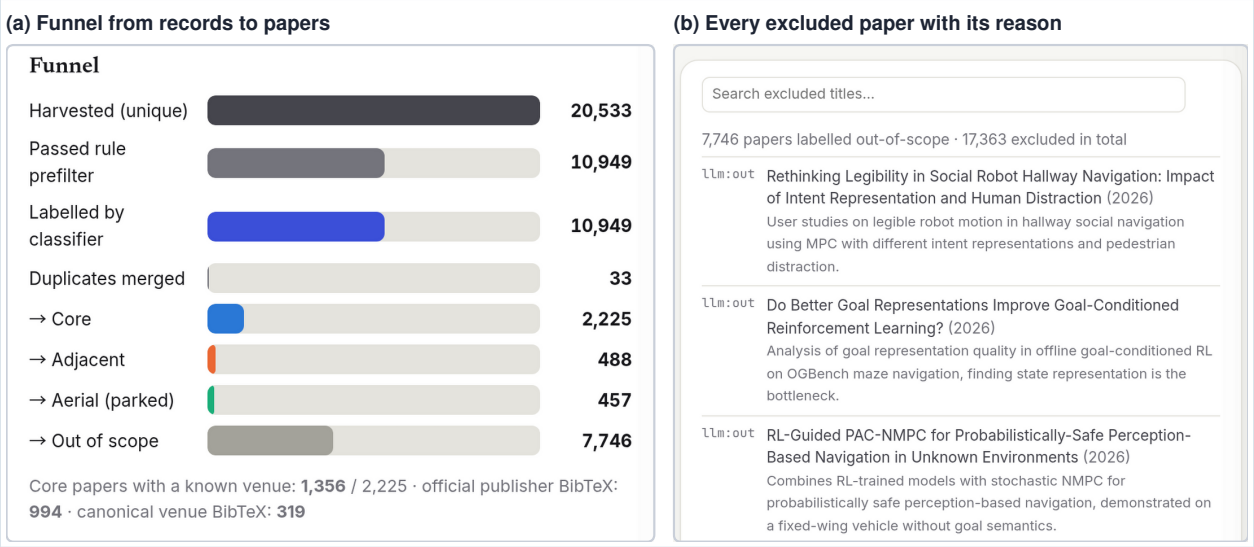


Figure 6 NavAtlas's pipeline page one day after the snapshot. Counts therefore differ slightly from Table 4. (a) Retrieved records and retained papers at each stage. (b) Searchable exclusions with the reasons given by the classifier or rules. The page also lists each query's yield, as in Table 5.

Task family	Benchmark entries	Count
Language-guided	R2R, R2R-CE, RxR, RxR-CE, REVERIE, SOON, CVDN, R4R	8
Object goals	HM3D ObjectNav v1, v2 and unversioned; MP3D, Gibson and RoboTHOR ObjectNav; HM3D-OVON	7
Multi-goal	GOAT-Bench	1
Image goals	Gibson ImageNav, Instance ImageNav	2
Point goals	Gibson PointNav, MP3D PointNav	2
Audio-visual	SoundSpaces	1
Outdoor	Touchdown	1
Household	ALFRED	1
Embodied QA	OpenEQA	1
Total		**24**

Table 7 Benchmark entries used to compare the frequency of subset results across learned and training-free methods. Unversioned HM3D ObjectNav results form a separate entry in this analysis and are excluded from the version-specific rankings.

results and 5.5% for learned ones.

Versions and costs. HM3D ObjectNav has two incompatible versions (Yadav et al., 2023). For 50 of the 172 papers reporting results on it, the resolver finds no version in the passages it reads. These results remain off both boards, although four, including three from training-free agents, exceed a group leader on at least one version. Cost reporting is even rarer. Many training-free agents query a large model several times per episode, yet a keyword search finds runtime, token usage or monetary cost among the results of only 7% of the 1,861 papers with results.

D Prompts and Judges

The system calls Claude through the Claude Code command line using the aliases `sonnet` and `opus`. During the build, these referred to Claude Sonnet 5.5 (`claude-sonnet-5-5`) and Claude Opus 5.5 (`claude-opus-5-5`). Build logs record only the aliases. A post-build command-line check returned these model identifiers. Search uses APIs and rules; the remaining stages use these Claude models, with deterministic code preparing inputs and checking outputs. Sonnet screens papers, and Opus relabels those with low-confidence labels. Reading processes each core and adjacent paper in a single model call, returning notes on the problem, key idea, ablation-supported insights, limitations and prior work, along with experimental settings and result rows. The judges are the same Opus model and Qwen3.8-27B. Qwen runs locally with 4-bit (Q4_K_M) weights, temperature 0, a 32k-token context and thinking disabled. All prompts are in the supplementary material, as part of the system's code and of the evaluation scripts. Below, we reproduce the system prompt for the protocol check introduced in §3.4 and describe the prompts of the judges.

Protocol check of the system. The check receives up to 60 numbered items of the form "benchmark, privileged information at test time, extra conditions" and returns a verdict and a reason of at most ten words for each. Its system prompt reads as follows, abridged and lightly

edited, with output-format instructions removed and the field's protocol notes inserted where marked.

> You audit results tables for a survey of {field}. For each reported result decide whether it was obtained under the benchmark's STANDARD evaluation protocol, i.e. directly comparable to other papers' numbers on that benchmark and split. Standard protocol notes per benchmark (information the protocol itself provides is NOT a deviation): {protocols}. Comparable: extra TRAINING data or pretraining of any kind (augmentation, synthetic data, web data, bigger models), different backbones, zero-shot use of foundation models, a single run of the agent. Not comparable: exploring or mapping the TEST environments before the episode (pre-exploration), beam search or picking among several trajectories, test-time feedback from humans, oracles or ground truth, privileged test-time information the protocol does not provide (GT semantic maps, GT object locations, GT goal position, oracle stop, oracle waypoints where the protocol has none), ensembles, test-time augmentation, a changed success radius, step budget, action space or sensor suite. IGNORE which episodes were evaluated (episode counts, scenes, subsets, splits): that is recorded separately. If the text says the condition was only used in training or only in an ablation that is not this row, it is comparable. When unsure, say comparable.

For the ablation of Table 2, the protocol notes are replaced by an instruction to use the model's own knowledge of each benchmark's standard protocol.

Scope. The judge receives the scope definition below and a numbered list of titles and abstracts. Opus processes 25 items per call for reference lists and 20 for screening samples, while Qwen processes 12. The prompt asks for a core, adjacent, aerial, out or unsure label, allowing the judge to use its knowledge of well-known papers. The definition is abridged here. The full prompt includes more examples of out-of-scope work.

> Field: embodied navigation research (2010 onward). Core: embodied navigation is the paper's main subject. An agent moves through a 3D environment (simulated or real) from egocentric perception towards a goal given as language, an object category, an image, a point, audio, dialogue, or a mix. Also counts: benchmarks, datasets, simulators, surveys and analyses built for such tasks, general visual-navigation foundation models, real-robot semantic or language navigation. Adjacent: navigation is a substantial but secondary part: embodied question answering, mobile manipulation or rearrangement, embodied instruction following, generalist embodied agents evaluated on navigation, 3D grounding built for navigation. Aerial: UAV or drone navigation. Out: anything else, such as general vision or language models, manipulation only, autonomous driving, web navigation, geometric planning or SLAM without goal semantics, legged locomotion.

Extraction. The judge receives the extracted row: method, benchmark, split, evaluation set, metric values and supporting quotation. It also receives the paper's first 1,500 characters and up to five 4,400-character passages around occurrences of the values. It checks whether all values belong to the stated method, benchmark, split and metric, whether the benchmark and split are correct, whether the method or variant is correct, and whether the row reports the paper's own result. A second pass omits the row's evaluation-set label. From passages around its values and up to six passages about the evaluation setup, the judge decides whether the result covers the full split or a subset, or whether this is unclear.

Protocol. The judge receives the result and a one-line standard protocol. For R2R-CE, this specifies Habitat VLN-CE, the full val-unseen split of 1,839 episodes, no ground-truth maps or segmentation, no pre-exploration and a 3 m success radius. It also receives passages around the result's values and up to six passages mentioning evaluation details, such as subsets, episodes, oracles, ground truth, pre-exploration or success criteria. The verdict is comparable, subset, non-standard or unclear. The system's protocol check, by comparison, receives only the extracted settings of up to 60 rows per call.

