HERBENCH: UNDERSTANDING FINE-GRAINED VISUAL RECOGNITION IN MULTIMODAL AGENTS THROUGH TRAIT-LEVEL EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Large Language Models (MLLMs) equipped with active visual tools have significantly advanced fine-grained visual recognition (FGVR) by inspecting local image regions. However, despite access to high-resolution visual evidence, multimodal agents still struggle with reliable specimen identification, leaving a fundamental question unanswered: *Do identification failures arise from selecting the wrong traits, misjudging subtle trait states, or failing to use those states to distinguish candidate categories?* To dissect this visual reasoning chain, we formulate a trait-level evaluation across three dimensions: *trait selection*, *trait judgment*, and *the use of trait evidence for identification*. We introduce **HerBench**, a specialized benchmark linking 851 megapixel-scale herbarium specimens (spanning 239 species) to 5,325 fine-grained trait questions derived from taxonomist-authored dichotomous keys (dKeys). Furthermore, inspired by professional taxonomic workflows, we propose a dKey-guided agentic framework that interleaves active visual inspection with expert key traversal to systematically guide trait selection and evidence integration. Extensive evaluations across nine state-of-the-art multimodal agents reveal a substantial *trait selection gap*: agents frequently answer isolated trait questions correctly yet completely omit or fail to make definite judgments on the same traits during end-to-end identification. Moreover, even when observation targets are explicitly specified, trait judgment accuracy remains modest (25.51%–64.55%). While dKey guidance improves species F1 for eight of nine models, controlled experiments with reference traits demonstrate that its performance gains are substantially amplified when accurate trait states are supplied, highlighting that the true utility of expert guidance depends fundamentally on the fidelity of the agent's perceptual evidence.

1 INTRODUCTION

Fine-grained visual recognition (FGVR) poses a longstanding challenge in computer vision: categories share remarkably similar global appearances while differing only in subtle, localized morphological traits (He et al., 2025; Zhang et al., 2024). Herbarium specimen identification epitomizes this challenge at its extreme (Tan et al., 2019). Decades of physical drying, pressing, and organ degradation frequently obscure or destroy diagnostic structures (such as leaf margins, indumentum or microscopic floral parts). Consequently, reliable specimen recognition demands more than generic recognition; it necessitates purposeful, targeted visual inspection of sparse remaining structures to gather discriminative evidence.

Recent advances in tool-augmented Multimodal Large Language Models (MLLMs) and agentic vision frameworks offer a promising path forward (Zheng et al., 2026; Yao et al., 2023). By actively acquiring zoomed-in crops and dynamically revisiting local details, multimodal agents can extend their visual acuity far beyond fixed-resolution representations. Yet, our empirical investigations reveal that access to local image crops alone does not translate to robust identification (Section 5.2). Solving FGVR requires mastering three tightly coupled, hierarchical capabilities: (1) **Trait Selection** (*identifying which diagnostic traits to observe*), (2) **Trait Judgment** (*accurately determining the subtle morphological states of those traits*), and (3) **Evidence Use** (*logically mapping contrasting trait states to prune candidate taxa*). Current evaluations either conflate these capabilities into

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

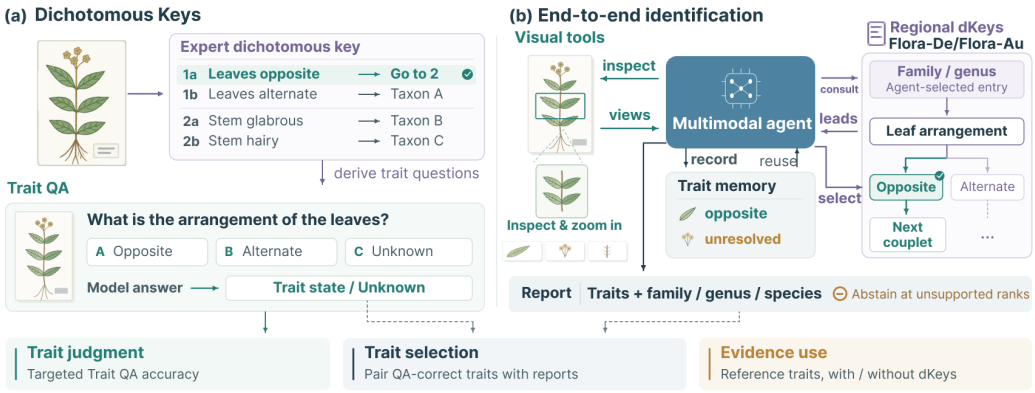

Figure 1: **HerBench overview.** (a) Taxonomist-authored dKeys provide structured morphological traits evaluated in Trait QA. (b) Our dKey-guided agentic framework combines iterative visual inspection, persistent trait memory, and regional keys for stepwise identification. The bottom row summarizes our analyses of trait judgment, trait selection, and evidence use.

an uninterpretable end-to-end score or measure isolated trait accuracy detached from identification, masking whether agent failures stem from unguided visual search, perceptual misjudgment, or defective taxonomic deduction.

To bridge this gap, we turn to the formal methodologies developed by botanists: *dichotomous keys* (dKeys) (Tofilski, 2018; Thiele & Klazenga, 2016). Compiled over centuries of taxonomic research, dKeys structure identification into successive, paired couplets of contrasting morphological statements. A dKey explicitly specifies *what* to inspect and *how* each observed state rules out competing taxa (Figure 1(a)). This motivates a fundamental question: **Can taxonomic knowledge illuminate where multimodal agents fail in visual reasoning chain, and can it guide agentic identification?**

To investigate this question, we present our study structured around three core challenges:

- **Q1: Trait Selection vs. Judgment — Do agents examine the right traits during identification, and can they judge them when explicitly queried?** Through paired evaluations between Trait Question Answering (Trait QA) and end-to-end identification reports, we find a severe *trait selection gap*: models often answer questions about diagnostic traits correctly, but fail to form any definite judgment about those identical traits during unguided identification. Even with target structures specified, judging subtle morphological states remains difficult, yielding accuracies between 25.51% and 64.55%.
- **Q2: Expert-Guided Traversal — Can dichotomous keys systematically guide agents to inspect relevant traits and use their evidence?** Inspired by professional taxonomic workflows, we propose a dKey-guided agentic framework that provides regional dKey graphs as interactive tools (Figure 1(b)). By interleaving active region cropping with key traversal, dKey guidance substantially reduces omitted traits and improves species F1 for eight of nine evaluated models.
- **Q3: Evidence Bottleneck — Does the efficacy of expert guidance depend on visual perception or taxonomic deduction?** In controlled experiments supplying reference trait states, we decouple perceptual error from taxonomic reasoning. We observe that while models struggle to map isolated trait states to species on their own, the F1 gains unlocked by dKey guidance increase drastically when reliable trait states are provided (e.g., yielding a +24.92 point surge on Kimi-K3 (Kimi Team, 2026) vs. only +2.20 without reference traits). This demonstrates that the true bottleneck of expert-guided reasoning lies in the fidelity of fine-grained visual evidence.

