Evaluating AI agents on the IUCN Red List assessor exam: Tradeoffs and opportunities for AI assistants in extinction risk assessments

Shane Weisz[1], Michael Dales[1], Sadiq Jaffer[1], Steven P. Bachman[2], Anil Madhavapeddy[1]

[1]*Department of Computer Science and Technology, University of Cambridge, Cambridge, United Kingdom*
[2]*Royal Botanic Gardens, Kew, Richmond, Surrey, United Kingdom*

Abstract

The IUCN Red List of Threatened Species is the world's most widely used system for assessing the global extinction risk of species, underpinning conservation policy and action worldwide. However, due to funding constraints and limited assessor capacity, the Red List faces significant gaps in species coverage (<20% of plant species have been assessed) and challenges in keeping assessments up to date (>30% are over 10 years old). Recent advances in AI open new opportunities for helping assessors increase assessment throughput and quality, but also raise concerns about trust and reliability, data security, cost, and environmental impact.

As a starting point for considering these tradeoffs, this paper evaluates the performance of AI agents on the official IUCN Red List assessor training exam. On a benchmark of 101 exam questions, we show that several frontier models from Anthropic and OpenAI reliably clear the exam's 75% pass mark. The top-performing model, Anthropic's Claude Opus 5, reaches 96% mean accuracy, increasing to 99% with majority vote over five attempts. Notably, the pass mark is now also reached by open-weight models running on local hardware: Alibaba's Qwen3.8 Flash, self-hosted on a single desktop Mac Studio, attains 81% mean accuracy (87% majority vote). This points to viable hosting alternatives where data security or sustainability concerns preclude sending data through external providers' data centres. Whilst the exam serves only as a capability floor, these results strengthen the case for thoughtful use of AI assistants in assessors' workflows, such as quality and consistency checks on draft assessments.

Keywords: IUCN Red List, extinction risk assessment, large language models, AI agents, sustainability of AI, local AI

1. Introduction

Amidst a global biodiversity crisis (Díaz et al., 2019), the IUCN Red List of Threatened Species (hereafter the Red List) is one of the world's most important resources for global species conservation (Rodrigues et al., 2006). It provides a standardised evidence-based system for classifying species into extinction risk categories, from Least Concern to Extinct, based on quantitative criteria (Mace et al., 2008). As a 'Barometer of Life', the Red List serves as a critical indicator of the health of the world's biodiversity, and guides policy and conservation action worldwide (Stuart et al., 2010).

As of the Red List update in July 2026, nearly 176,000 species have been assessed for the Red List (IUCN, 2026e). However, the Red List still faces significant data gaps, both taxonomic and geographic, as well as challenges in keeping existing assessments up to date (Bachman et al., 2019). Fewer than one in five described plant species have been assessed (IUCN, 2026e), whilst over 30% of assessed species have assessments older than the ten-year threshold at which the IUCN recommends reassessment (IUCN, 2026d,b).

The key constraint that prevents scaling up both new assessments and reassessments is limited expert assessor time – due to both lack of funding and limitations on the number of trained assessors (IUCN, 2026c; Bachman et al., 2019). Assessments are time- and knowledge-intensive, requiring assessors to synthesise often scattered literature and data, consult key experts, decide how to handle and represent uncertainty, and interpret available evidence against the quantitative Red List criteria.

As an indication of assessment time requirements, the global cacti assessment cited an average time of 6 hours per assessment (Goettsch et al., 2015). If we extrapolate, assessing the remaining 340,000+ described plant species (IUCN, 2026e) would require well over two million assessor-hours. Moreover, the more new species that are added to the Red List, the greater the resource demands will be to deliver reassessments at the IUCN's recommended ten-year cadence.

At the same time, recent AI capability advancements are opening new opportunities for knowledge work – particularly within an *agentic* paradigm, where agents use tools like document retrieval and code execution to extend beyond base parametric knowledge encoded in the underlying model weights. Agentic capabilities on expert knowledge work tasks are improving rapidly across domains, from finance to law to healthcare (Patwardhan et al., 2025; Kwa et al., 2026). Moreover, trends suggest AI is becoming cheaper for the same level of capability (Gundlach et al., 2026). With a step change in assessment throughput needed to meet the Red List's 2030 target[1],

Email addresses: sw984@cam.ac.uk (Shane Weisz), mwd24@cam.ac.uk (Michael Dales), sj514@cam.ac.uk (Sadiq Jaffer), s.bachman@kew.org (Steven P. Bachman), avsm2@cam.ac.uk (Anil Madhavapeddy)

[1]With a target of 260,000 species assessed by 2030 (IUCN, 2026c), and just

and keep pace with reassessment needs beyond it, thoughtful use of AI assistants could have strong potential to amplify limited expert assessor time and accelerate assessment throughput at scale – provided it can be shown to maintain (or even improve) the quality, trust and scientific rigour that underpins the Red List.

However, AI use also comes at a cost: financial costs (costly ChatGPT or Claude subscriptions, or per-token API charges), data security risks (data sent to and processed or retained by third-party providers), environmental impact (data centres' carbon emissions and water use concerns), and possible reputational risks (any errors or hallucinated outputs could degrade institutional trust). These costs and risks are particularly pertinent to a conservation nonprofit like the IUCN: limited budget for subscriptions or tokens, custodianship of sensitive species locality data, a commitment to environmental sustainability, and a trusted reputation to uphold as a global authority on nature.

The scale of both potential benefits and risks thus motivates the need for grounded appraisal of the use of AI assistants to support assessors in extinction risk assessment workflows.

1.1. Key framing: the IUCN Red List assessor exam

The official IUCN Red List Assessor Training exam (IUCN, 2026a) provides a useful starting point for exploring these capability-cost tradeoffs. Of course, achieving a high exam score would not imply that AI outputs could be trusted in real assessments without scrutiny by assessors – in practice, one of the Red List's key strengths is that assessors are species experts and take accountability for the judgment calls involved in reasoning over ambiguous, incomplete and uncertain data. Nonetheless, the ability of AI agents to reliably and accurately solve the exam questions would strengthen the case for their use as assistants to assessors in their Red List workflows, such as helping with quality and consistency checks. Moreover, the assessor training exam's deterministic grading mechanism is useful for removing subjectivity from evaluation.

1.2. Contributions

This work makes four key contributions:

1. **An agent evaluation framework comprising 101 questions from the official IUCN Red List assessor exam.** Using the official IUCN Red List Assessor Training exam, we construct a benchmark dataset for evaluating agentic capabilities to correctly apply the IUCN Red List Categories and Criteria. We outline our methodology and code implementation for constructing this evaluation framework, to support future work in domain-specific agentic evaluations.
2. **Frontier closed-source models consistently clear the exam pass mark threshold, but raise legitimate data security and cost concerns.** Using this evaluation framework, we show that several frontier agents (as of September 2026) consistently clear the exam's pass mark threshold of 75% – including Anthropic's Claude Opus 5, Opus 5.5 and Sonnet 5.5, and OpenAI's GPT-6 Sol and GPT-6 Luna. At the same time, we acknowledge use of such models can require costly subscriptions and carry data security concerns, along with rising public concern over the environmental impact of data centres.
3. **A self-hosted open-weight model now also meets the pass mark – offering possibilities for more data-secure and sustainable deployment alternatives.** We show that a self-hosted open-weight model, Alibaba's Qwen3.8 Flash (released in August 2026), can now also reliably clear the 75% threshold. This has important implications, as such models could in theory be self-hosted by IUCN Red List stakeholders on a renewables-powered server, with no per-token API charges and no assessment data leaving one's own infrastructure.
4. **Analysis of accuracy, cost and speed tradeoffs between agent configurations.** We highlight tradeoffs between accuracy, cost and speed across different agent configurations, and analyse the trends of how these have developed over the course of a single year.

The rest of this paper is structured as follows. In section 2, we introduce relevant background concepts around the IUCN Red List and the assessor-training exam, and related work. We then outline our methodology in section 3, and discuss our results and limitations in sections 4 and 5. Finally in section 6 we conclude and set directions for future work.

2. Background

2.1. The IUCN Red List of Threatened Species

Since its establishment in 1964, the IUCN Red List of Threatened Species has grown into one of the world's most comprehensive knowledge sources on the conservation status of plants and animals, building on contributions from thousands of species experts over the past six decades (Rodrigues et al., 2006; Akçakaya et al., 2026). The official Red List Categories and Criteria assessment framework for classifying species into data-driven threat categories has undergone several iterations, with the current version 3.1 (IUCN, 2012b) and its justification formally described by Mace et al. (2008).

Detailed documentation for the assessment framework is provided for assessors and reviewers in a series of official guideline documents, centred on the *IUCN Red List Categories and Criteria* (IUCN, 2012b) and its detailed 122-page companion *Guidelines for Using the IUCN Red List Categories and Criteria* (IUCN Standards and Petitions Committee, 2024).