Answer grading. Claude Opus grades each answer against the gold answer from the atlas as correct, partial, wrong or abstain. For result questions, the number must match within 0.1. Leader questions require both the method and number to match. Protocol questions require the verdict and reason to match.

Keyword rules. Keyword rules provide a fallback for rows with neither a head-audit nor a protocol-check verdict. Table 2 evaluates these rules separately. A row is flagged when its recorded privileged inputs or setting notes match one of 14 patterns, namely beam (search), pre-exploration, ensemble, test-time augmentation, ground truth, GT, oracle, privileged, pre-built map, prior map, known map, teleport, full observability, and a panoramic action space in continuous environments. A negated match, as in "no oracle" or "without ground truth", does not count.

E Human Annotation

Annotators and procedure. Three embodied-navigation researchers, including one author, volunteered without payment. All three independently labeled the 64 navigation protocol rows in tasks C and D, without seeing any verdict

or one another's labels. Remaining items were divided among them, with one label per item. Other domain experts wrote the owner corrections and the field file's protocol notes. None of the annotators contributed to either. Tasks A, B, C and G were sampled randomly from items already labeled by the judges, allowing comparisons on the same items. Task D covers the rows supporting the findings in §5. The interface hid system and judge labels and gave no indication of why an item was sampled. Annotators could see these labels only after completing every item in a task. Labels were saved on entry and could be revised before task completion, for example to correct a misclick. We use the final labels, taking the majority for the 64 protocol rows and a single label for each remaining item. We report all disagreements with the system and judges without adjustment. Human labels serve as an independent check rather than a gold standard.

Interface. Each item shows the paper's title, the content to judge and the passages supplied to the judges. LaTeX tables are rendered as tables. Extracted values are highlighted in yellow and protocol terms in blue. An optional Chinese machine translation appears below the passages, with instructions to base decisions on the original text. Raw passages remain available. Each task begins with the instructions reproduced below.

Task A, screening (60 papers). Forty papers that the system kept and twenty that it excluded, shown by title and abstract. *Instruction:* does the paper belong to an atlas of embodied navigation? *core*: navigation is the main subject, an agent moving through a 3D environment from egocentric perception toward a goal given as language, an object category, an image, a point, audio or dialogue, including benchmarks, simulators, surveys and analyses for such tasks. *adjacent*: navigation is a substantial but secondary part. *aerial*: navigation by UAVs or drones. *out*: anything else.

Task B, extraction (40 rows). Random verified rows, each shown with its method, metric values and the passages around them. *Instruction:* are these values reported for this method under this benchmark, split and metric (yes, partly, no, unclear)? Are the benchmark and split right (yes, no, not stated)? Is it the right method or variant (yes, no, unclear)?

Task C, protocol (40 rows). Random rows from the top ten of the 12 boards of Table 3, each shown with the benchmark's standard protocol. *Instruction:* is this result comparable to standard entries on the leaderboard, with the full split, the standard protocol and no privileged information at test time? *subset*: only part of the episodes or a re-curated split. *nonstandard*: the full split but a different protocol, such as pre-built or pre-explored maps of the test scenes, ground-truth perception, oracle or human help, beam search, a changed success criterion, or a different simulator or action space. *unclear*: the passages do not decide it.

Task D, the rows behind the findings (24 rows). Every group leader of Table 3 and every highest main result of its 12 boards that task C did not already cover, shown in random order with the same instruction and passages as task C. Together, tasks C and D cover all 22 group leaders and every board's highest main result. They include the next result on R2R, REVERIE and Instance ImageNav, where the top rows are oracle variants or an error on our part.

Task G, robot manipulation (40 rows). Twenty top rows and twenty random rows from the 300 manipulation rows in Appendix F, with the judges' instruction for that field. *comparable*: the standard tasks and evaluation protocol, where training data, pretraining, the model and the number of training demonstrations may differ. *subset*: a subset of the tasks, suites or episodes. *nonstandard*: all tasks but a different evaluation, such as fewer or different trials, privileged state at test time, human or oracle help, training on the evaluation tasks against the protocol, or a different robot, simulator version or success criterion.

	Agreement with the annotators	System	Opus	Qwen
A screening	Same scope label (of 60)	**57**	52	50
	Cohen's κ	**0.92**	0.78	0.72
	Same in-scope decision	**58**	54	52
B rows	Values correct (of 40)	40	39	39
	Benchmark and split correct	40	35	35
	Method correct	40	40	39
C protocol	Flag F1 (40 rows)	**0.93**	**0.93**	0.82
	Same flag decision	37	37	32
	Cohen's κ, flag	0.84	**0.94**	0.53
D	Same flag decision (of 24)	**24**	22	22
G manip.	Flag F1 (40 rows)	0.78	0.84	**0.93**
	Same flag decision	36	**37**	34
	Cohen's κ, flag	0.71	0.79	**0.92**

Table 8 System and judge agreement with human annotators, as counts unless stated. Tasks C and D use the first annotator's labels. Agreement across all three annotators is reported in the text. Some judge verdicts are unclear, so κ uses only rows decided by both sides. Qwen judged 35 of the 40 manipulation rows.

Results. Table 8 gives the agreement on each task. In screening, the system agrees with the annotators more often than either judge. Annotators consider all 40 retained papers in scope, along with 2 of the 20 excluded papers.

The latter are a core paper removed by the rules and a general autonomous-navigation review excluded by the classifier but labeled adjacent by its annotator. For extraction, annotators find every sampled row correct. Each judge leaves the benchmark or split unconfirmed for five rows, mainly because the supplied passages do not name the split. In the protocol task, the first annotator flags 21 of 40 rows: 8 subsets and 13 non-standard results. The complete system and Opus each achieve F1 0.93 against these labels. Opus agrees with the first annotator much more closely than Qwen does, motivating its use as the reference judge in the main text. Table 2 compares flag variants with the majority labels described below.

For the 24 rows supporting the findings, the first annotator agrees with the system on all rows and with each judge on 22. This annotator considers all 22 group leaders comparable, including seven not confirmed by at least one judge, and flags the highest main result on exactly the 8 boards identified in §5. For robot manipulation, annotators flag 9 of 40 rows. The protocol check achieves F1 0.78 (95% interval 0.50 to 0.95), compared with 0.50 for manipulation rules, 0.53 for navigation rules and 0.43 for subset labels. Its gains over manipulation rules and subset labels remain positive even in this small sample (paired intervals 0.04 to 0.57 and 0.07 to 0.68). Both judges agree with annotators more closely than the check does. The check is thus useful in a new field, but does not yet match a strong judge.

Agreement between annotators. On the 64 protocol rows, agreement on flagging reaches Fleiss' κ 0.94 (0.89 for task C and 1.00 for task D). Pairwise Cohen's κ values are 0.90, 0.90 and 1.00. The first annotator differs from the other two on three task C rows, accepting each result that the others and the system flag. The third annotator leaves two rows undecided. Majority labels agree with the system on all 64 rows, with Opus on 55 of its 59 decided rows and with Qwen on 44 of 57. Against majority labels for the 40 task C rows, F1 is 0.50 for subset labels, 0.83 for keyword rules, 0.86 for the protocol check, 0.81 without protocol notes, 0.86 for the open-weight check and 1.00 for the complete system. The Opus and Qwen judges score 0.91 and 0.77, respectively, counting unclear predictions as no flag. On the 36 rows it decides, Opus scores 0.95.

F Evaluation Details

Table 9 summarizes the evaluations in §4 alongside human judgments. Table 10 reports all checks and sample sizes.

Both judges receive the same prompts (Appendix D).

Coverage. We obtained reference lists from Semantic Scholar for six human-written surveys: the five VLN surveys in Figure 1 and a survey of LLMs for embodied navigation (Lin et al., 2023). After excluding three references that are not separate papers (Appendix C), 746 distinct references remain. Opus labels 329 core, 67 adjacent, 20 aerial and 316 out of scope, leaving 14 undecided. The atlas contains 320 of the 329 core references (97.3%) and 369 of all 416 in-scope references (88.7%). Recall is 96.5% on the 256 core references not manually listed as landmarks in the field file, and 94.3% when core references are defined by Qwen. Seven of the nine missing core papers were never retrieved. Coverage is weaker for adjacent references (33 of 67, Appendix C).