To ground these investigations in an authentic benchmark, we introduce **HerBench**, which aligns 851 megapixel-scale specimens (averaging 66.4 megapixels) across 239 species with regional dKeys and 5,325 validated trait questions. Our main contributions are summarized as follows:

- We introduce **HerBench**, the first expert-grounded benchmark for agentic FGVR, which links 851 high-resolution herbarium specimens (239 species) with validated dKey graphs and 5,325 source-

grounded trait questions, and pairs Trait QA with E2E identification on the same specimens to evaluate trait selection, trait judgment, and evidence use.

- We propose a **dKey-guided agentic framework** that interleaves active region inspection with key traversal, guiding agents on which structures to examine and how their states prune candidate taxa.
- Evaluating **nine open-source and proprietary models**, we reveal a trait selection gap and show that the benefit of expert guidance is bounded by the fidelity of fine-grained visual evidence.

2 Related Work

High-Resolution Perception and Agentic FGVR. Fine-grained visual recognition (FGVR) requires recognizing subtle local traits that distinguish visually similar categories (He et al., 2025; Zhang et al., 2024). Fixed-resolution encoders can lose such details when downsampling high-resolution images (Wang et al., 2024). Prior work mitigates this with cropping and multi-level visual search (Wu & Xie, 2023), recursive decomposition (Wang et al., 2024), and agentic visual tools for active region selection and iterative inspection (Wang et al., 2025; Xue et al., 2026; Zheng et al., 2026). Yet these methods mainly study local evidence acquisition rather than whether models select diagnostic traits and use them correctly. Our work instead decouples targeted trait inspection from E2E identification to study the gap between trait perception and taxonomic reasoning.

Fine-Grained and High-Resolution Benchmarks. Mainstream vision-language benchmarks largely use low-resolution images ($\leq$2K), leaving model capabilities on ultra-detailed scenes underexplored (Wang et al., 2024). HR-Bench evaluates 4K and 8K imagery for generic scene understanding (Wang et al., 2024), whereas biological FGVR demands even finer diagnostic fidelity. Existing biological benchmarks, including CUB-200-2011 (Wah et al., 2011), VLM4Bio (Maruf et al., 2024), and the Herbarium Challenge (Tan et al., 2019), focus on monolithic category classification or isolated attribute QA rather than connecting trait judgments to taxonomic decisions. HerBench fills this gap by aligning high-resolution specimens (averaging 66.4 megapixels) with taxonomic labels and source-grounded trait questions, enabling joint evaluation of trait selection, trait judgment, and trait evidence use.

Knowledge-Guided Visual Reasoning. External knowledge retrieval supports MLLMs in specialized domains (Li & Peng, 2026). Prior work uses unstructured descriptions from Wikipedia or Wikispecies to ground rare-taxon recognition (Lesperance et al., 2025; Chen et al., 2026). However, unstructured encyclopedic texts are often noisy and lack procedural logic for resolving visual ambiguities. **HerBench** instead harnesses taxonomist-authored dichotomous keys (dKeys) (Tofilski, 2018; Thiele & Klazenga, 2016), whose contrasting morphological choices provide both trait-level ground truth and structured identification paths.

3 HerBench: Dataset and Task Design

To systematically investigate how multimodal agents orchestrate *trait selection*, *trait judgment*, and *evidence use*, we construct **HerBench**, an expert-grounded benchmark built upon authentic botanical specimens and taxonomist-authored dichotomous keys (dKeys). **HerBench** contains **851 high-resolution specimens** covering **239 species** from Germany and Australia, together with standardized taxonomies and **5,325 trait questions** derived from published floras. By pairing Trait QA with end-to-end identification on the same specimens, HerBench supports trait-level analysis of the identification process. Figure 2 summarizes the construction pipeline.

3.1 Collection Preparation and Label De-Identification

We curated an initial collection of 2,484 digitized herbarium sheets through institutional collaborations, comprising 1,484 specimens from Germany and 1,000 from Australia. Herbarium sheets are scanned at megapixel fidelity, retaining their uncompressed acquisition resolution (averaging 66.4 megapixels in the final benchmark) to preserve microscopic diagnostic structures such as leaf pubescence, petal venation, and glandular trichomes. To prevent identification through specimen labels, we mask identifying text through automated detection, model-assisted inspection, and human review. Edits are restricted to text regions to preserve plant morphology.

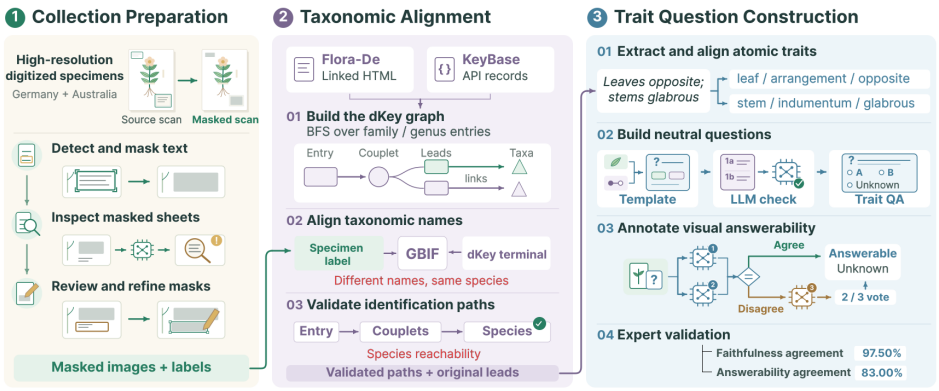

Figure 2: **HerBench construction pipeline.** (1) Collection preparation with multi-stage label masking; (2) taxonomic alignment, comprising dKey graph construction, GBIF name alignment, and path validation; (3) trait question construction, comprising atomic trait extraction, neutral question synthesis, multi-model answerability voting, and expert validation.

3.2 Dichotomous Key and Taxonomic Alignment

A **dichotomous key (dKey)** is a structured, stepwise identification guide organized into paired alternatives, termed *couplets* (Tofilski, 2018). Each branch within a *couplet*, or *lead*, describes one or more diagnostic morphological states (e.g., *leaves opposite*). A user navigates a dKeys by selecting the *lead* that matches the specimen and follow it to the next *couplet* or a taxonomic designation. Thus, dKeys explicitly formalize what structures to inspect and how specific visual traits prune candidate categories (Figure 1(a)). We use *Flora-De / Blumen in Schwaben* for Germany (Meyer, n.d.), and *Flora of Australia* with regional and specialist KeyBase projects for Australia (Thiele & Klazenga, 2016). We connect these resources to specimens through three steps.