2.2. IUCN Red List Assessor Training Course

To train Red List assessors, the IUCN provides an official online assessor training course: "Assessing Species' Extinction

under 176,000 assessed as of July 2026 (IUCN, 2026e), the remaining years to 2030 would require around 20,000 new species assessments per year, compared with an average rate of less than 10,000 per year over the past decade (IUCN, 2026f). The past-decade rate is calculated from IUCN Red List Table 1b, which shows that total species assessed on the Red List increased by 92,783 between 2015 and 2025 – from 79,837 at the end of 2015 to 172,620 at the end of 2025 – averaging 9,278 new assessments per year.

Risk using IUCN Red List Methodology" (IUCN, 2026a). The course culminates in a 2.5-hour, 25-question final exam, drawn from a larger question bank (see Section 3), to test an assessor's knowledge of how to apply the Red List categories and criteria to assess a species' extinction risk. The questions require complex domain knowledge, dynamic reasoning, and sometimes code or calculations. An illustrative question, modelled on the exam's generation length questions, is:

> The age of first reproduction for a species is 4 years and the oldest known reproductive individual bred at 15 years old. Gestation is long: a female gives birth to a single offspring once every two years, and fecundity is assumed to remain constant with age. Adults have an annual mortality rate of 25%. What is the generation length (in years) for this species?[2]

Ahead of considerations for AI assistants to support Red List workflows, such as consistency checks of draft assessments, the ability to reliably answer such questions is an important prerequisite.

2.3. Related work on LLMs for conservation

The ecology and conservation space has seen significant interest in leveraging large language models (LLMs) in recent years, with opportunities to apply AI for conservation captured in horizon scans by Reynolds et al. (2025) and Pollock et al. (2025). Initial research includes Iyer et al. (2025)'s work on LLM retrieval pipelines to ground evidence-based conservation decision-making, as well as Jaffer et al. (2025)'s AI-assisted system for maintaining living evidence databases to support systematic reviews in conservation science.

Research into LLM usage in the Red List domain in particular has been limited to two main works thus far: Dorm et al. (2026) benchmarked the parametric knowledge of LLMs for a variety of ecological tasks using Red List data, including range mapping and threat classification, whilst Uryu (2026) evaluated LLMs' ability to recall Red List fields, such as taxonomic classification and conservation status, on nearly 22,000 already-assessed species – again using only parametric knowledge. To our knowledge, no existing work in conservation has focused on *agentic* LLM paradigms in particular, with tool access and dynamic reasoning, as explored in this paper.

3. Methodology

3.1. Dataset construction

First, to construct a benchmark dataset of official IUCN Red List exam questions for our experiments, we requested the full set of exam questions and memos from the IUCN Red List Unit. These were shared as a list of 196 questions in Moodle GIFT file format. The full question bank included families of similar questions with only minor variations on the numerical parameters, so we reduce these to a single question per family.

The resulting dataset comprises a total of 101 questions and answers, from topics that include: understanding of key Red List terms (severe fragmentation, continuing decline, etc.), parameter calculations (population reduction, generation length), using supplied parameters to assign a Red List category, individual criteria-specific questions, as well as general questions around supporting information requirements, mapping standards, and regional assessments. The question types are split between category assignment (n = 45), multiple-select (n = 33), matching (n = 2), and parameter calculations (n = 21). The full breakdown of questions by Red List topic is listed in Appendix A (see Table A.2).

3.2. Evaluation framework

Next, to create our evaluation framework, we use Harbor Framework – an open-source framework that allows AI agents to be evaluated on tasks in configurable sandboxed container environments (Harbor Framework Team, 2026). Harbor Framework was released by the developers of Terminal-Bench, one of the most prominent benchmarks for evaluating the capabilities of coding agents (Merrill et al., 2026).

Using this framework, each of the 101 Red List exam questions is mapped to a corresponding Harbor *task*, where an attempt by a single agent at a task is referred to as a *trial*. In a given trial, the agent receives the corresponding exam question, along with an instruction on the required format for the answer, and any additional instructions. The agent then reasons inside an isolated Docker container, and must write its final answer to a file, which is then graded against the exam's official accepted answer keys.

To receive full credit, the answers must match exactly, ignoring case and whitespace. No partial credit is given for partially correct answers or answers with some correct options for multi-select cases.

Each task can be run with one of two conditions: (1) guidelines available, and (2) closed-book. For the guidelines condition, five official IUCN Red List assessment guideline documents, totalling 308 pages, are mounted read-only as markdown files in the container for the agent to read and search via tool calls:

1. *IUCN Red List Categories and Criteria*, 38 pages (IUCN, 2012b),
2. *Guidelines for Using the IUCN Red List Categories and Criteria*, 122 pages (IUCN Standards and Petitions Committee, 2024),
3. *Mapping Standards and Data Quality for IUCN Red List Spatial Data*, 32 pages (IUCN SSC Red List Technical Working Group, 2024),
4. *Supporting Information Guidelines*, 68 pages (IUCN, 2013), and

[2] The expected answer is 8 years. According to Section 4.4 of the Guidelines (IUCN Standards and Petitions Committee, 2024), when "annual mortality after the age of first reproduction is well known, and mortality and fecundity do not change after the age of first reproduction (i.e., there is no senescence)", generation length can be calculated as 1/(adult mortality) + age of first reproduction, i.e. $1/0.25 + 4 = 8$. In this case, the oldest breeder's age and the breeding frequency are distractors and not needed.

5. *Guidelines for Application of IUCN Red List Criteria at Regional and National Levels*, 46 pages (IUCN, 2012a).

In the closed-book condition, no documents are mounted – the agent must answer using its parametric knowledge of the guidelines only. In both conditions the agents' web search and fetch tools are disabled.

3.3. Agent configurations

Using this evaluation framework, we evaluate 10 models from 3 families (Anthropic's Claude, OpenAI's GPT, and Alibaba's Qwen), all released between September 2025 and September 2026. Our primary comparison comprises the latest release from each family as of September 2026: Anthropic's Claude Opus 5.5 and Claude Sonnet 5.5, OpenAI's GPT-6 Sol and GPT-6 Luna, and Alibaba's open-weight Qwen3.8 Flash (a self-hostable 125-billion-parameter mixture-of-experts model with 6 billion active parameters).

To assess change across model generations, we also evaluate the previous generation releases of Anthropic's models (Claude Opus 4.5 and 5, and Sonnet 4.5 and 5), and one previous self-hostable Qwen mixture-of-experts variant (Qwen3.6-35B-A3B).

To run the Anthropic models, we use Claude Code as the agent harness, accessed through a $100 monthly Claude Max subscription. For the OpenAI models we use Codex CLI via a $20 monthly ChatGPT Plus subscription. For the open-weight Qwen models we use the OpenCode agent harness, and serve them via Ollama at 4-bit precision[3], running on a single local Apple M2 Ultra Mac Studio with 192 GB unified memory.

Reasoning effort was set to high for the Anthropic and OpenAI models, and to low for Qwen3.8 Flash (following recommendations by Willison (2026), consistent with our pilot runs).

3.4. Evaluation metrics

For each model configuration, we run five independent sweeps over the suite of 101 exam questions. For each trial, Harbor records the full trajectory of messages, tool use, timings and costs. Using these recorded trajectories, we compute a series of evaluation metrics across the pool of 101 exam questions: accuracy metrics (mean accuracy, pass@5, pass^5, maj@5), speed (seconds per question, tokens per second), token use (input, output, cache-reads, cache-writes), and cost (dollars per question). We define these metrics as follows.

Introduced by Chen et al. (2021), pass@k is the percentage of questions where the agents produced the correct answer on *at least one* of k attempts. In contrast, pass^k (Yao et al., 2025) is the percentage of questions where the agent got *all* of the k attempts correct. maj@k is computed as the percentage of questions answered correctly by the majority-vote answer across the k attempts, with tied votes scored as the share of tied attempts that are correct.

[3]Qwen3.8 Flash in NVFP4 format via Ollama's MLX backend (`https://ollama.com/library/qwen3.8-flash-next:125b-mlx`); Qwen3.6-35B-A3B in Q4_K_M format (`https://ollama.com/library/qwen3.6:35b-a3b-q4_K_M`).

Alongside mean accuracy μ over the five runs, we report 95% confidence intervals as $\mu \pm 1.96s$, where s is the run-to-run standard error, following the agent-benchmarking practices set by Huang et al. (2026) and Merrill et al. (2026). The standard error is computed as $s = \sigma/\sqrt{5}$, where σ is the standard deviation of the mean accuracies over the five repeat runs.