	Check	Opus	Qwen	Human
complete	Core references of six surveys	97.3	94.3	–
	Kept core papers in scope	98.7	95.3	100
	Excluded papers in scope (↓)	5.0	9.0	10
correct	Rows with all values correct	99.3	96.6	100
	Benchmark and split correct	85.6	83.6	100
	Rows with every field correct	73.3	61.6	–
	Paradigm group correct	93.0	90.0	–

Table 9 NavAtlas evaluation in October 2026, in percent. Judge samples comprise 329 survey references, 150 core papers, 100 excluded papers, 146 rows and 100 method papers. Human samples comprise 40 retained papers (core or adjacent), 20 excluded papers and 40 rows (Appendices E and F).

Screening. The judges labeled samples from each screening outcome. Of 150 random core papers, Opus considers 148 in scope and Qwen 143. Of 100 classifier exclusions, Opus would retain 5 (2 aerial, 2 adjacent and 1 core), and Qwen 9. Extrapolating Opus's rate to all 7,717 exclusions suggests about 390 missed in-scope papers (95% Wilson interval 170 to 860), including about 80 core papers. Of 100 rule exclusions, Opus would retain none (95% upper bound 3.7%), while Qwen would retain 5, including a model of ant route following. Rerunning the classifier on 200 random papers reproduces 196 scope labels and 31 of the 34 core papers' paradigm labels. We also judged 100 random core papers from each of the embodied AI and LLM agent atlases using their field files' scope definitions. Reading was still underway in both atlases. Opus considers 96 of 98 decided embodied AI papers in scope, including 88 core papers. For LLM agents, it considers 98 of 100 in scope, including 83 core papers. Qwen's corresponding counts are 94 and 88 for embodied AI, and 96 and 87 for LLM agents, each out of 100. Other in-scope papers are labeled adjacent or applied. These results indicate precise screening, but we did not estimate recall for either atlas.

Paradigms. For 100 random core method papers in NavAtlas, judges received titles, abstracts and the field

	Check	n	Opus	Qwen
recall	Core survey references found	329, 333	97.3	94.3
	All in-scope references found	416, 420	88.7	86.0
screening	Kept core papers in scope	150	98.7	95.3
	Classifier exclusions in scope	100	5.0	9.0
	Rule drops in scope	100	0.0	5.0
extraction	All values correct	146	99.3	96.6
	Method and own result correct	146	100.0	97.3
	Benchmark and split correct	146	85.6	83.6
	Full split or subset correct*	138, 125	89.9	86.4
	Every field correct	146	73.3	61.6
top ten	System flags confirmed	71, 72	83.1	76.4
	Comparable rows accepted	42	97.6	78.6
	Agreement with the system	113, 114	88.5	77.2
random	System flags confirmed	34, 36	85.3	77.8
	Judge flags found	38	76.3	73.7
	Agreement with the system	95, 89	85.3	79.8

Table 10 NavAtlas checks against two reference judges, in percent: recall against six surveys' references, screening of random samples, extraction of 146 verified rows, and protocol flags on the top ten rows of 12 boards plus 100 random lower rows. Where decided sample sizes differ, n lists both. *Judged without the system's label. The judges agree on 86.2% of decided random rows. Of the 75 top-ten flags on the boards in Table 3, 2 concern unverified rows and 2 are undecided by Opus.

file's paradigm definitions. Opus agrees with the system's paradigm labels on 87 papers and Qwen on 84. Agreement on the broader group (learned, training-free or no method) is 93 and 90, respectively. Opus's disagreements most often concern the distinction between task-specific learning and large-scale pretraining, which leaves the group unchanged.

Extraction. The verbatim check establishes that a number occurs in the paper, but not that it belongs to the stated method, benchmark and split. After changing success rates by 0.1 to 5 points in 400 verified rows, 16 to 29% still occur elsewhere in their papers, with no trend by perturbation size. We therefore checked attribution directly. Each judge received 150 verified rows, one per paper, together with the title page and up to five passages around the values. Both judges could check 146 rows. Opus confirms all values in 145 rows and Qwen in 141. Both confirm the method and attribution to the paper's own results in at least 142 rows. They confirm the benchmark and split in 125 and 122 rows, respectively. Most remaining rows come from papers that omit the split or use their own episodes. In a second pass, judges assess full-split coverage from evaluation-setup passages without seeing the system's label. Opus agrees with the system on 124 of 138 decided rows and Qwen on 108 of 125. The system finds 13 of the 22 subsets identified by Opus and labels 5 full-split rows as subsets. Eight of these 14 disagreements fall outside the 24 common benchmarks in §5. Opus confirms every field in 107 rows and Qwen in 90. For Opus, most of the gap between fully correct rows and the 145 rows with correct values concerns the split. In 19 rows, the supplied passages leave the benchmark or split unconfirmed. Only 2 have an incorrect benchmark or split. Full-split labels account for most remaining discrepancies. Human annotators, reading the same passages, judge the benchmark and split correct in all 40 sampled rows (Appendix E).

Recall of extraction. The preceding checks cover only extracted rows. To estimate omissions, Opus read full texts from two sets of papers. For each paper and board pair, it identified the highest success rate reported for the paper's own method or variants and assessed whether it covered the full split under the standard protocol. The first set contains 40 random pairs of a paper and one of the 12 boards on which it has a result. One paper has no cached text. In another, Opus finds no result for the paper's own method. For 36 of the remaining 38 pairs, the atlas contains the paper's best own result. For the other two, it contains variants scoring 0.2 and 0.5 points lower. The second set contains all 74 pairs in which a read paper's screening labels name the benchmark of one of the 12 boards but the paper has no row on that board. For 21 pairs, the atlas contains the result under an unknown HM3D ObjectNav version or another split. Among the other 53, Opus finds own-method results in 16, including 7 it judges to use the full split and standard protocol. None exceeds its board's group leader. Two of these 7 are judge errors: a result on another Gibson benchmark and a result from cached text belonging to a different paper, which the owner had already excluded from the atlas. The audit cost $28 in Opus calls.

Protocol flags. We sampled the top ten rows by reported success rate (SR) from each of the 12 boards in Table 3 (120 rows), plus 100 random lower rows. From evaluation-setup excerpts, judges classified results as comparable, subset or non-standard. On decided top-ten rows, the system agrees with Opus on 88.5% and Qwen on 77.2%. Agreement between judges is 82.1%. On the 111 rows decided by all three, Cohen's κ is 0.78 between the system and Opus, 0.54 between the system and Qwen, and 0.65 between judges. Flags from the head audit or owner corrections are confirmed least often: 8 of 13 by Opus and 7 of 14 by Qwen. Often, the excerpts omit the relevant setting, such as a 30° turn angle. Of the random rows, 88 retain the protocol check's verdict. Opus confirms 29 of the 34 system flags it decides. The system detects 29 of 38 departures flagged by Opus. Of the 9 missed departures, 4 are subsets recorded as full splits, 2 are results assigned to the wrong benchmark or split, and 3 were missed by the

protocol check.

Robot manipulation. To test transfer, we sampled extracted results from 12 manipulation boards in the embodied AI atlas: the ten best papers per board and 180 random lower rows, totaling 300 rows. The protocol check received the engine's prompt and one line describing each benchmark's standard protocol, written as an owner would in the field file. Judges received the prompt in Appendix D, with categories adapted to manipulation. Training data, pretraining and the number of training demonstrations may differ, but task sets, evaluation trial counts, robots and privileged test-time inputs must match the standard protocol. Opus judged all 300 rows. Qwen's local server became unavailable after the first 236: all 120 top rows and 116 of the 180 random rows. We report Qwen's results on this subset. Opus flags 24% of decided rows, and agreement between judges is F1 0.72 (Table 11). Neither navigation rules nor rules written for manipulation add much to subset labels, because papers describe departures in varied terms. The protocol check improves F1 over subset labels by 0.17 against Opus and 0.15 against Qwen, with paired bootstrap intervals excluding zero. Protocol notes make no measurable difference. Manipulation rules were written after judging the first 120 rows but before the remaining 180. On those 180 rows alone, the check still leads: F1 against Opus is 0.77, compared with 0.56 for rules and 0.55 for subset labels. Three Opus-flagged rows report adversarial-attack success rather than policy success. These are extraction errors, not protocol departures.