Unified dKey graph construction. We convert German HTML pages and Australian API records into directed graphs linking entries, couplets, leads, and terminal taxa. The graphs preserve transitions within and across keys, enabling retrieval of complete identification paths and all sibling leads presented at each decision point.

Nomenclatural harmonization via GBIF. Taxonomic revisions can give the same species different names in specimens and dKeys. To eliminate nomenclature-induced trajectory breaks, we harmonize scientific names with Global Biodiversity Information Facility (GBIF) backbone taxonomy (GBIF Secretariat, 2023), mapping infraspecific taxa to parent species for species-level evaluation.

End-to-end path verification. Source keys commonly feature UI navigational links, cyclic cross-references, or broken transitions that can corrupt automated graph traversal. We trace each candidate path from a family- or genus-level entry, checking that selected leads match the source text, links reach the intended nodes, and the terminal species matches the specimen label. The diagnostic traits along these validated paths are used to construct Trait QA.

3.3 Trait Question Construction

The dKey trajectories establish the authoritative sequence of morphological decisions required to identify each specimen. We use these traits as observation targets for Trait QA, assessing whether models can judge their states from the image when explicitly told which organ and attribute to inspect. Construction comprises the following four stages.

Atomic trait decomposition. A lead may combine several observations, such as *leaves opposite; stems glabrous*. To assess individual visual judgments, we use Qwen3.8-Max (Qwen Team, 2026a) to decompose the source leads into atomic `organ–attribute–state` records (e.g., `leaf–arrangement–opposite`). We align contrasting states of the same organ and attribute across sibling leads, preserving semantic qualifiers and negations. We retain comparable visual traits and exclude non-visual conditions such as geographic distribution and flowering season.

Neutral questions and source-derived options. We instantiate canonical question templates for attributes such as arrangement, shape, and count. Qwen3.8-Max checks each question against the source leads and refines its wording while preserving the target trait. Answer options use the aligned contrasting states plus *Cannot determine from the image* (Unknown) for insufficient visual evidence.

Visual answerability annotation. The validated dKey trajectory provides each item's reference answer, which is retained for Answerable items. Qwen3.8-Max and Kimi-K3 independently answer each Trait QA item under standard visual QA prompts without disclosing the annotation purpose. Model responses are binarized into *Answerable* (any concrete option) or *Unknown*. Disagreements are resolved by GPT-5.6-Sol (OpenAI, 2026) and binary majority vote.

Expert taxonomic audit. To validate question construction and answerability annotation, a plant taxonomy expert (doctoral researcher in systematic botany) independently audited a stratified sample of 200 Trait QA items. Using full-resolution images, the expert checked each question against its source lead and answered it blinded to the study purpose, target taxa, reference answers, and model votes. Expert responses were mapped to Answerable/Unknown and compared with model-derived labels, yielding **97.50%** source faithfulness (195/200) and **83.00%** answerability agreement (166/200). Expert corrections were applied to the audited items. Full construction details, dataset statistics, and task definitions are given in Appendix A.

4 METHOD

Standard multimodal agents equipped with active perception tools can iteratively crop local regions to examine subtle morphological details. However, unconstrained visual exploration often fails in fine-grained botanical identification for two fundamental reasons: (1) agents lack prior knowledge of *which* diagnostic structures discriminate between visually similar taxa (leading to omitted or non-diagnostic inspections), and (2) even when fine details are perceived, agents struggle to systematically map subtle trait states to taxonomic ranks (Section 5.4).

In biological taxonomy, specialists resolve this bottleneck via dichotomous keys (dKeys), which frame identification as an evidence-driven traversal over a decision graph of contrasting morphological states (Tofilski, 2018). Inspired by this taxonomic practice, we propose a unified agentic framework that integrates iterative high-resolution visual inspection with structured dKey guidance. As illustrated in Figure 1(b), our method exposes regional dKeys as interactive environment tools, enabling agents to interleave active visual exploration, persistent trait logging, and key-directed hypothesis pruning. We first formalize the base agentic visual reasoning framework (Section 4.1), and then detail the dKey-guided traversal and identification workflow (Section 4.2).

4.1 AGENTIC VISUAL REASONING WORKFLOW

Fine-grained visual recognition requires dynamic visual exploration: the choice of subsequent inspection targets depends strictly on the morphology observed in previous steps. Following the "thinking with images" paradigm (Zheng et al., 2026), we formulate specimen identification as a tool-mediated, multi-turn reasoning process that interleaves active visual observation with hypothesis refinement within a ReAct-style loop (Yao et al., 2023).

Agent State Representation. Given a high-resolution specimen image I and geographic provenance metadata m, the agent is initialized with a global whole-sheet overview $I^{(0)}$, which is preserved across all interaction steps as an invariant coordinate reference for specimen layout. At step t, the agent's observable state is formalized as:

$$H_t = \big(I^{(0)}, m, \mathcal{C}_t, O_{<t}, \mathcal{H}^{\text{tool}}_{<t}\big), \tag{1}$$

where $\mathcal{C}_t$ denotes a bounded visual buffer holding up to K recent high-resolution crops, $O_{<t}$ represents the chronological trajectory of structured morphological observations recorded prior to step t, and $\mathcal{H}^{\text{tool}}_{<t}$ archives the historical sequence of tool calls and returned execution outputs. Conditioned on H_t and a fixed task instruction P, the policy model π_θ selects an action a_t:

$$a_t \sim \pi_\theta(a \mid H_t, P), \qquad a_t \in \mathcal{A}_{\text{agent}}, \tag{2}$$

where the base action space comprises three specialized tools:

$$\mathcal{A}_{\text{agent}} = \{\texttt{inspect_region}, \texttt{record_observation}, \texttt{submit_report}\}. \tag{3}$$

Dynamic Visual Evidence Acquisition ($\mathcal{C}_t$). To resolve micro-morphological traits that are unclear in the global view $I^{(0)}$, the agent invokes `inspect_region` with normalized bounding box $b_j = (x_0, y_0, x_1, y_1) \in [0,1]^4$ and `reason` specifying the target organ or structure. Optional parameters γ_j support localized rotation and contrast/brightness enhancements. Let n_t denote the cumulative number of crops acquired by step t. We maintain the latest K crops in a FIFO buffer $\mathcal{C}_t$:

$$\mathcal{C}_t = \left[\text{Resize}\left(T_{\gamma_j}\left(I[b_j]\right)\right)\right]_{j=\max(1,n_t-K+1)}^{n_t}, \tag{4}$$

where $I[b_j]$ extracts the crop at full resolution, T_{γ_j} applies morphological transformations, and Resize rescales the crop to a maximum dimension of 1,024 pixels. Bounding the buffer to K views prevents visual token explosion while preserving immediate cross-organ comparative context.