For cost reporting, since we use a subscription for the Anthropic and OpenAI models, we compute what the equivalent token use would cost on their metered API, by applying each model's published per-token prices (as of September 2026) to the recorded token counts for each attempt (separately applying the respective rates for input tokens, output tokens, cache-reads and cache-writes). We note that the self-hosted models running on the local Mac Studio do not have a per-token price – after the fixed cost of the hardware itself, the only variable cost for self-hosted models is the electricity use of the server.

3.5. Experimental design

Using the above configurations, our experiments investigate the following effects:

1. **Open-weight vs closed-source agents.** How do self-hosted open-weight Qwen models compare to the frontier closed models from OpenAI and Anthropic? In other words, can self-hosting offer a data-secure alternative for conservation deployments?
2. **Model size.** What effect does using larger and more expensive models (Claude Opus 5.5, GPT-6 Sol) have on accuracy and cost, compared to their smaller and cheaper alternatives (Claude Sonnet 5.5, GPT-6 Luna)?
3. **Model generations.** How have accuracy, cost and speed changed with a model family's subsequent generations between September 2025 and September 2026?
4. **Knowledge source.** How does performance differ when models rely on parametric knowledge only (faster and cheaper), compared to agentically searching over the official Red List guideline documents?
5. **Answer consistency.** How much do answers differ between repeated attempts for a given agent, and what effect does the unanimity of answers have on accuracy? For questions where repeat attempt answers differ, how much does accuracy improve by taking majority vote?

4. Results and Discussion

We now present the results from running the above experiments and discuss key findings. For the full results for each agent configuration across the full suite of 101 Red List exam questions, see Table A.1 in Appendix A.

4.1. Closed models set the frontier; self-hosted open models present a data-secure alternative

In Figure 1, we see that Anthropic and OpenAI's frontier proprietary models (as of September 2026), GPT-6 Sol and Claude Opus 5.5, attain mean accuracies of 91% and 88% across the five runs respectively in the guidelines condition – consistently

clearing the 75% threshold that serves as the exam pass mark required for assessor certification. Anthropic's Claude Opus 5 (July 2026) scores even higher, reaching 96% mean accuracy with guidelines access.[4] This rises even further to 99% with majority vote over five attempts, and 100% pass@5 (every question answered correctly at least once).

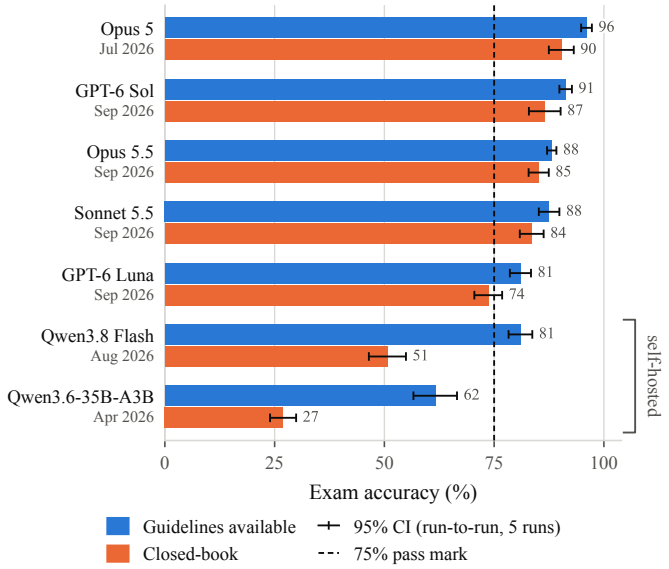


Figure 1: Mean accuracy for each model, averaged over five independent attempts on each of the 101 exam questions – closed-book (red) and with the official IUCN guidelines documents available (blue). The whiskers are 95% run-to-run confidence intervals (±1.96 standard errors, across the five repeat runs of the full suite). The dashed line reflects the 75% exam pass mark required for assessor certification.

The respective cheaper variants, Claude Sonnet 5.5 and GPT-6 Luna, also clear the pass mark with mean accuracies of 88% and 81% with guidelines respectively. Sonnet 5.5 matches Opus 5.5 almost exactly (both 88%) at around 30% lower cost per question, and whilst GPT-6 Luna's mean accuracy is 10 percentage points lower than GPT-6 Sol, it is over 20 times cheaper (see Figure 2).

At this level, many of the remaining errors reflect understandable alternative interpretations of ambiguous cases (such as the border between Near Threatened and Least Concern, which has no formal quantitative thresholds), or strict marking where there is a justifiable case for adjusting the answer key, rather than genuine capability limitations. For more details on the full breakdown of errors by topic and question type, see Table A.2 and Table A.3 respectively.

Most notably, we see that the self-hosted Qwen3.8 Flash agent attains a mean accuracy of 81% with guidelines access. This reflects rapid progress, with an improvement of nearly 20 percentage points over Qwen3.6-35B-A3B, Alibaba's previous generation's self-hostable mixture-of-experts model released just four months prior. Qwen3.8 Flash matches GPT-6 Luna's accuracy, and surpasses it with majority vote (87.3% against 84.2%), whilst running entirely on local hardware.

Local deployment is particularly relevant due to the data security advantages it offers when working with sensitive species data. Red List assessments often use precise locality data for threatened species, which can be important to handle with care: unrestricted access to species locality information has been shown to be exploited by poachers, in some cases within months of a species being described (Lindenmayer and Scheele, 2017). Using a cloud-hosted model from an external AI provider to process such data can add a degree of risk, as it requires trusting the provider's terms on retention and use – whereas self-hosting keeps the data on infrastructure within the assessors' own control. Similarly, self-hosting could enable sustainability reassurances through control over the energy supply for the deployment, rather than relying on data centres controlled by Anthropic or OpenAI (although it of course does not negate the energy cost of training the model itself).

4.2. Guidelines access improves accuracy, especially for the self-hosted open models

Another clear difference between the open and closed models is shown in the importance of guidelines access. For Qwen3.8 Flash, mean accuracy increases by 30 percentage points from 51% to 81% when guidelines access is allowed, whereas having guidelines available only improves mean accuracy by 3 to 7 percentage points (Opus 5.5 and GPT-6 Luna respectively) for the current-generation closed models. This suggests that the proprietary large models have stronger internal parametric knowledge of the Red List guidelines. The weaker parametric knowledge of Qwen3.8 Flash means the model relies extensively on repeated tool calls, averaging 12 calls per attempt (searching and reading the guidelines), compared to fewer than two per attempt on average for Opus 5.5 and Sonnet 5.5. See Table A.4 for the full breakdown of tool usage by model.

We also note that even if access to the guidelines yields only minor accuracy improvements from the proprietary models, use of and citation to the official guidelines is nonetheless important for trust – since it allows assessors to verify for themselves the sources used for claims, by tracing the agent's tool calls that accessed the guidelines.

4.3. No free lunch: self-hosting has speed tradeoffs

Figure 2 highlights the tradeoffs between accuracy, cost and speed between the closed models and the self-hosted open Qwen3.8 Flash.

We can see that although Qwen3.8 Flash has approached the frontier on accuracy (matching GPT-6 Luna with 81%) whilst running on self-hosted infrastructure with no per-token API cost, this comes at a speed sacrifice – the extensive tool calls and back-and-forth reasoning result in a mean time per question of just under 10 minutes (561 seconds). For workflows where speed matters, GPT-6 Luna may thus be a more viable candidate in terms of scalability, answering around 15 times faster on

[4]Opus 5.5's lower score is largely an artefact of overly strict marking: on 28 attempts across six Criterion B questions asking for category only, it gave the correct category with a full code including additional subcriteria, e.g. CR B1ab(ii,iii) where the key accepts only CR or CR B1ab(iii) – arguably defensible, interpreting a species disappearing from sites as suggesting decline in area of occupancy. Counting these as correct, Opus 5.5 would score 94%, just below Opus 5.