Components and the extended head audit. Table 12 expands the comparison in Table 2 with additional variants, agreement scores, intervals and separate results for the two samples. F1 is lower on the random rows for every variant except the subset label and the judge prompt with Sonnet. The head audit increases agreement with Opus but decreases it with Qwen. Judge results therefore do not establish that the Opus-run audit improves flags overall. It does affect leaders, changing 5 of the 22 group leaders in Table 3. Each judge confirms 3 changes. We also ran an extended head audit on every group leader, starting without the owner's protocol and paradigm corrections. The prompt adds one question: is the decision model trained on navigation data? The audit reached 39 rows. Three runs ($3.97 total) returned identical verdicts on 35 rows. We use majority verdicts. It selects the leader in Table 3 for 18 of 22 groups and agrees with all four paradigm corrections and both step-budget corrections it encounters. Like the Opus judge, it accepts the English-only RxR-CE result and exhaustive REVERIE exploration that the owner excludes.

G Robustness of the Insights

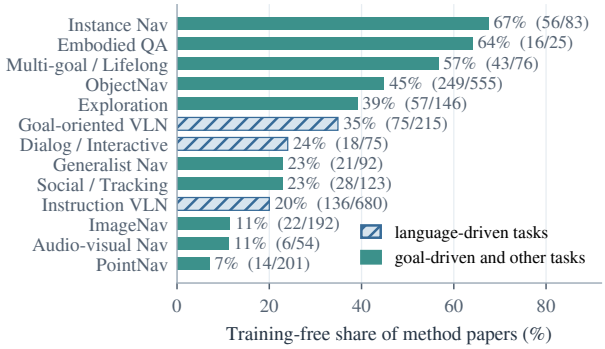


Figure 7 Share of training-free papers among the core method papers of each task with at least 25 such papers. Training-free work is common where a vision-language model can recognize the goal and rare on point, image-view and audio goals.

We test whether the findings in §5 depend on ambiguous decisions behind the published rankings in Table 3. The 12 boards cover 11 benchmarks, with two versions of HM3D ObjectNav: R2R (Anderson et al., 2018b), R2R-CE (Krantz et al., 2020), RxR-CE (Ku et al., 2020), REVERIE (Qi et al., 2020), SOON (Zhu et al., 2021a), HM3D ObjectNav (Ramakrishnan et al., 2021; Yadav et al., 2023), MP3D ObjNav (Chang et al., 2017; Batra et al., 2020), HM3D-OVON (Yokoyama et al., 2024), GOAT-Bench (Khanna et al., 2024), Gibson ImageNav (Xia et al., 2018; Mezghani et al., 2022) and Instance ImageNav (Krantz et al., 2022). They are the success-rate boards of the atlas's summary tables with more than 20 papers, plus the two image-goal boards with the most papers. The head audit covers the first ten, so the two ImageNav boards rely on the protocol check and owner corrections.

Table 13 reports both judges' verdicts on the 22 group leaders in Table 3, using the protocol prompt in Appendix D. Opus flags the training-free REVERIE leader as a subset result and the learned Instance ImageNav leader for using a robot different from the benchmark's specification. We also checked 21 runners-up, including two additional rows from the leading papers. Opus flags two as subsets: the training-free R2R runner-up (783 of 2,349 episodes) and a learned MP3D ObjNav result (1,394 of 2,195 episodes). It leaves a third undecided. Qwen flags four runners-up and cannot decide three.

Table 14 shows which protocol decisions determine each group leader. Table 12 reports agreement with judges at each step. The head audit changes 5 of the 22 group leaders in Table 3, with each judge confirming 3 changes. It also removes the only training-free SOON result.

Table 15 reports gaps under five sources of decisions. Without owner protocol and paradigm corrections, the

	Opus (291 rows)				Qwen (225 rows)			
Flags from	P	R	F1	Interval	P	R	F1	Interval
Subset label	.74	.46	.57	.46 to .67	.54	.50	.52	.38 to .64
Label + navigation rules	.74	.49	.59	.48 to .69	.54	.53	.53	.39 to .65
Label + manipulation rules	.73	.47	.57	.46 to .68	.54	.53	.53	.39 to .66
Label + check	.77	.71	.74	.65 to .82	.60	.78	.67	.56 to .78
no protocol notes	.73	.69	.71	.61 to .79	.60	.78	.67	.56 to .78
Other judge			.72				.72	

Table 11 Protocol flags on 12 manipulation boards in the embodied AI atlas: the ten best papers per board and 180 random lower rows. P, R, F1 and 95% intervals follow Table 12 and use each judge's decided rows. Navigation rules: 14 patterns from Appendix D. Manipulation rules: 15 patterns covering task subsets, fewer trials, privileged inputs, human help and test-time training, written after judging the first 120 rows but before the other 180. Paired bootstrap intervals for F1 differences (Opus, then Qwen) are as follows. Check minus subset label: .09 to .27 and .05 to .27. Check minus manipulation rules: .08 to .26 and .04 to .25. Check minus check without notes: −.01 to .09 and −.03 to .03.

		Opus (208 rows)						Qwen (203 rows)						
	Flags from	Agree	P	R	F1	Interval	Top	Random	Agree	P	R	F1	Interval	Top	Random
rules	Copy reported numbers	52.9	n/a	.00	.00	.00 to .00	.00	.00	49.8	n/a	.00	.00	.00 to .00	.00	.00
	Subset label	71.2	.93	.42	.58	.47 to .67	.57	.58	66.5	.87	.39	.54	.44 to .63	.51	.60
	Label + rules	78.8	.86	.66	.75	.67 to .81	.79	.69	77.3	.85	.67	.75	.67 to .81	.76	.72
SurveyAtlas	Label + check	85.6	.89	.80	.84	.78 to .89	.87	.78	80.3	.84	.75	.79	.73 to .85	.80	.78
	no protocol notes	85.1	.87	.81	.84	.77 to .89	.87	.77	81.3	.84	.77	.81	.74 to .86	.82	.78
	open-weight check	83.7	.88	.76	.81	.75 to .87	.86	.74	78.8	.84	.72	.77	.70 to .83	.79	.74
	Rules or check	84.1	.84	.83	.83	.77 to .89	.88	.76	80.8	.81	.80	.81	.75 to .86	.82	.79
	Check + head audit	87.5	.85	.90	.87	.82 to .92	.91	.81	77.8	.77	.80	.78	.72 to .84	.80	.76
	+ owner corrections	87.0	.84	.90	.87	.81 to .91	.90	.81	78.3	.77	.81	.79	.73 to .85	.81	.76
ref.	Judge prompt, Sonnet	92.3	.98	.86	.91	.87 to .95	.91	.91	84.7	.90	.78	.84	.78 to .89	.83	.85

Table 12 Full protocol-flag comparison. Copy reported numbers: no flags. Rules or check: flag if either component does. Agree: percent agreement after collapsing the judge's three-way verdict to flagged or unflagged. P, R, F1: flag precision, recall and F1. Interval: 95% F1 interval from 2,000 row-level bootstrap resamples. Top, Random: F1 on top-ten and random rows separately. Inter-judge F1 is 0.84 on 198 jointly decided rows. Paired bootstrap intervals for F1 differences (Opus, then Qwen) are as follows. Check minus rules: 0.03 to 0.16 and −0.02 to 0.11. Check minus check without notes: −0.05 to 0.05 and −0.07 to 0.03. Head audit minus check: −0.01 to 0.08 and −0.04 to 0.03. Check minus judge prompt with Sonnet: −0.13 to −0.02 and −0.10 to 0.01.

classifier places two papers among the training-free group leaders: a token cache on a trained vision-language-action model (VLA) for R2R-CE, and an LLM planner commanding R2R-trained followers on REVERIE and SOON. The RxR-CE leader becomes an English-only result, narrowly ahead of a token-pruning method on a trained VLA. The extended head audit adds a question about whether the decision model is trained on navigation data. A trained low-level controller that merely executes chosen waypoints does not count as a trained decision model. All three runs label the four relevant papers (two token caches, the token-pruning method and the LLM planner) as trained. They also reject the 1,000-step and 260-step R2R-CE results, matching the owner. The first run additionally rejects four rows accepted by both later runs: the English-only RxR-CE result, exhaustive REVERIE exploration, the atlas's training-free RxR-CE leader (40.7), and the learned MP3D ObjNav leader. Using only the first run would yield a RxR-CE gap of 44.2 points. For the judge columns, we proceeded down each group from its highest verified full-split row until the judge accepted a comparable result, judging previously unchecked rows as needed. Both judges accept a learned Instance ImageNav result flagged by the protocol check for an oracle-visibility success criterion. Both also accept a training-free HM3D-OVON result flagged by the head audit for macro-steps following an A* path.