Persistent Episodic Trait Memory ($O_{<t}$). To preserve findings from an inspected region b_j, the agent calls `record_observation` at step t to store

$$o_t = \left(\text{organ}, \text{findings}, \text{uncertainty}, \text{confidence}, b_j\right), \tag{5}$$

where findings documents the perceived morphological state, uncertainty logs ambiguities (e.g., whether organs are damaged or missing). These structured textual records persist permanently in $O_{<t}$, retaining cumulative anatomical evidence across the entire inference horizon.

Hierarchical Prediction and Epistemic Abstention (y). Upon gathering sufficient diagnostic evidence, the agent invokes `submit_report` to produce a structured taxonomic prediction:

$$y = \left(y_f, y_g, y_s, \text{rationale}\right), \tag{6}$$

where y_f, y_g, and y_s denote family, genus, and species predictions. The agent is instructed to report supported taxonomic levels and abstain at finer ranks with insufficient evidence by leaving those predictions empty. The report includes the recorded traits and their supporting image regions, with rationale explaining the classification decisions and any abstention.

4.2 dKey-Guided Identification

While the visual reasoning agent in Section 4.1 can actively inspect image regions, it relies on unguided visual search. To inject professional taxonomic criteria into this search process, we expose the regional dKey decision graphs compiled in Section 3 via an expanded action interface:

$$\mathcal{A}_{\text{key-guided}} = \mathcal{A}_{\text{agent}} \cup \{\texttt{consult_key}, \texttt{key_step}\}. \tag{7}$$

We formalize a regional dKey as a directed decision graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Each couplet node $v \in \mathcal{V}$ defines a set of k_v mutually exclusive diagnostic leads, denoted as $\mathcal{L}_v = \{(d_i, \tau_i)\}_{i=1}^{k_v}$, where d_i is a formal morphological condition (e.g., *"stems glandular-hairy"*) and τ_i represents the transition pointer to a successor couplet node $v' \in \mathcal{V}$ or a terminal taxon designation T.

Taxonomic Entry Retrieval (`consult_key`). The agent uses the global specimen overview $I^{(0)}$ and preliminary crops to formulate an initial family- or genus-level hypothesis. It then invokes `consult_key`(taxon, rank) to query the regional dKey library. If the queried taxon is covered, the runtime initializes graph traversal at root node v_0 and returns the contrasting leads $\mathcal{L}_{v_0}$; otherwise, it returns `no_key` and the agent falls back to visual reasoning.

Trait-Guided Stepwise Traversal (`key_step`). At any active couplet v, the contrasting leads in $\mathcal{L}_v$ serve as explicit instructions indicating *what* morphological structures the agent must examine. Instead of guessing, the agent calls `inspect_region` to zoom into the specific structures referenced in $\{d_i\}_{i=1}^{k_v}$. Once sufficient evidence is recorded in $O_{<t}$, the agent selects the supported branch $\hat{i}$:

$$v' = \texttt{key_step}(\hat{i}) = \tau_{\hat{i}}. \tag{8}$$

The runtime updates the tool history $\mathcal{H}_{<t}^{\text{tool}}$ with the leads of the successor node v'. When a family-level trajectory reaches a terminal genus node with its own key, traversal automatically continues from that genus key. Repeated couplet choices are cached for consistency.

Terminal Verification and Guided Abstention. When a branch terminates at a leaf node $\tau_{\hat{i}} = T$, the runtime returns the comprehensive botanical diagnosis of taxon T. The agent conducts a final verification by cross-referencing this description against the specimen image. If a mandatory diagnostic structure in lead d_i is missing, or ambiguous, the agent terminates the key walk and abstains at that rank, reporting the deepest confidently verified parent taxon in `submit_report`.

5 EXPERIMENTS

We organize our experiments around the three questions raised in the introduction. Sections 5.2 and 5.3 jointly address **Q1: Do agents examine the right diagnostic traits and judge them when queried?** and **Q2: Can dKey guidance direct agents toward the relevant traits and their use?** Section 5.4 addresses **Q3: Is dKey guidance bounded by perceptual fidelity or taxonomic deduction?** Appendices C and D further show that a correct genus entry improves species F1 for all nine models, whereas adaptive beam search over alternative key branches yields no consistent gain.

5.1 EXPERIMENTAL SETUP

Benchmark data and evaluated models. We evaluate on the full **HerBench** benchmark: 851 high-resolution specimens across 239 species, paired with 5,325 source-grounded trait questions. We benchmark nine representative agentic models spanning diverse architectural scales and deployment paradigms: seven state-of-the-art open-source models, including Qwen3.5 (Qwen Team, 2026b) (9B, 27B, and 35B-A3B mixture-of-experts), Gemma 4 (Gemma Team, 2026) (E4B-IT, 31B-IT, and 26B-A4B-IT), and Kimi-K3; and two frontier proprietary models, GPT-5.6-Sol and Qwen3.8-Max. Fine-grained Trait QA, report pairing, and reference-trait analyses cover the seven open-source models.

Experimental settings. Trait QA evaluates *trait judgment* with specified observation targets, and we assess *trait selection* by comparing judgments of the same traits between Trait QA and E2E identification reports. For E2E identification, *Baseline* uses our agentic visual reasoning workflow (Section 4.1), and *dKey-guided* adds regional dKey traversal to direct trait selection and evidence use (Section 4.2). To probe *evidence use*, *Reference Traits (± dKey)* supplies reference trait states extracted from each specimen's validated dKey graph, decoupling perceptual error from taxonomic reasoning. The specimen image and visual tools remain available in all E2E settings.

Evaluation metrics. For end-to-end (E2E) identification, we report species coverage, precision, recall, F1, and hierarchical F1 (hF1) (Riehl et al., 2023). Precision measures accuracy among submitted species predictions; hF1 evaluates the family–genus–species path. Trait QA reports accuracy on answerable questions, Unknown questions, and all questions. To quantify trait selection, Section 5.3 tracks the proportions of correct, contradictory (incorrect), and unmentioned (no definite judgment) traits in identification reports among traits accurately answered in Trait QA. Implementation, pairing procedures, and metric formulas are detailed in Appendix B.