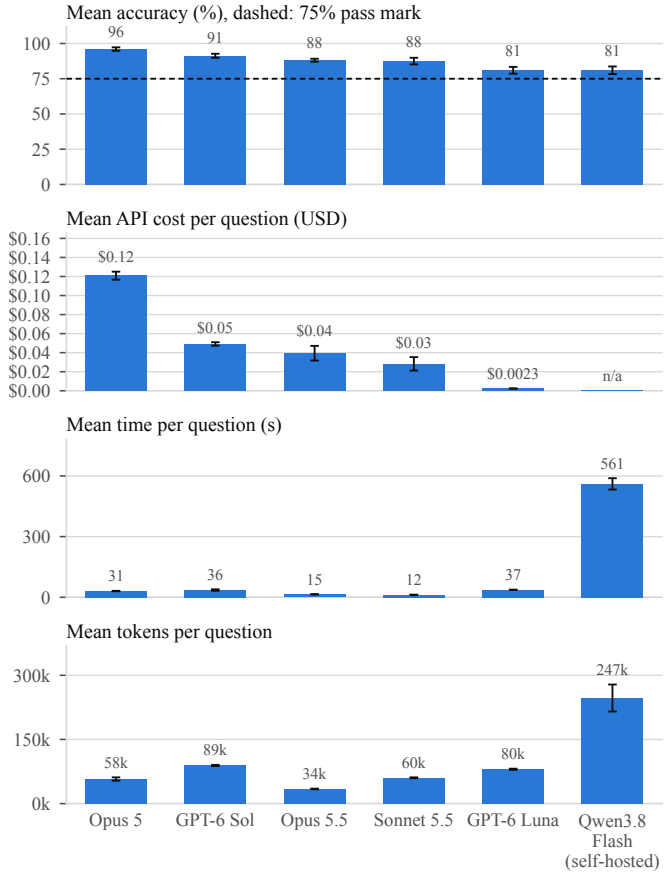


Figure 2: Mean accuracy, API cost in dollars, time and tokens per question, for Claude Opus 5, GPT-6 Sol, Claude Opus 5.5, Claude Sonnet 5.5, GPT-6 Luna, and Qwen3.8 Flash, for the guidelines-available condition only.

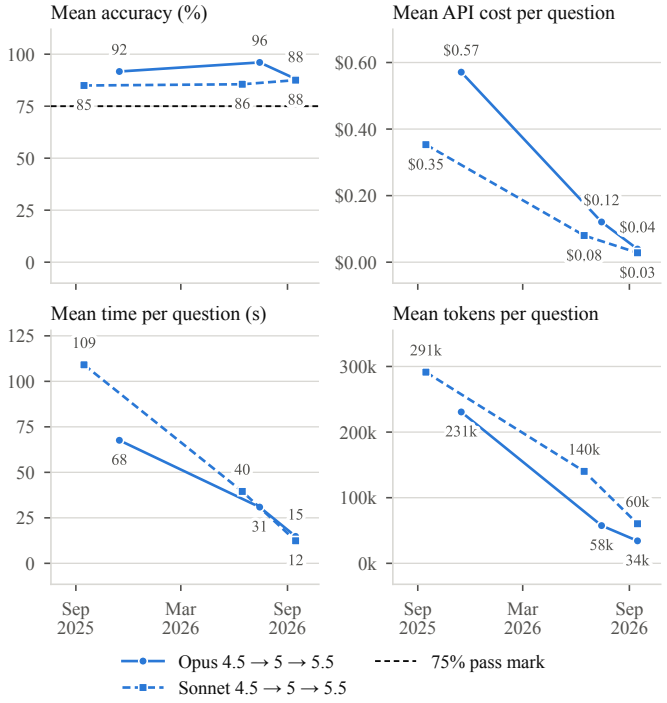


Figure 3: Mean accuracy, API cost per question in dollars, time per question, and tokens per question for three releases of Claude Opus and Claude Sonnet between September 2025 and September 2026 (4.5, 5 and 5.5), for the guidelines-available condition.

average (37 mean seconds per question for GPT-6 Luna compared to 561 for Qwen3.8 Flash). Cloud-hosted models also offer scalability advantages due to the ease of deploying additional parallel workflows – whereas additional servers would need to be purchased for self-hosted solutions due to memory bandwidth limitations.

4.4. Trends towards cheaper and faster answers

Another interesting observation can be made with the trends in accuracy, cost and speed between different releases of the Anthropic models between September 2025 and 2026 (for the guidelines condition), reflected in Figure 3.

We see that whilst the accuracy curve has plateaued since September 2025 (between 85% and 96%, approaching the exam's ceiling), cost per question and time per question have been decreasing rapidly, matching cost trends observed in AI more widely (Gundlach et al., 2026). From the 4.5 to the 5.5 releases, cost per question fell 12–15-fold (e.g. from $0.57 to $0.04 for Opus) and time per question 4.5–9-fold (e.g. from 109 to 12 seconds for Sonnet). Should these trends continue, this has promising implications for the viability of carefully-leveraged low-cost high-throughput parallel AI workflows at scale.

4.5. Agreement across attempts signals answer confidence

In building trust in AI outputs, an important first step is having the ability to inspect reasoning traces, and verify citations to official sources, so that reasoning can be independently validated. Even more useful, however, is having a reliable indication of answer confidence in a given response.

Agreement across repeated attempts proves to be a valuable signal of answer confidence. Pooled across all agents with guidelines available, when all five attempts are unanimous the answer is correct on 96% of questions, but as the attempts diverge the accuracy of the majority-vote answer declines: 79% for a 4–1 split, 65% for 3–2, 56% for 3–1–1, and 44% where no answer gets more than two votes (see Table A.5).

Table 1 shows this holds for every individual agent: for example, Sonnet 5.5 gets 98% of questions correct across the 82 questions where its attempts are unanimous, but only 47% of the remaining 19 questions with attempts that diverge. In possible deployment settings, such as automated consistency checking of draft assessments, this suggests a validation workflow where non-unanimous verdicts across repeat runs are flagged for extra scrutiny from human experts.

5. Limitations

Firstly, we reiterate that the exam only tests knowledge of the Red List methodology, rather than the judgement applied by assessors in real assessments. The exam serves only as a

Table 1: Answer agreement across five attempts with guidelines available: the number of questions on which each agent's attempts were unanimous, and the accuracy of the majority-vote answer when attempts were unanimous versus split. The pooled row gives the mean count across agents. Per-split detail and the closed-book condition are in Table A.5.

		Majority-vote answer correct (%)	
Model	Unanimous (of 101)	When unanimous	When split
Opus 5.5	85	94	62
Sonnet 5.5	82	98	47
Opus 5	85	100	94
Sonnet 5	75	96	58
Opus 4.5	84	99	56
Sonnet 4.5	73	97	66
GPT-6 Sol	89	96	71
GPT-6 Luna	66	92	69
Qwen3.8 Flash	58	98	73
Qwen3.6	33	85	67
All models (pooled)	73.0	96	67

capability floor ahead of considerations for use of AI assistants in assessment workflows.

Another consideration with LLM benchmarks is data contamination: if the exam questions and answers appeared in a model's training data, correct answers could reflect memorisation rather than independent reasoning. However, a benefit of the exam question bank being confidential is that it is out of reach of web crawlers that collect training data for LLMs, which reduces this contamination risk. To verify no public exposure, we ran exact-phrase web searches for each question, as well as targeted searches of question-and-answer sites. Whilst this cannot rule out questions reaching model training data by other means (e.g. from exam candidates pasting them into consumer AI chatbots), the questions would still not be accompanied by the official answer keys, which would be needed if the models were to memorise correct answers during training.

To avoid per-token API costs, we ran the Claude and GPT models via their subscription-based agent harnesses (Claude Code and Codex CLI respectively), whilst the Qwen models were run using OpenCode on a local desktop Mac Studio. The model, agent harness and serving infrastructure are therefore confounded: results thus reflect these specific deployment configurations, rather than purely isolating the effect of the underlying model. Evaluating further deployment settings and providers (e.g. DeepSeek, Google's Gemini, xAI's Grok) is left to future work.

Finally, the sustainability case for self-hosting is based on having control over the energy supply (e.g. a renewables-powered server) instead of depending on third-party AI providers' decisions on the energy mix running their data centres. We did not measure energy usage itself for these experiments – and we acknowledge that external providers can be more energy-efficient per query than self-hosting due to more efficient hardware and batch computations. Similarly, the absence of per-token charges for self-hosted models does not account for hardware and electricity costs, which depend on local rates, and may exceed per-query costs for cheaper closed models such as GPT-6 Luna.

6. Conclusions

This paper has used the official IUCN Red List assessor training exam as a starting point for exploring the capabilities and costs of AI agents, ahead of their consideration for use in IUCN Red List assessment workflows. We show that recent frontier AI models consistently surpass the capability floor represented by the exam: with access to the official IUCN guidelines, Anthropic's Claude Opus 5 reaches 96% mean accuracy (99% taking majority vote over five attempts), and OpenAI's GPT-6 Sol 91%, compared to the 75% exam pass mark. Moreover, self-hosted open-weight models now clear the threshold too: Alibaba's Qwen3.8 Flash, running on a single local Mac Studio desktop, reaches 81% (87% by majority vote), nearly 20 percentage points above its predecessor released four months prior. Although this comes with a speed tradeoff – self-hosted Qwen3.8 Flash around 15 times slower than GPT-6 Luna – it points to viable deployment options where data security or sustainability concerns preclude sending data to external AI providers.