Both judges also assessed the best row affected by each of the owner's seven protocol corrections. They flag both R2R results that learn from ground-truth episode outcomes and the 200-step R2R-CE result, although partly for different reasons from the owner. Opus rejects the 1,000-step result, while Qwen rejects exhaustive REVERIE exploration. Both accept the English-only RxR-CE result and the 260-step result with 30° turns. Neither prompt specifies the standard step budget of 500 or turn angle of 15°.

	Board	Group	SR	Opus	Qwen
instructions	R2R	learned	86.0	comp.	comp.
		training-free	49.5	comp.	unclear
	R2R-CE	learned	77.3	comp.	comp.
		training-free	48.8	comp.	comp.
	RxR-CE	learned	78.0	comp.	comp.
		training-free	40.7	unclear	comp.
	REVERIE	learned	60.4	comp.	comp.
		training-free	33.2	subset	unclear
	SOON	learned	50.9	comp.	comp.
object and image goals	HM3D ObjNav v1	learned	79.1	unclear	unclear
		training-free	72.3	comp.	comp.
	HM3D ObjNav v2	learned	83.7	comp.	comp.
		training-free	84.7	comp.	non-std.
	MP3D ObjNav	learned	64.0	comp.	unclear
		training-free	52.3	comp.	comp.
	HM3D-OVON	learned	66.3	comp.	comp.
		training-free	54.9	comp.	comp.
	GOAT-Bench	learned	62.7	comp.	comp.
		training-free	83.7	comp.	comp.
	Gibson ImageNav	learned	94.0	comp.	comp.
	Inst. ImageNav	learned	63.0	non-std.	comp.
		training-free	82.6	comp.	comp.

Table 13 Blind protocol judgments on the comparable group leaders in Table 3. Opus confirms 18 of 22 and Qwen 17. Both confirm 15. comp.: comparable, non-std.: non-standard.

Counts behind §5. Task-specific policies trained from scratch accounted for 95% of core method papers in 2018 and 29% in 2026. Over this period, fine-tuned foundation models grew from none to 23%. Training-free methods account for 67% of method papers on instance navigation and 20% on instruction-following VLN. The highest reported result is flagged on 10 of 12 boards. Of the 8 flagged main results, Opus confirms 7 departures, Qwen 6 and human annotators all 8. The largest gaps to comparable leaders are 21.6 points on MP3D ObjectNav and 19.2 on HM3D-OVON. Overall, 75 of 120 top-ten rows are flagged. Later papers mark the departure when copying a flagged result in 9 of the 11 cases we found. On R2R-CE, 30 of 41 training-free papers report only subset results, including 18 on Open-Nav's 100-episode subset. Nine of ten training-free group leaders are zero-shot modular pipelines. Two of the three that match or beat learned policies choose actions using frozen detectors and a semantic map.

H Question Answering

We evaluated Ask on template questions with gold answers derived from the atlas and on research questions written by the annotators.

Board	Group	Check	+Audit	+Owner
R2R-CE	learned	81.4	**77.3**	77.3
	training-free	60.8	60.8	**48.8**
RxR-CE	training-free	53.9	53.9	**40.7**
REVERIE	learned	61.1	61.1	**60.4**
	training-free	62.4	**33.2**	33.2
SOON	training-free	45.2	n/a	n/a
HM3D ObjNav v1	learned	86.6	**79.1**	79.1
	training-free	80.0	**72.3**	72.3
HM3D-OVON	training-free	63.0	**54.9**	54.9

Table 14 Group leaders (SR) affected by protocol decisions. Check: protocol check alone. +Audit: adds the head audit as run. +Owner: adds owner protocol corrections, as in Table 3. Owner paradigm corrections apply throughout. Table 15 also removes these. The other 14 group leaders in Table 3 do not change.

Template questions. Table 16 reports results for all question sets.

Each set contains 12 questions of each type. *Results* questions ask for the SR reported for a paper's full method on a standard board, using papers posted since June 2026. *Leaders* questions ask which method has the highest SR among a board's full-split, standard-protocol results. *Protocol* questions ask whether a result follows the standard protocol. They cover six results whose flags Opus confirmed and six judged comparable by both the system and Opus. The final set's 12 leader questions use boards absent from development questions. Nine final questions concern papers targeted during development, including four protocol questions about results used as gold answers to development leader questions. The retrieval baseline splits all core and adjacent full texts into 134,825 passages of about 1,300 characters. BM25 ranks the passages, and Sonnet receives the top 20 with instructions to answer only from them and cite paper identifiers. Retrieval is not restricted to the paper named in a question. The closed-book baseline uses the same Sonnet model, prompted to answer as an embodied-navigation expert and acknowledge when it does not know. Two of Ask's five errors on the final set involve protocol questions whose numbers differ from those recorded in the notes. One error is an incorrect protocol verdict. Another is a result answer that leads with the wrong value among several reported numbers. The remaining error answers a leader question using a RoboTHOR ObjectNav (Deitke et al., 2020) test board containing only one row. We did not manually verify the gold leaders. On Gibson PointNav, the gold leader is a training-data condition in a dataset paper rather than a method.

Research questions. Before any answers were generated, annotators wrote 15 questions in Chinese that they would ask as navigation researchers. Retrieval used English

	Board	Atlas	None	Audit$^+$	Opus	Qwen
instructions	R2R	36.5	36.5	36.5	36.5	38.0
	R2R-CE	28.5	14.2	28.5	24.0	20.6
	RxR-CE	37.3	24.1	24.1	24.1	24.1
	REVERIE	27.2	21.3	39.1	39.1	38.4
	SOON	n/a	27.8	n/a	n/a	n/a
goals	HM3D ObjNav v1	6.8	6.8	6.8	3.7	3.7
	HM3D ObjNav v2	−1.0	−1.0	−1.0	−1.0	−0.5
	MP3D ObjNav	11.7	11.7	11.7	11.7	1.5
	HM3D-OVON	11.4	11.4	11.4	3.3	3.7
	GOAT-Bench	−21.0	−21.0	−21.0	−21.0	−21.0
	Inst. ImageNav	−19.6	−19.6	−21.8	−5.0	−5.0

Table 15 SR of the best comparable learned result minus that of the best comparable training-free result, under five sources of decisions. Negative values: training-free agents lead. Atlas: Table 3. None: no protocol or paradigm correction by the owner, with the classifier's paradigms (scope corrections and removed papers cannot be undone from the snapshot). Audit$^+$: None plus the extended head audit on every group leader (majority of three runs). All columns use the verified rows of the snapshot and the group rule of §5. Opus, Qwen: the owner's paradigms and the judge's verdict on each verified full-split row that could lead a group.