5.2 E2E IDENTIFICATION PERFORMANCE AND TRAIT JUDGMENT

E2E identification bottleneck: Frontier agents plateau while open models collapse. As presented in Table 1, E2E identification under the unguided *Baseline* setting proves exceptionally challenging across all evaluated architectures. Frontier proprietary models exhibit comparable but plateaued performance, with species F1 scores ranging from 50.83% to 57.87% (Qwen3.8-Max). Among open-source models, Kimi-K3 demonstrates competitive performance at 50.68% F1. In stark contrast, smaller and medium open-source models (Qwen3.5 and Gemma 4 variants) exhibit a severe performance collapse, scoring between 0.00% and 9.35% species F1. Even for the strongest agent, an F1 of 57.87% underscores that fully unguided, reliable botanical identification remains far beyond the reach of contemporary vision-language agents.

dKey guidance raises species F1 but trades away precision. Compared with Baseline, dKey-guided improves species F1 for eight of the nine models. Qwen3.5-27B improves from 9.35% to 14.20%, and Kimi-K3 from 50.68% to 52.88%. The exception, Gemma4-E4B-IT, abstains on most specimens and makes no correct species prediction. Coverage increases for all nine models, while precision falls for every model with nonzero species F1. For GPT-5.6-Sol, coverage rises by 31.97 percentage points while precision falls by 16.21 points. Agents identify more specimens correctly, but also submit more incorrect species predictions. For Qwen3.5-35B-A3B and Gemma4-26B-A4B, the large increase in incorrect species predictions is accompanied by fewer correct genus predictions. These changes lower hF1 despite the improvement in species F1.

Targeted trait judgment remains an acute perceptual barrier. When the observation target is explicitly specified in Trait QA, accuracy on answerable questions remains surprisingly modest,

Table 1: End-to-end identification and Trait QA on **HerBench**. Ans./Unk./overall: accuracy on answerable questions, Unknown questions, and all questions.

	Model	End-to-end identification					Trait QA acc		
		Setting	Coverage	Precision	Recall	Sp. F1	hF1	Ans.	Unk.	Overall
Proprietary	GPT-5.6-Sol	Baseline	49.35	**76.90**	37.96	50.83	66.01	–	–	–
		dKey-guided	**81.32**	60.69	49.35	54.44	66.19			
	Qwen3.8-Max	Baseline	65.88	72.86	48.00	57.87	69.92	–	–	–
		dKey-guided	70.00	70.92	**49.65**	**58.41**	**70.94**			
Open-source	Kimi-K3	Baseline	56.29	70.35	39.60	50.68	66.48	**64.55**	**90.52**	**72.32**
		dKey-guided	**61.34**	69.54	**42.66**	**52.88**	**67.31**	–	–	–
	Qwen3.5-27B	Baseline	10.58	48.89	5.17	9.35	30.42	57.40	79.41	63.98
		dKey-guided	34.08	27.93	9.52	14.20	31.40	–	–	–
	Qwen3.5-9B	Baseline	1.53	76.92	1.18	2.31	26.45	52.92	76.90	60.10
		dKey-guided	14.45	22.76	3.29	5.75	27.15	–	–	–
	Qwen3.5-35B-A3B	Baseline	2.94	**80.00**	2.35	4.57	34.76	58.12	69.62	61.56
		dKey-guided	32.08	22.34	7.17	10.85	32.10	–	–	–
	Gemma4-31B	Baseline	5.29	24.44	1.29	2.46	17.02	55.14	76.84	61.63
		dKey-guided	15.04	15.62	2.35	4.09	18.66	–	–	–
	Gemma4-26B-A4B	Baseline	15.16	33.33	5.05	8.78	23.54	45.87	89.45	58.91
		dKey-guided	39.60	18.10	7.17	10.27	22.32	–	–	–
	Gemma4-E4B-IT	Baseline	0.00	0.00	0.00	0.00	2.93	25.51	84.37	43.12
		dKey-guided	8.23	0.00	0.00	0.00	4.30	–	–	–

spanning 25.51% to 64.55% across the seven evaluated open models. Even the top-performing model, Kimi-K3, resolves only 64.55% of answerable questions, despite achieving 72.32% overall accuracy bolstered by a 90.52% recognition rate on visually unanswerable (Unknown) questions. This confirms that discerning microscopic morphological states from herbarium sheets remains an acute visual challenge even when the model is explicitly told exactly where and what to inspect.

Rank inversion: Trait accuracy does not guarantee identification success. Crucially, model rankings on isolated Trait QA diverge from their E2E identification performance. Gemma4-31B scores nearly ten points higher than Gemma4-26B-A4B on answerable Trait QA (55.14% vs. 45.87%) but achieves less than one-third of its Baseline species F1 (2.46% vs. 8.78%). Similar rank reversals appear for Qwen3.5-35B-A3B and Qwen3.5-27B, showing that accurate isolated trait judgment does not necessarily translate into stronger identification. This motivates our investigation of how models select and use trait evidence during identification.

5.3 Diagnosing Trait Selection: What to Look For

To uncover why strong trait perceivers stumble in E2E identification, we investigate *trait selection* by linking individual Trait QA judgments with the models' final E2E reports on identical specimens. Specifically, we isolate the subset of answerable traits that each model correctly resolved in Trait QA, and evaluate whether those diagnostic features are accurately reported, misjudged, or completely omitted in its E2E identification rationale (Figure 3).

The trait selection gap: Diagnostic features are widely omitted during unguided reasoning. As illustrated in Figure 3, within the subset of traits judged correctly in direct Trait QA, correct judgments consistently outnumber incorrect ones in the final E2E reports. However, the overwhelming majority of these traits receive *no definite trait judgment* across most models under the Baseline setting. In other words, even when an agent possesses the perceptual capability to discern a diagnostic structure correctly upon request, it routinely fails to actively inspect or report that trait when navigating identification autonomously. This severe *trait selection gap* demonstrates that unguided agents struggle with visual attention allocation—they do not know *which* traits are diagnostically decisive for the specimen at hand.

dKeys serve as structured attention directors. dKey guidance markedly mitigates this selection bottleneck. Across all seven evaluated models, it substantially expands the proportion of correct trait judgments within E2E reports. For Kimi-K3, the share of QA-correct traits correctly reported

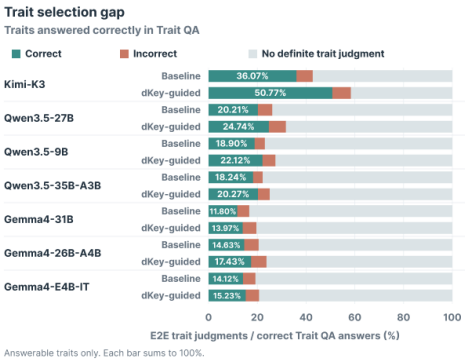

Figure 3: **Trait selection gap.** For answerable traits each model answers correctly in Trait QA, how the same traits are judged in its E2E reports under Baseline vs. dKey-guided.