Whilst we stress that the exam serves only as a capability floor and does not test the judgement required in real assessments, these results do strengthen the case for the thoughtful use of AI assistants in assessors' workflows. Where AI tools are used, we recommend simple practices to help assessors verify their outputs: providing the model with the official IUCN guidelines and asking it to cite the relevant passages, so its reasoning and sources for claims can be checked; and asking the same question again in a fresh conversation, treating any disagreement as a flag for closer scrutiny (supported by our finding that agreement across repeated attempts is a strong signal of answer confidence).

Looking ahead, future work could move beyond the exam to evaluating agent behaviour in real-world deployment settings, such as quality and consistency checks of draft assessments, or helping assessors understand commonly confused Red List concepts (like locations and severe fragmentation). In so doing, lessons on how interactive AI assistants can best help assessors could be consolidated, standardised and shared.

In the longer term, should AI continue to become cheaper and faster, more ambitious agentic workflows to assist assessors throughout the assessment lifecycle could become increasingly viable. For example, one could envisage each assessor using agents to continually monitor their assessed species for newly-published occurrence records or literature, flagging when an assessment may need updating, and organising available evidence for species not yet assessed. Of course, such workflows would go well beyond what the exam tests, and would require rigorous research into how reliably agents can retrieve and interpret evidence from occurrence databases and online literature, how to preserve assessor accountability and guard against over-reliance on AI outputs, how to make AI use transparent, traceable and auditable, and who would bear the cost. But given the scale of the challenges facing the Red List – fewer than 20% of plant species assessed, and over 30% of assessed species not

reassessed in over a decade – the potential of carefully leveraging AI agents as part of a scalable solution suggests these are questions worth considering.

Data availability

The full evaluation framework and analysis code underlying this work is available at `https://github.com/shaneweisz/ai-redlist-exam`. Since the exam is used for official assessor certification, the IUCN Red List Unit requires that the questions and answer keys are kept confidential. For researchers wishing to reproduce the evaluation, access can be granted on a case-by-case basis by the IUCN Red List Unit.

Every run reported in this study has been archived in full on Harbor Hub with the full agent transcripts. Clicking on a model label in any of the tables in the paper links to the transcripts from the corresponding run. For the same reason as above, to preserve confidentiality, these transcript archives must be kept private (as they contain the questions and accepted answers), but access can also be granted on a case-by-case basis.

Acknowledgements

Thanks to Caroline Pollock (Senior Programme Coordinator at the IUCN Red List Unit) for facilitating access to the official exam question bank for use in this study.

CRediT authorship contribution statement

Shane Weisz: Conceptualization, Methodology, Software, Investigation, Formal analysis, Data curation, Writing – original draft, Writing – review & editing. **Michael Dales:** Writing – review & editing. **Sadiq Jaffer:** Writing – review & editing. **Steven P. Bachman:** Writing – review & editing. **Anil Madhavapeddy:** Supervision, Funding acquisition, Writing – review & editing.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work, the authors used Claude Code (Anthropic) in order to assist with code development and LaTeX formatting, and suggest edits to wording and structure. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Appendix A. Full results

Appendix A.1. Overall results

Table A.1: Results for every agent configuration and condition (guidelines vs closed-book) on the 101-question exam, with five attempts per question. Metrics are as defined in §3.

	Accuracy (%)				Latency		Tokens / question					API cost
								cache				
Model and condition	mean ± SE	maj@5	pass^5	pass@5	mean s	tok/s	in	write	read	out	total	USD / q
ANTHROPIC'S CLAUDE, CLAUDE CODE CLI												
Opus 5.5 *released Sep 2026*												
Guidelines	88.1 ± 0.5	89.1	83.2	93.1	15	55	6	2.1k	31k	822	34k	0.039
Closed-book	85.2 ± 1.2	87.5	77.2	91.1	12	56	4	803	18k	692	20k	0.024
Sonnet 5.5 *released Sep 2026*												
Guidelines	87.5 ± 1.2	88.1	79.2	97.0	12	61	6	2.3k	57k	758	60k	0.028
Closed-book	83.6 ± 1.4	85.3	70.3	94.1	10	58	4	789	39k	609	40k	0.017
Opus 5 *released Jul 2026*												
Guidelines	96.0 ± 0.6	99.0	87.1	100.0	31	52	16	5.5k	50k	1.6k	58k	0.121
Closed-book	90.3 ± 1.4	91.8	79.2	98.0	18	52	4	1.9k	13k	914	16k	0.049
Sonnet 5 *released Jun 2026*												
Guidelines	85.5 ± 1.4	86.1	72.3	96.0	40	67	12	6.9k	131k	2.6k	140k	0.080
Closed-book	77.4 ± 1.0	78.2	62.4	88.1	27	73	5	2.6k	38k	2.0k	43k	0.038
Opus 4.5 *released Nov 2025*												
Guidelines	91.7 ± 0.5	91.6	85.2	96.0	68	51	52	39k	188k	3.4k	231k	0.571
Closed-book	81.8 ± 1.6	83.7	69.3	90.1	36	55	19	4.4k	30k	2.0k	36k	0.109
Sonnet 4.5 *released Sep 2025*												
Guidelines	85.0 ± 0.9	88.6	70.3	95.0	109	45	73	34k	252k	4.9k	291k	0.353
Closed-book	71.7 ± 1.3	77.8	50.5	87.1	50	51	18	5.9k	27k	2.6k	36k	0.082
OPENAI'S GPT, CODEX CLI												
GPT-6 Sol *released Sep 2026*												
Guidelines	91.3 ± 0.7	92.6	86.1	96.0	36	22	13k	0	75k	780	89k	0.049
Closed-book	86.5 ± 1.8	88.1	73.3	95.0	23	22	4.0k	0	21k	509	25k	0.017
GPT-6 Luna *released Sep 2026*												
Guidelines	81.0 ± 1.2	84.2	62.4	91.1	37	26	12k	0	68k	965	80k	0.002
Closed-book	73.7 ± 1.6	78.7	52.5	90.1	28	25	3.6k	0	24k	685	28k	0.001
ALIBABA'S QWEN, OPENCODE (SELF-HOSTED)												
Qwen3.8 Flash *released Aug 2026*												
Guidelines	81.0 ± 1.4	87.3	58.4	97.0	561	16	238k	0	0	8.9k	247k	n/a
Closed-book	50.7 ± 2.2	63.0	15.8	82.2	236	20	32k	0	0	4.6k	37k	n/a
Qwen3.6-35B-A3B *released Apr 2026*												
Guidelines	61.6 ± 2.5	73.0	28.7	84.2	374	19	331k	0	0	7.0k	338k	n/a
Closed-book	26.9 ± 1.5	29.2	8.9	51.5	108	44	28k	0	0	4.2k	32k	n/a

Appendix A.2. Accuracy by topic

Table A.2: Exam accuracy by topic for each agent configuration, (a) with guidelines available and (b) closed-book.

(a) Guidelines available

		Accuracy (%)										
Topic	*n*	Opus 5.5	Sonnet 5.5	Opus 5	Sonnet 5	Opus 4.5	Sonnet 4.5	GPT-6 Sol	GPT-6 Luna	Qwen3.8 Flash	Qwen 3.6	Mean
General rules	5	100	100	100	100	100	100	100	100	100	92	99
Red List categories	9	100	93	93	76	78	76	100	53	69	58	80
Reduction calculations	15	99	99	99	100	100	100	100	97	91	76	96
Severe fragmentation	7	100	100	100	97	100	94	91	80	94	91	95
Key terms*	9	96	98	96	69	98	89	71	87	80	67	85
Criterion A	7	83	49	83	94	60	60	94	71	51	6	65
Criterion B	8	30	70	95	85	90	75	85	78	72	30	71
Criterion C1	9	89	89	98	98	100	91	100	73	89	60	89
Criteria C2 and D	4	100	70	100	85	100	80	100	75	90	60	86
Criterion E	9	89	82	96	56	98	98	69	96	80	69	83
Supporting information	4	100	100	100	70	80	85	100	90	95	55	88
Mapping standards	5	72	100	96	92	88	68	76	76	72	68	81
Regional assessments	10	88	82	96	86	90	72	100	72	76	62	82
All questions	101	88	88	96	86	92	85	91	81	81	62	85

(b) Closed-book

		Accuracy (%)										
Topic	*n*	Opus 5.5	Sonnet 5.5	Opus 5	Sonnet 5	Opus 4.5	Sonnet 4.5	GPT-6 Sol	GPT-6 Luna	Qwen3.8 Flash	Qwen 3.6	Mean
General rules	5	100	84	88	80	80	84	84	76	76	76	83
Red List categories	9	89	84	82	67	60	56	89	49	33	24	63
Reduction calculations	15	100	99	99	99	100	99	100	99	65	15	87
Severe fragmentation	7	100	97	100	100	94	74	86	89	69	49	86
Key terms*	9	87	91	100	53	60	60	67	62	42	11	63
Criterion A	7	66	63	74	69	54	31	83	60	17	0	52
Criterion B	8	38	70	92	80	88	65	80	80	48	5	64
Criterion C1	9	87	87	98	98	89	87	96	69	51	13	77
Criteria C2 and D	4	100	80	95	75	70	65	100	75	10	15	68
Criterion E	9	89	71	100	49	100	100	82	87	56	47	78
Supporting information	4	95	80	75	75	60	50	75	55	60	45	67
Mapping standards	5	60	80	64	56	80	40	68	60	68	32	61
Regional assessments	10	90	82	82	84	92	70	96	68	52	48	76
All questions	101	85	84	90	77	82	72	87	74	51	27	73

* Generation length (1), continuing decline (3), EOO/AOO (1), mature individuals (1), extreme fluctuation (1) and locations (2).