Set	System	Results	Leaders	Protocol	All
Dev. 1	Ask, no fixes	10	7	7	24
Dev. 2	Ask, 2 of 3 fixes	10	6	9	25
Final	Ask, all fixes	**11**	**11**	**9**	**31**
	Full-text retrieval	4	0	6	10
	Closed book	0	0	0	0

Table 16 Correct answers out of 12 per type and 36 overall. Development sets identified faults in Ask. The final set was run once after fixes. Final-set 95% Wilson intervals: Ask 71 to 94%, retrieval 16 to 44%, closed book 0 to 10%.

renderings:

1. What are the main technical approaches to zero-shot VLN, and how well do they actually perform on R2R-CE?
2. How do agentic VLN methods differ from zero-shot VLN methods?
3. Since vision-and-language navigation began, which major shifts in technical approach has the field gone through, and what were they?
4. Compared with a serial observe-think-act loop, what do continuous perception, interruptible planning and asynchronous execution gain in embodied navigation? How do existing studies jointly evaluate success rate, reaction speed and compute cost in dynamic environments?
5. Does VLN performance improvement come from better spatial reasoning, or from stronger visual recognition, language priors and training data coverage? Which experiments separate these factors?
6. Do navigation models really understand spatial relations in instructions such as turn right after the second door or go to the chair behind the sofa? Which works rule out shortcuts from visual matching or dataset bias?
7. What does embodied navigation need to remember? In which tasks are raw images, language summaries, topological graphs, geometric maps and object memory useful, and what evidence shows a memory representation is irreplaceable?
8. What is the main bottleneck of long-horizon navigation failures: missed perception, localization drift, memory failure or planning errors? How do studies locate where an error starts and how it propagates?
9. When does long-term memory across tasks help navigation and when does it mislead decisions? How do methods handle environment changes, stale memory and conflicts between tasks?
10. How much value do explicit maps still have for VLM-based navigation? Under the same observations, model and compute budget, where do topological maps, metric maps and pure history context each have an advantage?
11. What should drive exploration in navigation: unknown regions (frontiers), the probability of the target, or uncertainty about the instruction? Which works study when each strategy applies and the trade-off between exploration gain and movement cost?
12. Can uncertainty in navigation be estimated reliably? Which methods distinguish not seeing clearly, not knowing where the agent is, not understanding the instruction and not identifying the target instance, and choose different actions accordingly?
13. Can a navigation agent recover from errors on its own without an oracle or reset? Which studies evaluate detecting a wrong turn, backtracking, relocalization and replanning?
14. What is the most reliable interface between high-level language planning and low-level motion control in navigation: subgoals, waypoints, pixel goals, trajectories or programmatic policies, and how do they affect execution success and replanning after failure?
15. If foundation models keep getting stronger, which problems in embodied navigation still need navigation research? What evidence shows that spatial memory, active perception, closed-loop error correction or long-term adaptation are not solved by model scale alone?

Only the first question names a benchmark. For the others, Ask uses retrieved notes and the field overview without adding board results. Sonnet answered each question under three conditions: Ask as deployed, full-text retrieval, and no supplied material. Full-text retrieval used the top 80 BM25 passages truncated to Ask's context length (about 33,000 characters). All conditions used the same instructions: cite only supplied material, identify weak evidence and answer in at most about 450 Chinese characters. Annotators saw each question's three answers in random order, with conditions hidden. They scored correctness, coverage, citation support and usefulness for deciding what to read or do next from 1 to 5, then selected the best and worst answers. Opus graded the same answers using the same rubric and the material supplied to each answer.

	Annotators						Opus	
Condition	Corr.	Cov.	Supp.	Use.	Best	Worst	Use.	Best
Ask	4.93	4.93	4.53	4.67	12	0	4.53	13
Full text	4.60	4.13	4.27	4.20	0	4	2.93	1
Closed book	4.47	4.20	4.07	4.07	3	11	3.47	1

Table 17 Blind grades of the answers to the annotators' 15 questions, as mean scores from 1 to 5 and as counts of best and worst answers.

Across the four criteria, Ask's mean score exceeds full-

text retrieval on 13 questions, with one tie and one loss. Its average gain is 0.47 points (paired bootstrap interval 0.30 to 0.62). Against closed-book answers, Ask wins on 11 questions, ties on two and loses on two, with an average gain of 0.57 (0.22 to 0.88). Opus gives lower scores for coverage and to retrieval-based answers, but ranks conditions in the same order. It prefers Ask on 13 questions and agrees with annotators on the best answer for 10 of 15. Closed-book answers, which can cite papers only by name and year, were rated worst on 11 questions.

I Cost and Freshness

	Stage	Items	$ per item	$ total
initial build	Screen, Sonnet	10,897	0.0035	37.8
	Relabel, Opus	1,810	0.0121	22.0
	Read full text, Sonnet	1,773	0.1634	289.8
	Read abstract, both models	549	0.0247	13.6
	Read full text, Opus	371	0.2897	107.5
	Protocol check, Sonnet	8,894	0.0008	7.4
	Head audit, Opus	292	0.0679	19.8
	Initial build			**497.8**
	First live update			**3.6**

Table 18 Costs at API list prices. Screening processes 40 papers per call, and the protocol check processes 60 rows. The first live update screened 52 records, followed by a reading run on 25 papers. Unit costs use replayed calls (3 screening batches of 40, 1 Opus batch of 30, 3 Opus readings and 1 protocol batch of 60) and that night's 25 Sonnet readings. For the head audit, we average first-call costs over 50 evaluation rows (39 in the extended audit and 11 in an earlier run without the decision-model question), then apply this mean to all 292 build-time audit calls. The extended audit's 78 repeated calls used the prompt cache and are excluded. Rounding prevents row totals from summing exactly. The atlas was built through a Claude subscription.

The pipeline requires no GPU. Table 18 reports costs by stage. A Sonnet reading takes a median of 22 seconds. With four parallel calls, 2,693 readings take about four hours of wall-clock time. The first live update added 23 papers, including 3 aerial papers that were not read. That night's reading run processed the other 20 new papers and 5 earlier ones. For the cost-reporting analysis in §5, a keyword search of metric names and setting notes finds time, token, energy or monetary information in 136 of 1,861 papers. A broader search allowing any occurrence of "time" or "cost", including "test-time", finds 252 (13.5%).

Reconstructing weekly updates. We reconstructed the eight weeks before the snapshot by assigning papers to the week of their first arXiv version (Table 19). The last week is incomplete: papers posted near its end entered the atlas only in the first live update. A complete week adds 11 to 26 papers and up to 120 results. Three boards received a new comparable leader in the last week alone. Across the 12 boards, comparable leaders changed 6 times in 2020, 6 in 2021, 15 in 2022, 8 in 2023, 17 in 2024, 16 in 2025 and 16 in the first nine months of 2026. Of the 32 leaders established in 2025 or 2026, 22 were replaced before the snapshot and 10 remained. Treating these 10 as censored, a Kaplan-Meier estimate gives a median tenure of 104 days, with 42% replaced within 90 days. For the 27 leaders established in 2020 to 2022, all subsequently replaced, the corresponding estimates are 209 days and 11%. The difference is not significant (log-rank $p = 0.12$). We therefore report change counts and the staleness of frozen reviews, without claiming shorter tenure. The reconstruction assumes papers enter the atlas in the week they are posted, consistent with the delay of one to five days measured in the first live update.

Live updates. Since the snapshot, NavAtlas has received live updates, initially triggered by hand and now scheduled each morning. Every run logs its steps, failures, retries, new papers, delays since posting, changed leaders and estimated cost. The first update to read its new papers automatically illustrates the time and cost involved. It retrieved 16 candidate records and retained 2 core papers posted one and three days earlier. Reading both LaTeX sources took 35 seconds. The protocol check then assessed 117 new combinations of benchmark and setting, its unit of judgment, and the head audit reread 10 new top rows. Including the website rebuild, the update took about 10 minutes. At the unit costs in Table 18, this totals about $1.2. Of this, about $0.4 is spent on the two new papers. Most checks and audits processed papers read in the preceding update. Even without new papers, a run completes pending work and costs a few tens of cents. The first scheduled run, for example, cost about $0.3. Failures so far have come from search APIs. When arXiv rate-limited one retrieval run, it stopped after that step. A new run an hour later completed the update.

J Other Atlases and Fields Beyond AI

The same engine built two further atlases, covering embodied AI and LLM agents (Table 20). Our three atlases cover AI fields because we can assess their papers ourselves and because shared benchmarks make protocol-aware extraction particularly relevant. To test transfer to a field with different publishing practices, we built a pilot atlas of perovskite tandem solar cells, an active area of materials science and photovoltaics published mainly in journals.

Week	Papers	Core	Results	$	New comparable leader
−8	19	18	47	3.14	**HM3D ObjNav v2**
−7	18	14	66	2.99	none
−6	11	8	39	1.83	none
−5	22	19	84	3.66	none
−4	14	12	50	2.33	none
−3	21	17	69	3.49	none
−2	26	21	120	4.34	none
−1	23	20	114	3.85	**GOAT-Bench, R2R-CE, RxR-CE**

Table 19 Reconstructed weekly updates for the eight weeks before the snapshot. Week −1 ends at the snapshot and is incomplete. Papers: core and adjacent papers first posted that week. Results: their atlas board rows. $: reading at $0.1634 per paper plus protocol checking at $0.0008 per row (Table 18), excluding screening at $0.0035 per record. The last column lists boards from Table 3 whose comparable leaders would change.