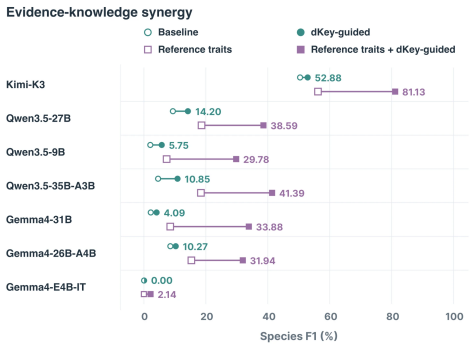

Figure 4: **Evidence-knowledge synergy.** Species F1 under Baseline vs. dKey-guided, without (circles) and with (squares) reference trait states.

during identification rises from 36.07% under Baseline to 50.77% under dKey guidance. By presenting contrasting morphological couplets, dKeys act as an external taxonomic attention mechanism, directing the agent's visual tools toward the structures needed to discriminate candidate taxa.

5.4 Decoupling Trait Perception from Taxonomic Deduction

To determine whether agent failures stem from perceptual errors or defective taxonomic deduction, we conduct controlled decoupling experiments by supplying models with ground-truth trait states extracted along each specimen's validated dKey path. Figure 4 illustrates the impact of providing verified trait evidence with and without dKey guidance.

Morphological evidence alone is insufficient without structured taxonomic deduction. Providing reference trait states improves Baseline species F1 for all seven models (e.g., Qwen3.5-35B-A3B advances from 4.57% to 18.40%). Yet, absolute performance remains surprisingly depressed: six models remain trapped below 20% species F1, with only Kimi-K3 reaching 56.21%. This demonstrates that even when perceptual ambiguity is eliminated and perfect morphological evidence is handed to the agent, models fundamentally struggle to map these visual observations to species. Internal knowledge within modern LLMs is insufficient to execute fine-grained taxonomic deduction.

Evidence-knowledge synergy: The utility of expert guidance scales with perceptual fidelity. When dKey guidance is coupled with verified reference traits, species F1 undergoes a dramatic resurgence across every evaluated architecture. Crucially, the performance gains unlocked by dKey guidance are drastically larger in the presence of reference traits than in their absence. For Kimi-K3, dKey guidance yields a modest 2.20 percentage point gain on raw images, but sparks a remarkable **24.92 percentage point surge** (advancing from 56.21% to 81.13%) when verified trait states are supplied. This striking synergy reveals the underlying mechanics of agentic FGVR: *dKeys provide the necessary deductive logic to navigate complex taxonomic spaces, but their real-world efficacy is fundamentally bounded by the fidelity of the agent's fine-grained visual perception.*

6 Conclusion

We introduced **HerBench**, which grounds high-resolution herbarium specimens and 5,325 trait questions in taxonomist-authored dKeys to diagnose trait selection, trait judgment, and evidence use in multimodal agents. Our analyses reveal a *trait selection gap*: agents often answer isolated trait questions correctly yet leave the same traits unjudged during identification, while targeted trait judgment itself remains difficult (25.51%–64.55% on answerable questions). Our dKey-guided framework improves species F1 for eight of nine models, and its gains grow substantially when accurate reference traits are supplied, showing that the value of expert guidance is bounded by the fidelity of fine-grained visual perception. Future work will explore perceptual verification loops and extend key-guided traversal to open-world biological identification.

AI USE STATEMENT

In this work, we used generative AI tools to assist with translation, language polishing, refinement of formal descriptions in Section 4, and creation or modification of scientific figures. We also used LLMs and MLLMs to support dataset preparation, Trait QA construction, and semantic matching of trait statements in identification reports. These procedures are described in Section 3. Dataset validation included source checks and an expert audit of sampled questions. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We will release **HerBench** and the code for dataset construction, agent inference, and evaluation, as well as the prompts and configurations, concurrently with the arXiv release of this paper. Section 3 describes dataset construction and validation, and Section 4 presents the identification method. Appendix B provides implementation details, report-matching procedures, and metric definitions. The release will preserve source attribution and follow the redistribution terms of the underlying resources.

REFERENCES

Junhan Chen, Zilu Zhou, Yujun Tong, Dongliang Chang, Yitao Luo, and Zhanyu Ma. Seeing as experts do: A knowledge-augmented agent for open-set fine-grained visual understanding. 2026. URL https://arxiv.org/abs/2603.03762.

DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence, 2026. URL https://arxiv.org/abs/2606.19348.

GBIF Secretariat. Gbif backbone taxonomy, 2023. URL https://www.gbif.org/dataset/d7dddbf4-2cf0-4f39-9b2a-bb099caae36c.

Gemma Team. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

Hulingxiao He, Geng Li, Zijun Geng, Jinglin Xu, and Yuxin Peng. Analyzing and boosting the power of fine-grained visual recognition for multi-modal large language models. In *International Conference on Learning Representations*, volume 2025, pp. 89927–89944, 2025. URL https://arxiv.org/abs/2501.15140.

Kimi Team. Kimi K3: Open frontier intelligence, 2026. URL https://arxiv.org/abs/2607.24653.

Svetlana Kiritchenko, Stan Matwin, Richard Nock, and Fazel Famili. Learning and evaluation in the presence of class hierarchies: Application to text categorization. In *Canadian AI*, 2006. URL https://api.semanticscholar.org/CorpusID:15269416.

Nathaniel Lesperance, Sujeevan Ratnasingham, and Graham W. Taylor. Taxonomic reasoning for rare arthropods: Combining dense image captioning and rag for interpretable classification. 2025. URL https://arxiv.org/abs/2503.10886.

Geng Li and Yuxin Peng. FIKA-Bench: From fine-grained recognition to fine-grained knowledge acquisition. 2026. URL https://arxiv.org/abs/2605.13193.

M Maruf, Arka Daw, Kazi S Mehrab, Harish B Manogaran, Abhilash Neog, Medha Sawhney, Mridul Khurana, James P Balhoff, Yasin Bakış, Bahadir Altintas, et al. VLM4Bio: A benchmark dataset to evaluate pretrained vision-language models for trait discovery from biological images. *Advances in Neural Information Processing Systems*, 37:131035–131071, 2024. URL https://arxiv.org/abs/2408.16176.

Thomas Meyer. Flora-DE: Flora von deutschland. Online photographic identification key, n.d. URL https://www.blumeninschwaben.de/.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition, 2026. URL https://openai.com/index/gpt-5-6/.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026a. URL https://qwen.ai/blog?id=qwen3.8.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026b. URL https://qwen.ai/blog?id=qwen3.5.