Appendix A.3. Outcomes and errors by question type

Table A.3: Outcome and error breakdown by question type for each agent configuration, (a) with guidelines available and (b) closed-book.

(a) Guidelines available

Question type and outcome	Opus 5.5	Sonnet 5.5	Opus 5	Sonnet 5	Opus 4.5	Sonnet 4.5	GPT-6 Sol	GPT-6 Luna	Qwen3.8 Flash	Qwen 3.6	All
Category questions (n = 45)											
Correct	179	168	211	184	194	174	203	161	167	97	1738 (77%)
Category one step off on the threat scale	11	25	8	19	16	20	6	31	21	51	208 (9%)
Category further off, or a non-threat category	0	8	0	10	9	13	5	20	14	17	96 (4%)
Right category and criterion, extra subcriteria	28	11	2	3	4	8	6	8	8	1	79 (4%)
Right category, criteria added when not asked for	1	3	0	2	0	0	0	0	2	20	28 (1%)
Right category, subcriteria or numerals wrong	0	7	2	0	1	6	0	0	5	12	33 (1%)
Right category, criterion wrong	6	1	2	7	1	4	5	5	6	12	49 (2%)
Right category, spacing only	0	2	0	0	0	0	0	0	0	0	2 (0%)
No answer written	0	0	0	0	0	0	0	0	2	15	17 (1%)
Multiple-select questions (n = 33)											
Correct	154	161	162	143	155	146	156	146	140	128	1491 (90%)
Right options plus an extra	7	0	1	10	3	9	6	8	18	13	75 (5%)
A right option missed	0	0	0	6	4	3	0	1	1	9	24 (1%)
Other wrong set	4	4	2	6	3	7	3	10	6	9	54 (3%)
No answer written	0	0	0	0	0	0	0	0	0	6	6 (0%)
Calculation questions (n = 21)											
Correct	102	103	102	95	104	100	92	97	94	78	967 (92%)
Wrong number	3	2	3	10	1	5	13	8	11	26	82 (8%)
No answer written	0	0	0	0	0	0	0	0	0	1	1 (0%)
Matching questions (n = 2)											
Correct	10	10	10	10	10	9	10	5	8	8	90 (90%)
Pairs wrong	0	0	0	0	0	1	0	5	2	2	10 (10%)

(b) Closed-book

Question type and outcome	Opus 5.5	Sonnet 5.5	Opus 5	Sonnet 5	Opus 4.5	Sonnet 4.5	GPT-6 Sol	GPT-6 Luna	Qwen3.8 Flash	Qwen 3.6	All
Category questions (n = 45)											
Correct	169	165	199	162	172	154	198	153	87	39	1498 (67%)
Category one step off on the threat scale	18	23	16	30	23	28	14	39	63	63	317 (14%)
Category further off, or a non-threat category	0	8	1	16	19	14	1	18	33	31	141 (6%)
Right category and criterion, extra subcriteria	22	12	3	4	4	14	7	7	8	0	81 (4%)
Right category, criteria added when not asked for	5	4	1	3	0	0	2	5	7	26	53 (2%)
Right category, subcriteria or numerals wrong	1	3	1	1	2	4	0	1	13	4	30 (1%)
Right category, criterion wrong	10	10	3	8	4	11	3	2	14	10	75 (3%)
Right category, criteria missing	0	0	0	0	0	0	0	0	0	2	2 (0%)
Right category, spacing only	0	0	1	1	1	0	0	0	0	0	3 (0%)
No answer written	0	0	0	0	0	0	0	0	0	50	50 (2%)
Multiple-select questions (n = 33)											
Correct	152	147	143	131	139	103	138	122	103	77	1255 (76%)
Right options plus an extra	10	9	18	16	16	27	19	21	37	33	206 (12%)
A right option missed	1	4	0	9	8	9	0	6	3	13	53 (3%)
Other wrong set	2	5	4	9	2	26	8	16	22	35	129 (8%)
No answer written	0	0	0	0	0	0	0	0	0	7	7 (0%)
Calculation questions (n = 21)											
Correct	99	100	104	88	92	96	91	91	64	16	841 (80%)
Wrong number	6	5	1	17	13	9	14	14	41	76	196 (19%)
No answer written	0	0	0	0	0	0	0	0	0	13	13 (1%)
Matching questions (n = 2)											
Correct	10	10	10	10	10	9	10	6	2	4	81 (81%)
Pairs wrong	0	0	0	0	0	1	0	4	8	5	18 (18%)
No answer written	0	0	0	0	0	0	0	0	0	1	1 (1%)

Appendix A.4. Tool usage

Table A.4: Tool calls per attempt for each agent configuration and condition.

Model and condition	Tool calls per attempt								Mean calls per attempt, by what the call did					
	mean	attempts making *n* calls (%)						max	search	read	list	compute	answer	other
		0	1	2	3–5	6–10	11+		(`grep`)	(`Read`, `sed`)	(`ls`, `find`)	(`python`)	(`Write`)	
ANTHROPIC'S CLAUDE, CLAUDE CODE CLI														
Opus 5.5 *released Sep 2026*														
Guidelines	1.90	–	62	10	27	1	–	6	0.33	0.24	0.32	0.00	1.00	0.00
Closed-book	1.07	–	93	7	–	–	–	2	0.00	0.07	0.00	0.00	1.00	0.00
Sonnet 5.5 *released Sep 2026*														
Guidelines	1.82	–	68	8	22	3	–	10	0.36	0.31	0.14	0.00	1.00	0.00
Closed-book	1.07	–	93	7	–	–	–	2	0.00	0.07	0.00	0.00	1.00	0.00
Opus 5 *released Jul 2026*														
Guidelines	4.67	–	24	3	34	36	4	15	1.33	0.66	1.48	0.11	1.10	0.00
Closed-book	1.11	–	90	9	1	–	–	3	0.00	0.07	0.00	0.03	1.01	0.00
Sonnet 5 *released Jun 2026*														
Guidelines	5.04	–	16	14	34	28	8	28	2.00	0.95	0.78	0.23	1.08	0.01
Closed-book	1.28	–	75	22	2	–	–	5	0.00	0.08	0.00	0.04	1.16	0.00
Opus 4.5 *released Nov 2025*														
Guidelines	5.65	–	–	–	60	36	4	18	0.99	1.90	1.20	0.01	1.55	0.00
Closed-book	1.25	–	76	24	<1	–	–	3	0.00	0.08	0.00	0.00	1.17	0.00
Sonnet 4.5 *released Sep 2025*														
Guidelines	8.33	–	–	–	32	46	22	33	2.08	2.28	2.22	0.12	1.64	0.00
Closed-book	1.29	–	73	26	1	–	–	5	0.00	0.07	0.01	0.02	1.19	0.00
OPENAI'S GPT, CODEX CLI														
GPT-6 Sol *released Sep 2026*														
Guidelines	4.72	–	2	6	63	29	1	12	1.99	1.56	0.01	0.04	1.01	0.10
Closed-book	1.09	–	91	8	1	–	–	3	0.00	0.07	0.00	0.01	1.01	0.00
GPT-6 Luna *released Sep 2026*														
Guidelines	4.29	–	5	4	74	18	–	9	1.87	1.08	0.24	0.03	1.02	0.05
Closed-book	1.26	–	82	11	6	–	–	5	0.01	0.07	0.09	0.01	1.02	0.06
ALIBABA'S QWEN, OPENCODE (SELF-HOSTED)														
Qwen3.8 Flash *released Aug 2026*														
Guidelines	12.29	–	–	1	7	41	50	108	4.99	3.48	2.21	0.46	1.12	0.04
Closed-book	2.03	–	50	40	8	1	2	35	0.11	0.17	0.16	0.36	1.21	0.01
Qwen3.6-35B-A3B *released Apr 2026*														
Guidelines	11.09	–	–	<1	7	51	42	51	1.82	4.62	2.45	0.31	1.87	0.02
Closed-book	2.19	7	22	42	25	3	1	14	0.02	0.32	0.05	0.13	1.66	0.00

Appendix A.5. Answer consistency across attempts

Table A.5: How each question's five attempts split into distinct answers for each agent configuration, (a) with guidelines available and (b) closed-book. For example, 5 means all five attempts yielded the same answer, whereas 3-2 means three attempts were one distinct answer, and two attempts another produced answer. Each cell gives the number of questions with that split on which the majority-vote answer was correct, over the number of questions with that split. The last two rows of each panel average these counts over all agents, and give the resulting percentage correct for each split.