	Papers			Core/week	
Field	Records	Kept	Core	2025	2026
Embodied navigation	20,480	3,147	2,212	9	16
Embodied AI	29,331	17,452	12,682	60	127
LLM agents	36,162	21,245	13,635	69	223

Table 20 Three atlases built by the same engine, October 2026. Weekly counts of core papers use first arXiv versions.

The pilot. We wrote a 128-line field file covering perovskite/silicon, all-perovskite, perovskite/CIGS and perovskite/organic tandems. It defines three scopes, ten topics, four settings and five device-architecture paradigms (two-terminal, four-terminal, three-terminal, subcell only and no device). We disabled reading and ran the engine unchanged apart from one fix described below. In about 18 minutes, it retrieved 2,912 records, including only 125 from arXiv, retained 2,396 papers and labeled 2,153 core. These correspond to about nine new core papers per week in 2025 and eleven in 2026. Given the field file's scope definitions, Opus considers all 100 random core papers in scope (99 core). Of 99 classifier exclusions, it would retain 4, all adjacent. Initial recall is weaker: Opus identifies 177 core references in two widely cited reviews (Li & Zhang, 2020; Aydin et al., 2024). The atlas contains 106: 55% of the 2020 review's core references and 74% of the 2024 review's. Of 71 missing references, 66 were never retrieved. Most date from 2014 to 2019 and describe "semi-transparent perovskite cells for tandems" or "four-terminal" stacks using terms absent from our queries. The other 5 were retrieved but not retained. As in NavAtlas, whose queries underwent several revisions, these misses point to query coverage. Owners can inspect each query's yield on the pipeline page. Adding queries for these terms increased retrieval to 7,049 records and retention to 3,947 papers, without engine changes. Recall rose to 159 of 177 core references (90%). This estimate is optimistic because we wrote the queries after inspecting misses and did not reassess precision.

What carries over. Retrieval, merging, screening, venue resolution, BibTeX, the library, research map, timeline, digest and scheduled updates required no changes. The industry-author lookup failed on the first build because it expected pages that did not yet exist. We fixed this error.

What does not carry over yet. Reading requires openly available full texts. NavAtlas reads a fifth of its papers from abstracts. In the pilot, only 76 retained papers have arXiv versions, and most journals in the field are not open access. Most pilot papers would therefore be read from abstracts. The field also has the stated evaluation protocols needed for protocol-aware extraction. Solar-cell efficiency depends on certification by an accredited laboratory, the measured area and mask, and the direction and speed of the voltage scan (Snaith et al., 2014). These conditions could be recorded as standard protocols in a field file. However, the protocol check, head audit and website metric columns currently assume agent success rates, so applying them here would require engine changes.

Fields without benchmarks. Medicine already maintains living systematic reviews (Elliott et al., 2014), with risk-of-bias assessment serving as a counterpart to our protocol flags. Without shared benchmarks, protocol-aware extraction has little to assess. An atlas can still maintain a current library, a research map and cited answers.

K A Walk through the Living Review

This walkthrough follows a researcher new to embodied navigation through NavAtlas on the day after the snapshot, before that day's update. Numbered figure markers correspond to the numbers in the text and captions.

Getting the lay of the land. The research map (Figure 8) organizes papers across 15 tasks and six paradigms. Learned policies dominate instruction following (1), while object search has nearly as many zero-shot modular papers as task-specific ones (2). Some cells are empty, such as audio-visual navigation with fine-tuned foundation models (3). Clicking a cell opens its papers in the library. Readers can filter by task, paradigm, setting, benchmark, year, author or venue and export selections as BibTeX. The timeline shows annual paper counts by paradigm and each year's most cited papers.

Reading a paper. A paper page (Figure 9) begins with the venue and citation count (1), links to the paper, DOI, scholarly search engines and code search (2), a one-sentence summary (3), and taxonomy labels (4). Reading notes identify the model, source and date (5), then describe the problem, motivation, key idea, method and insights supported by the paper's ablations (6). The setting records information needed for comparison, including whether the decision model is trained, the backbone, training data and privileged inputs (7). A results summary (8) precedes individual result rows (9). Each row indicates whether all its values were found in the paper. Hovering over the check mark reveals their source passages (10).

Checking the state of the art. A board (Figure 10) ranks results for one benchmark and split (1). Readers can display all evaluation sets, full splits only or subsets only (2), separate benchmark-trained from zero-shot agents (3), hide non-standard and unverified rows (4), and filter by paradigm (5). Any metric can determine the ranking (6). Check marks identify rows whose values all occur in the paper (7), and tags mark industry authors (8). The filters shown retain 266 of 461 R2R-CE results. The comparable leader, with SR 77.3, comes from a paper posted three days before the snapshot. When all sets are shown, orange labels identify subset rows and name the subset on hover. Protocol flags show the reason for each departure.

Asking a question. Ask (Figure 11) accepts a plain-language question (1) and answers with numbered links to atlas entries (2). Because this question names a benchmark, Ask also receives its full-split rows and flags, from which it builds a table of learned leaders and gaps (3). The full answer explains why other candidate results were excluded. That discussion is omitted from the figure. Every question receives a shareable link.

Staying current. After scheduled updates, new papers are marked with a badge for ten days and can be sorted to the top of the library. A digest groups them by task. The pipeline page (Figure 6) shows how the atlas was built. Readers can return after each update to see new work, its claims and whether its results are comparable.

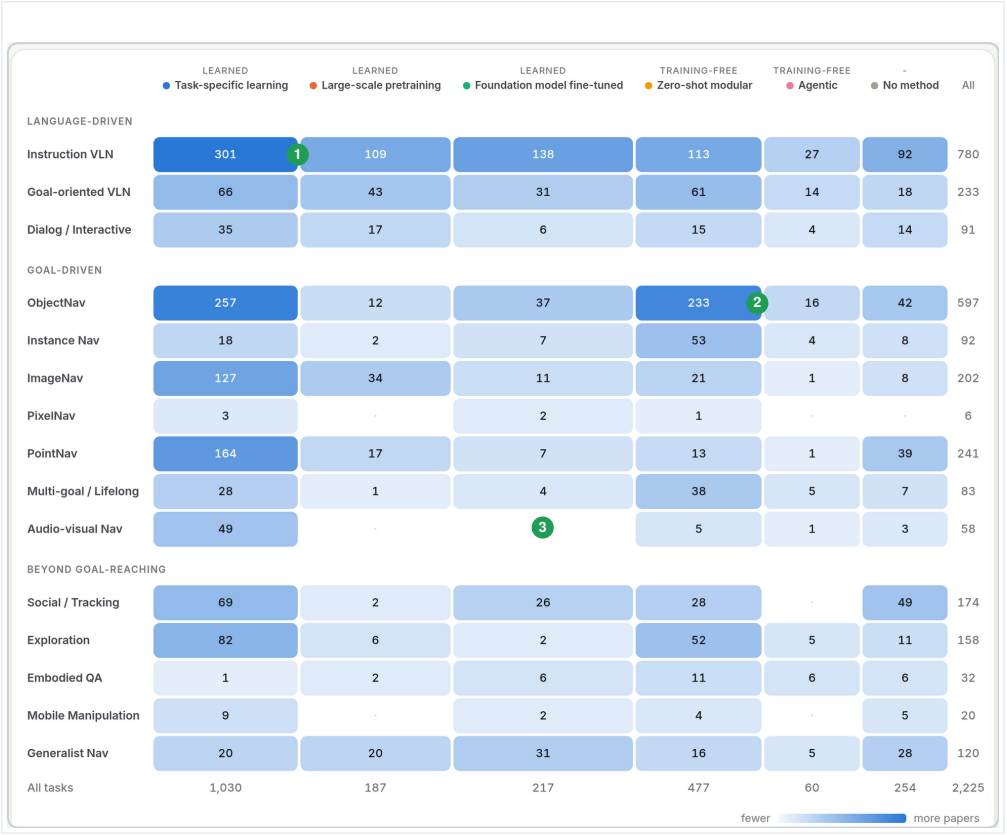


Figure 8 NavAtlas's research map in October 2026, captured after the first live updates. Counts slightly exceed the snapshot's. Cells show core papers by task and paradigm, with darker shading for larger counts. (1) Task-specific instruction following. (2) Zero-shot modular object navigation. (3) An empty cell: audio-visual navigation with fine-tuned foundation models.