Kevin Riehl, Michael Neunteufel, and Martin Hemberg. Hierarchical confusion matrix for classification performance evaluation. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 72(5):1394–1412, July 2023. ISSN 1467-9876. doi: 10.1093/jrsssc/qlad057. URL http://dx.doi.org/10.1093/jrsssc/qlad057.

Stephen G. Saupe. Plant identification (featuring taxonomic keys). Lecture notes, BIOL 308 Plant Taxonomy, College of Saint Benedict and Saint John's University, 2009. URL https://www.employees.csbsju.edu/ssaupe/biol308/Lecture/keys.htm. Accessed September 24, 2026.

Kiat Chuan Tan, Yulong Liu, Barbara Ambrose, Melissa Tulig, and Serge Belongie. The herbarium challenge 2019 dataset. 2019. URL https://arxiv.org/abs/1906.05372.

Kevin Thiele and Niels Klazenga. Keybase - teaching old keys new tricks. *Australasian Plant Conservation: journal of the Australian Network for Plant Conservation*, 25:3–4, 2016. URL https://www.biodiversitylibrary.org/part/373656.

Adam Tofilski. Dkey software for editing and browsing dichotomous keys. *ZooKeys*, 735:131–140, 2018. ISSN 1313-2989. doi: 10.3897/zookeys.735.21412. URL https://doi.org/10.3897/zookeys.735.21412.

Robert Turnbull, Emily Fitzgerald, Karen M Thompson, and Joanne L Birch. Hespi: a pipeline for automatically detecting information from herbarium specimen sheets. *BioScience*, 75(8):637—-648, July 2025. ISSN 1525-3244. doi: 10.1093/biosci/biaf042. URL http://dx.doi.org/10.1093/biosci/biaf042.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. July 2011. URL https://authors.library.caltech.edu/records/cvm3y-5hh21.

Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models, 2024. URL https://arxiv.org/abs/2408.15556.

Ziyue Wang, Chi Chen, Fuwen Luo, Yurui Dong, Yuanchi Zhang, Yuzhuang Xu, Xiaolong Wang, Peng Li, and Yang Liu. ActiView: Evaluating active perception ability for multimodal large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7605–7633, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.376. URL https://aclanthology.org/2025.acl-long.376/.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. 2023. URL https://arxiv.org/abs/2312.14135.

Yu Xue, Haoxuan Qu, Zhuoling Li, Yihang Lou, Yan Bai, Hossein Rahmani, and Jun Liu. ToolFG: Towards well-grounded fine-grained image classification. 2026. URL https://arxiv.org/abs/2606.02518.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. 2023. URL https://arxiv.org/abs/2210.03629.

Yuhui Zhang, Alyssa Unell, Xiaohan Wang, Dhruba Ghosh, Yuchang Su, Ludwig Schmidt, and Serena Yeung-Levy. Why are visually-grounded language models bad at image classification? 2024. URL https://arxiv.org/abs/2405.18415.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. DeepEyes: Incentivizing "thinking with images" via reinforcement learning. 2026. URL https://arxiv.org/abs/2505.14362.

A Dataset Construction Details and Tasks

This appendix expands the construction stages summarized in Section 3 and reports the final dataset composition and task definitions.

A.1 Collection Preparation and Label De-Identification

We assembled 2,484 digitized specimens through an institutional collaboration, grouped by collection country: 1,484 from Germany and 1,000 from Australia. Herbarium sheets are digitized at megapixel fidelity and retain their uncompressed acquisition resolution, averaging 66.4 megapixels in the final dataset, so that microscopic diagnostic structures such as leaf pubescence, petal venation, and glandular trichomes remain inspectable.

Mounted herbarium labels frequently carry printed or handwritten taxonomic determinations, which pose a risk of OCR-based text leakage that would let a model identify a specimen without examining its morphology. To remove such non-visual shortcuts while preserving plant structures, we apply a three-stage cascade de-identification protocol:

1. **Automated text-region detection.** We use Hespi (Turnbull et al., 2025), a specialized layout-analysis tool for herbarium sheets, to localize text elements, institutional stamps, and accession metadata for initial masking.
2. **Model-assisted residual auditing.** We prompt Qwen3.8-Max to re-inspect the masked full-sheet scans and flag any faint, fragmented, or partially obscured annotations missed by the automated detector.
3. **Expert visual review.** Human reviewers examine all flagged sheets at full native resolution and apply final pixel-precise masks, ensuring that textual cues are obscured while no anatomical structures are altered.

All edits are restricted to text regions, so plant morphology is preserved throughout.

A.2 dKey Graph Construction and Taxonomic Alignment

We select dKeys according to collection region and taxonomic coverage: *Flora-De / Blumen in Schwaben* for Germany (Meyer, n.d.), and *Flora of Australia* with supplementary regional and specialist projects on KeyBase for Australia (Thiele & Klazenga, 2016). To connect each specimen to a complete dKey path that ends at its recorded species, and then turn the traits along that path into questions, we standardize the two source formats, align their taxonomic names with the collection records, and verify that each species is reachable from the selected dKey.

Unified dKey graph construction. The German resource is organized as linked HTML pages, whereas the Australian resource provides structured API records. We normalize both resources into directed graphs. Starting from each family- or genus-level dKey entry, we perform a breadth-first traversal over all reachable identification steps. We represent each decision point (*couplet*) and each alternative choice (*lead*) as a graph node; transitions to the next couplet or a terminal taxon are represented as directed edges. The traversal follows links across dKeys, including transitions from family-level to genus-level keys. The resulting graph preserves complete identification paths and all sibling leads presented as alternatives at each couplet. This representation allows us to retrieve paths from a selected entry to terminal taxa, together with the sibling leads needed for question construction.

Nomenclatural harmonization via GBIF. The same biological species may be represented by different scientific names across collection records and dKeys because accepted names change through taxonomic revisions, and older synonyms remain in use. We therefore align these names through GBIF, a global biodiversity infrastructure that links scientific names to accepted taxa, synonyms, and taxonomic identifiers (GBIF Secretariat, 2023). We remove authorship strings and formatting variants to obtain canonical scientific names, use GBIF relationships and identifiers to confirm that two names denote the same taxon, and map infraspecific records to their parent species for species-level evaluation. Approximate string similarity is used only to retrieve possible matches. This alignment prevents differences in nomenclature from breaking the link between a specimen and the corresponding dKey endpoint.