	Majority-vote answer correct / questions, by split (of 101)						
Model	5	4–1	3–2	3–1–1	2–2–1[†]	2–1–1–1	all differ[†]
(a) Guidelines available							
Opus 5.5	80/85	4/7	5/6	0/1	1/1	0/1	–
Sonnet 5.5	80/82	5/10	2/3	1/4	1/2	–	–
Opus 5	85/85	7/7	7/7	1/2	–	–	–
Sonnet 5	72/75	12/17	1/6	2/3	–	–	–
Opus 4.5	83/84	6/8	2/5	–	1.5/2	0/1	0/1
Sonnet 4.5	71/73	10/15	3/5	2/3	1.5/3	2/2	–
GPT-6 Sol	85/89	4/6	2/3	2/2	0.5/1	–	–
GPT-6 Luna	61/66	21/24	1/4	0/1	1/2	1/4	–
Qwen3.8 Flash	57/58	12/14	10/13	4/5	3/7	2/3	0.2/1
Qwen3.6	28/33	25/27	7/10	7/13	1.5/3	5/12	0.2/3
Mean	70.2/73.0	10.6/13.5	4.0/6.2	1.9/3.4	1.1/2.1	1.0/2.3	<0.1/0.5
% majority correct (pooled)	96%	79%	65%	56%	52%	43%	8%
(b) Closed-book							
Opus 5.5	75/79	5/8	6/7	2/3	0/2	0/1	0.4/1
Sonnet 5.5	70/74	9/14	4/9	3/3	–	–	0.2/1
Opus 5	78/80	10/11	3/6	1/2	0.5/1	–	0.2/1
Sonnet 5	63/75	13/17	1/5	1/1	1/2	0/1	–
Opus 4.5	67/74	12/14	3/7	1/2	0.5/2	1/2	–
Sonnet 4.5	51/59	16/18	6/10	1/3	1/2	3/5	0.6/4
GPT-6 Sol	70/75	12/15	7/11	–	–	–	–
GPT-6 Luna	52/56	14/21	7/11	4/6	1.5/4	1/1	0/2
Qwen3.8 Flash	16/21	20/24	12/15	7/14	3/10	5/14	0.6/3
Qwen3.6	8/17	9/21	4/8	4/17	1.5/7	2/19	1/12
Mean	55.0/61.0	12.0/16.3	5.3/8.9	2.4/5.1	0.9/3.0	1.2/4.3	0.3/2.4
% majority correct (pooled)	90%	74%	60%	47%	30%	28%	12%

[†] A tied split: no single answer has the most votes. Majority vote is then scored as the share of the tied answers that are correct, i.e. the expected score under random tie-breaking.

References

Akçakaya, H.R., Milner-Gulland, E.J., Hoffmann, M., Regan, H.M., Naujokaitis-Lewis, I., Punt, A.E., Gaston, K.J., Keith, D.A., Butchart, S.H.M., Stuart, S.N., Collar, N.J., Di Marco, M., Bohm, M., Berryman, A.J., Mallon, D.P., Hochkirch, A., Dulvy, N.K., Rivers, M., Brooks, T.M., Westrip, J.R.S., Rodríguez, J.P., Hilton-Taylor, C., Scott, J., Tarr, S., Ledger, S.H., Linardich, C., 2026. Three decades of classifying threatened species: Lessons learned from and about the IUCN Red List criteria for quantifying extinction risk. Proceedings of the Royal Society B: Biological Sciences 293, 20260316. doi:10.1098/rspb.2026.0316.

Bachman, S.P., Field, R., Reader, T., Raimondo, D., Donaldson, J., Schatz, G.E., Lughadha, E.N., 2019. Progress, challenges and opportunities for Red Listing. Biological Conservation 234, 45–55. doi:10.1016/j.biocon.2019.03.002.

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H.P.d.O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., Ryder, N., Pavlov, M., Power, A., Kaiser, L., Bavarian, M., Winter, C., Tillet, P., Such, F.P., Cummings, D., Plappert, M., Chantzis, F., Barnes, E., Herbert-Voss, A., Guss, W.H., Nichol, A., Paino, A., Tezak, N., Tang, J., Babuschkin, I., Balaji, S., Jain, S., Saunders, W., Hesse, C., Carr, A.N., Leike, J., Achiam, J., Misra, V., Morikawa, E., Radford, A., Knight, M., Brundage, M., Murati, M., Mayer, K., Welinder, P., McGrew, B., Amodei, D., McCandlish, S., Sutskever, I., Zaremba, W., 2021. Evaluating Large Language Models Trained on Code. doi:10.48550/arXiv.2107.03374, arXiv:2107.03374.

Díaz, S., Settele, J., Brondízio, E.S., Ngo, H.T., Agard, J., Arneth, A., Balvanera, P., Brauman, K.A., Butchart, S.H.M., Chan, K.M.A., Garibaldi, L.A., Ichii, K., Liu, J., Subramanian, S.M., Midgley, G.F., Miloslavich, P., Molnár, Z., Obura, D., Pfaff, A., Polasky, S., Purvis, A., Razzaque, J., Reyers, B., Chowdhury, R.R., Shin, Y.J., Visseren-Hamakers, I., Willis, K.J., Zayas, C.N., 2019. Pervasive human-driven decline of life on Earth points to the need for transformative change. Science 366, eaax3100. doi:10.1126/science.aax3100.

Dorm, F., Millard, J., Purves, D., Harfoot, M., Mac Aodha, O., 2026. Large language models possess some ecological knowledge, but how much? Ecological Informatics 95, 103699. doi:10.1016/j.ecoinf.2026.103699.

Goettsch, B., Hilton-Taylor, C., Cruz-Piñón, G., Duffy, J.P., Frances, A., Hernández, H.M., Inger, R., Pollock, C., Schipper, J., Superina, M., Taylor, N.P., Tognelli, M., Abba, A.M., Arias, S., Arreola-Nava, H.J., Baker, M.A., Bárcenas, R.T., Barrios, D., Braun, P., Butterworth, C.A., Búrquez, A., Caceres, F., Chazaro-Basañez, M., Corral-Díaz, R., Del Valle Perea, M., Demaio, P.H., Duarte De Barros, W.A., Durán, R., Yancas, L.F., Felger, R.S., Fitz-Maurice, B., Fitz-Maurice, W.A., Gann, G., Gómez-Hinostrosa, C., Gonzales-Torres, L.R., Patrick Griffith, M., Guerrero, P.C., Hammel, B., Heil, K.D., Hernández-Oria, J.G., Hoffmann, M., Ishihara, M.I., Kiesling, R., Larocca, J., León-de La Luz, J.L., Loaiza S., C.R., Lowry, M., Machado, M.C., Majure, L.C., Ávalos, J.G.M., Martorell, C., Maschinski, J., Méndez, E., Mittermeier, R.A., Nassar, J.M., Negrón-Ortiz, V., Oakley, L.J., Ortega-Baes, P., Ferreira, A.B.P., Pinkava, D.J., Porter, J.M., Puente-Martinez, R., Gamarra, J.R., Pérez, P.S., Martínez, E.S., Smith, M., Manuel Sotomayor M. Del C., J., Stuart, S.N., Muñoz, J.L.T., Terrazas, T., Terry, M., Trevisson, M., Valverde, T., Van Devender, T.R., Véliz-Pérez, M.E., Walter, H.E., Wyatt, S.A., Zappi, D., Alejandro Zavala-Hurtado, J., Gaston, K.J., 2015. High proportion of cactus species threatened with extinction. Nature Plants 1, 15142. doi:10.1038/nplants.2015.142.

Gundlach, H., Lynch, J., Mertens, M., Thompson, N., 2026. The Price of Progress: Price Performance and the Future of AI. doi:10.48550/arXiv.2511.23455, arXiv:2511.23455.

Harbor Framework Team, 2026. Harbor: A framework for evaluating and optimizing agents and models in container environments. URL: https://doi.org/10.5281/zenodo.20953922, doi:10.5281/zenodo.20953922.

Huang, W., Lee, C., Tng, L., Ge, S., 2026. DeepSWE: Measuring Frontier Coding Agents on Original, Long-Horizon Engineering Tasks. doi:10.48550/arXiv.2607.07946, arXiv:2607.07946.