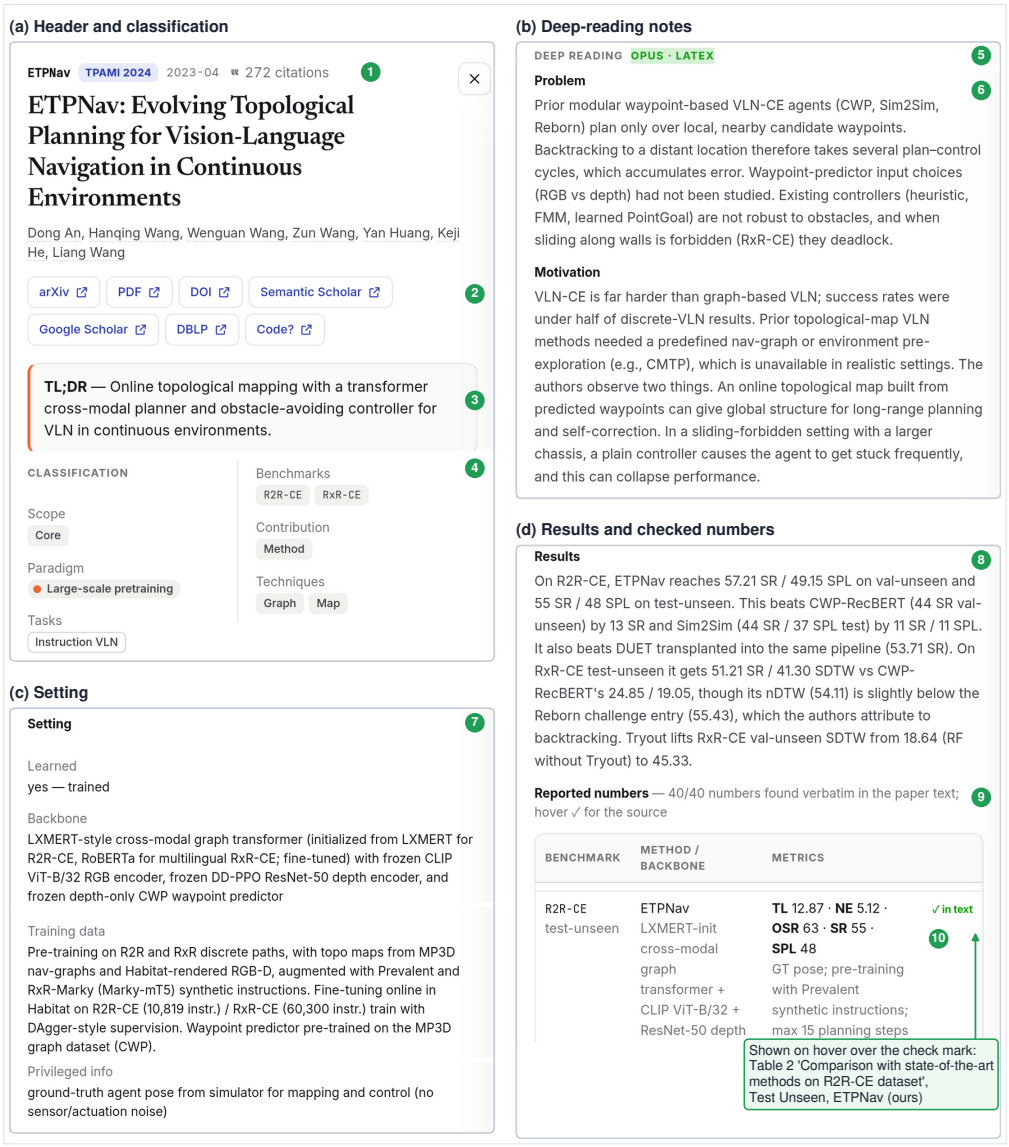


Figure 9 ETPNav's paper page (An et al., 2025), captured in a narrow window for legibility. The displayed year is the journal's early-access year. (1) Venue and citations. (2) Links. (3) Summary. (4) Classification. (5) Reading model, date and source. (6) Notes. (7) Setting. (8) Results summary. (9) Result rows and counts of values found in the text. (10) Verification check mark. Its tooltip is reproduced in a box because screenshots do not capture it. The abstract and some setting fields are omitted. The protocol check accepts ETPNav's use of simulator pose.

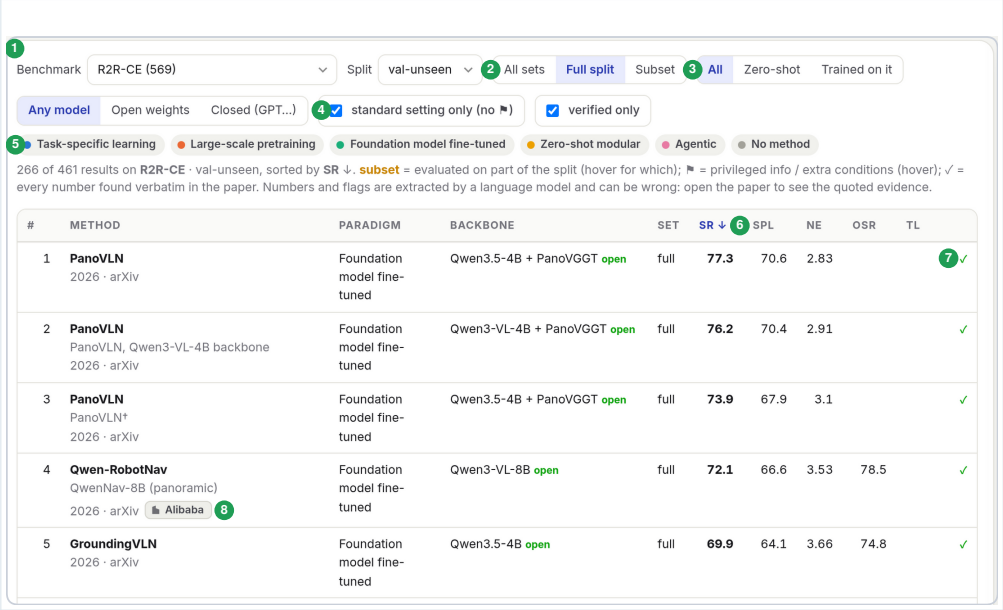


Figure 10 The R2R-CE val-unseen board with only full-split, standard and verified rows. (1) Benchmark and split. (2) Evaluation sets. (3) Training regime. (4) Filters for standard settings and verified rows. (5) Paradigm filter. (6) Sorting by any metric. (7) Check mark for numbers found in the paper. (8) Industry tag.

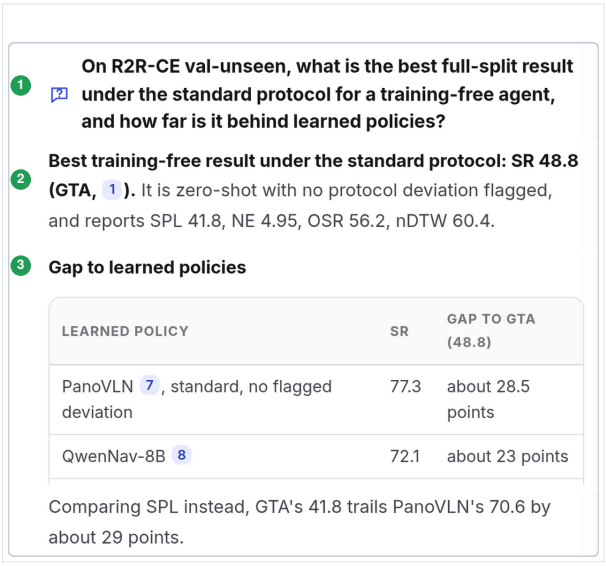


Figure 11 An Ask answer about R2R-CE, without its discussion of other candidates. (1) Question. (2) Cited answer. (3) Table built from the board.