IUCN, 2012a. Guidelines for Application of IUCN Red List Criteria at Regional and National Levels: Version 4.0. Technical Report. IUCN. Gland, Switzerland and Cambridge, UK. URL: https://www.iucnredlist.org/resources/regionalguidelines.

IUCN, 2012b. IUCN Red List Categories and Criteria: Version 3.1. Second edition. IUCN. Gland, Switzerland and Cambridge, UK. URL: https://www.iucnredlist.org/resources/categories-and-criteria.

IUCN, 2013. Documentation Standards and Consistency Checks for IUCN Red List Assessments and Species Accounts. Version 2. IUCN. URL: https://www.iucnredlist.org/resources/supporting-information-guidelines.

IUCN, 2026a. Assessing species' extinction risk using IUCN Red List methodology. Online training course, hosted by The Nature Conservancy on ConservationTraining.org. URL: https://www.conservationtraining.org/course/index.php?categoryid=23. accessed 2026-06-08.

IUCN, 2026b. Assessment process: Raw data to red list. The IUCN Red List of Threatened Species. URL: https://www.iucnredlist.org/assessment/process. accessed 2026-10-13.

IUCN, 2026c. Barometer of Life. URL: https://www.iucnredlist.org/about/barometer-of-life. accessed 2026-06-08.

IUCN, 2026d. The IUCN red list of threatened species. version 2026-1. URL: https://www.iucnredlist.org. accessed 2026-10-13.

IUCN, 2026e. Table 1a: Number of species evaluated in relation to the overall number of described species, and numbers of threatened species by major groups of organisms. The IUCN Red List of Threatened Species. Summary Statistics. Version 2026-1. URL: https://www.iucnredlist.org/resources/summary-statistics. accessed 2026-10-13.

IUCN, 2026f. Table 1b: Numbers of threatened species by major groups of organisms (1996–2026). The IUCN Red List of Threatened Species. Summary Statistics. Version 2026-1. URL: https://www.iucnredlist.org/resources/summary-statistics. accessed 2026-10-13.

IUCN SSC Red List Technical Working Group, 2024. Mapping Standards and Data Quality for IUCN Red List Spatial Data. Version 1.20. IUCN. URL: https://www.iucnredlist.org/resources/mappingstandards.

IUCN Standards and Petitions Committee, 2024. Guidelines for Using the IUCN Red List Categories and Criteria. Version 16. Standards and Petitions Committee. URL: https://www.iucnredlist.org/resources/redlistguidelines.

Iyer, R., Christie, A.P., Madhavapeddy, A., Reynolds, S., Sutherland, W., Jaffer, S., 2025. Careful design of Large Language Model pipelines enables expert-level retrieval of evidence-based information from syntheses and databases. PLOS ONE 20, e0323563. doi:10.1371/journal.pone.0323563.

Jaffer, S., Morgan, W., Reynolds, S., Christie, A., Madhavapeddy, A., Sutherland, W., 2025. AI-assisted Living Evidence Databases for Conservation Science. doi:10.33774/coe-2025-rmsqf.

Kwa, T., West, B., Becker, J., Deng, A., Garcia, K., Hasin, M., Jawhar, S., Kinniment, M., Rush, N., Arx, S.V., Bloom, R., Broadley, T., Du, H., Goodrich, B., Jurkovic, N., Miles, L.H., Nix, S., Lin, T., Painter, C., Parikh, N., Rein, D., Sato, L.J.K., Wijk, H., Ziegler, D.M., Barnes, E., Chan, L., 2026. Measuring AI Ability to Complete Long Software Tasks. doi:10.48550/arXiv.2503.14499, arXiv:2503.14499.

Lindenmayer, D., Scheele, B., 2017. Do not publish. Science 356, 800–801. doi:10.1126/science.aan1362.

Mace, G.M., Collar, N.J., Gaston, K.J., Hilton-Taylor, C., Akçakaya, H.R., Leader-Williams, N., Milner-Gulland, E., Stuart, S.N., 2008. Quantification of Extinction Risk: IUCN's System for Classifying Threatened Species. Conservation Biology 22, 1424–1442. doi:10.1111/j.1523-1739.2008.01044.x.

Merrill, M.A., Shaw, A.G., Carlini, N., Li, B., Raj, H., Bercovich, I., Shi, L., Shin, J.Y., Walshe, T., Buchanan, E.K., Shen, J., Ye, G., Lin, H., Poulos, J., Wang, M., Nezhurina, M., Jitsev, J., Lu, D., Mastromichalakis, O.M., Xu, Z., Chen, Z., Liu, Y., Zhang, R., Chen, L.L., Kashyap, A., Uslu, J.L., Li, J., Wu, J., Yan, M., Bian, S., Sharma, V., Sun, K., Dillmann, S., Anand, A., Lanpouthakoun, A., Koopah, B., Hu, C., Guha, E., Dreiman, G.H.S., Zhu, J., Krauth, K., Zhong, L., Muennighoff, N., Amanfu, R., Tan, S., Pimpalgaonkar, S., Aggarwal, T., Lin, X., Lan, X., Zhao, X., Liang, Y., Wang, Y., Wang, Z., Zhou, C., Heineman, D., Liu, H., Trivedi, H., Yang, J., Lin, J., Shetty, M., Yang, M., Omi, N., Raoof, N., Li, S., Zhuo, T.Y., Lin, W., Dai, Y., Wang, Y., Chai, W., Zhou, S., Wahdany, D., She, Z., Hu, J., Dong, Z., Zhu, Y., Cui, S., Saiyed, A., Kolbeinsson, A., Hu, J., Rytting, C.M., Marten, R., Wang, Y., Dimakis, A., Konwinski, A., Schmidt, L., 2026. Terminal-Bench: Benchmarking Agents on Hard, Realistic Tasks in Command Line Interfaces. doi:10.48550/arXiv.2601.11868, arXiv:2601.11868.

Patwardhan, T., Dias, R., Proehl, E., Kim, G., Wang, M., Watkins, O., Fishman, S.P., Aljubeh, M., Thacker, P., Fauconnet, L., Kim, N.S., Chao, P., Miserendino, S., Chabot, G., Li, D., Sharman, M., Barr, A., Glaese, A., Tworek, J., 2025. GDPval: Evaluating AI Model Performance on Real-World Economically Valuable Tasks. doi:10.48550/arXiv.2510.04374, arXiv:2510.04374.

Pollock, L.J., Kitzes, J., Beery, S., Gaynor, K.M., Jarzyna, M.A., Mac Aodha, O., Meyer, B., Rolnick, D., Taylor, G.W., Tuia, D., Berger-Wolf, T., 2025. Harnessing artificial intelligence to fill global shortfalls in biodiversity knowledge. Nature Reviews Biodiversity 1, 166–182. doi:10.1038/s44358-025-00022-3.

Reynolds, S.A., Beery, S., Burgess, N., Burgman, M., Butchart, S.H.M., Cooke, S.J., Coomes, D., Danielsen, F., Minin, E.D., Durán, A.P., Gassert, F., Hinsley, A., Jaffer, S., Jones, J.P.G., Li, B.V., Aodha, O.M., Madhavapeddy, A., O'Donnell, S.A.L., Oxbury, W.M., Peck, L., Pettorelli, N., Rodríguez, J.P., Shuckburgh, E., Strassburg, B., Yamashita, H., Miao, Z., Sutherland, W.J., 2025. The potential for AI to revolutionize conservation: A horizon scan. Trends in Ecology & Evolution 40, 191–207. doi:10.1016/j.tree.2024.11.013.

Rodrigues, A.S.L., Pilgrim, J.D., Lamoreux, J.F., Hoffmann, M., Brooks, T.M., 2006. The value of the IUCN Red List for conservation. Trends in Ecology & Evolution 21, 71–76. doi:10.1016/j.tree.2005.10.010.

Stuart, S.N., Wilson, E.O., McNeely, J.A., Mittermeier, R.A., Rodríguez, J.P., 2010. The Barometer of Life. Science 328, 177–177. doi:10.1126/science.1188606.

Uryu, S., 2026. Benchmarking large language models for biodiversity assessment: Performance and biases across IUCN Red List species. Ecological Informatics 96, 103839. doi:10.1016/j.ecoinf.2026.103839.

Willison, S., 2026. Qwen 3.8 27b is excellent, but it defaults to wildly overthinking things. https://simonwillison.net/2026/Aug/16/qwen-38-27b/. Accessed 2026-10-20.

Yao, S., Shinn, N., Razavi, P., Narasimhan, K., 2025. τ-bench: A benchmark for tool-agent-user interaction in real-world domains, in: International Conference on Learning Representations (ICLR). URL: https://proceedings.iclr.cc/paper_files/paper/2025/hash/1b126cc38b8638e07bef37e7b2bb72bf-Abstract-Conference.html.