End-to-end path verification. The original resources contain interface links as well as identification links, and extraction can also produce missing nodes or incorrect edges. These errors can break identification paths or lead to the wrong terminal taxon. We search from a family- or genus-level entry to the specimen's aligned species and trace each candidate path step by step. We check that every node exists, each selected lead corresponds to the original text, transitions between dKeys reach the intended entry, and the final endpoint matches the specimen's recorded species. We discard paths containing webpage navigation, broken nodes, or links that continue beyond a species endpoint, because these add decisions that are not part of species identification. If several dKeys have the same entry name, we retain a version only when the recorded species is reachable. These checks ensure that each retained specimen's recorded species is reachable from its selected dKey entry through valid identification links.

A.3 Trait Question Generation and Answerability Annotation

Atomic trait decomposition. Before writing questions, we merge couplets repeated in family- and genus-level paths and separate each lead into individual visual judgments. A lead often combines several organs and attributes: *leaves opposite; stems glabrous*, for example, describes both leaf arrangement and stem indumentum. Turning the full lead into one question would mix these observations and require a compound answer. We use Qwen3.8-Max to parse each set of sibling leads using only their source text and convert them into atomic organ–attribute–state records, such as `leaf--arrangement--opposite` and `stem--indumentum--glabrous`. It retains the supporting phrase, negation, and conjunction or disjunction relations. We then align alternative states that describe the same organ and attribute across sibling leads. Comparable visual traits are retained, while conditions such as geographic distribution and flowering season are excluded.

Neutral questions and source-derived options. We define templates for attributes such as arrangement, shape, color, and count, and instantiate them with the extracted organ and attribute. For example, a leaf-arrangement record yields *What is the arrangement of the leaves?* Qwen3.8-Max checks each instantiated question against the original sibling leads and adjusts its wording for fluency without changing the organ or attribute being asked about. Concrete options come directly from the aligned states in those leads. Validation checks that the question and options preserve the source meaning and that the options describe comparable states. Options retain stable identifiers and are shuffled for presentation. Each question includes *Cannot determine from the image* (Unknown), accommodating missing, obscured, or insufficiently resolved structures.

Visual answerability annotation. The validated dKey path provides the concrete reference answer for each Trait QA item, while the specimen image determines whether that answer can be judged visually. Qwen3.8-Max and Kimi-K3 independently answer each complete Trait QA item by selecting a concrete option or Unknown. To avoid an explicit cue to prefer Unknown, their prompts present an ordinary image question without disclosing the annotation purpose. We then map any concrete choice to Answerable and an Unknown response to Unknown. If the two binary labels agree, we retain their consensus; otherwise, GPT-5.6-Sol independently answers the same item, and the binary majority determines the answerability label. This voting process determines visual answerability. The concrete reference answer for an Answerable item remains the trait specified by the validated dKey path.

A.4 Expert Taxonomic Audit

To assess the reliability of question construction and visual answerability annotation, a plant taxonomy expert (a doctoral researcher in systematic botany) independently audited a stratified sample of 200 Trait QA items. The audit evaluated whether the questions preserved their source leads and whether the model-derived Answerable/Unknown labels agreed with human judgments from the specimen images.

The expert received the full-resolution specimen image, the complete Trait QA item with shuffled options, and the source sibling leads. Target taxa, dKey-derived reference answers, and model votes were hidden. For *source faithfulness*, the expert checked whether the organ, attribute, options, and qualifiers correctly expressed the original leads. For *visual answerability*, the expert directly answered the Trait QA item. Concrete responses were mapped to Answerable and Unknown responses to Unknown, and these binary labels were compared with the model-derived labels.

The audit yielded **97.50% source faithfulness (195/200)** and **83.00% visual answerability agreement (166/200)** within the stratified sample. Expert corrections were applied to the audited questions and labels; items outside the audit retained their model-derived answerability labels.

A.5 Dataset Composition and Task Definitions

Construction begins with 2,484 source specimens. Taxonomic alignment and path validation retain 1,467 specimens with exactly one valid path that reaches a species endpoint. Requiring every morphological decision on the path to yield at least one comparable visual-trait question retains 890 specimens. We then remove specimens for which every retained question is labeled Unknown, producing the final shared set of **851 specimens** spanning **239 species**, with **5,325 trait questions**. Table 2 summarizes its composition.

Table 2: **HerBench** statistics.

	Germany	Australia	Total
Specimens / identification instances	616	235	851
Trait questions	3,459	1,866	5,325
Answerable	2,646	1,086	3,732
Unknown	813	780	1,593
Mean image resolution (MP)	63.58	73.80	66.40

Trait QA presents an image, an organ-and-attribute question, and answer options to evaluate targeted trait judgment and recognition of insufficient evidence. **End-to-end identification** asks models to identify the family, genus, and species from the same specimen images.

B Experimental Details

B.1 Implementation Details

Visual input and context. The whole-sheet overview remains in the agent's context throughout the whole conversation. Inspection requests crop the original high-resolution scan and return views with a maximum side length of 1,024 pixels. The context retains the most recent $K = 8$ crops together with accumulated textual observations (Section 4.1). Collection metadata provided for identification contains only the specimen's geographic collection location. Key queries access locally organized regional libraries through a shared tool interface.

Inference settings. We set the maximum number of interaction rounds to 30 and the reasoning budget to 81,920 tokens. The sampling temperature is 1.0, with `top_p` set to 0.95.

Controlled inputs. Reference Traits supplies atomic trait states extracted from the selected leads along the specimen's validated dKey path, alongside the image and visual tools. Reference Traits + dKey-guided applies the method to the same image and traits. Oracle Family and Oracle Genus provide the correct entry and initialize its key. Agents then judge subsequent conditions from the image and select their own branches.

Evaluation coverage. The identification runs cover 851 specimens per model, except for Qwen3.8-Max, which has 850 evaluated specimens. Trait QA covers 5,325 questions per model. Trait QA accuracies use valid answers as their denominators.

B.2 Dichotomous-Key Runtime

Key version selection. The agent supplies a proposed family or genus name and rank to `consult_key`. The lookup prioritizes the specimen-associated key validated during dataset construction when the requested entry name matches its stored accepted name or a verified synonym. Otherwise, it queries the regional index using the agent's requested name and rank. During dataset construction, versions sharing an entry name are retained only when the specimen's recorded species is reachable through a valid path (Section 3.2). This version-selection rule determines which graph is returned for a matching request. An unavailable entry returns `no_key`.

Traversal and repeated couplets. The runtime follows the link associated with the agent-selected lead. When a family-level path reaches a genus with an available genus-level key, traversal automatically resumes at the initial couplet of that key. If an identical complete lead set recurs within the same run, the runtime reuses the earlier branch choice to maintain consistency.

Returned information and terminal checks. Each retrieved couplet presents all competing leads, including attached taxonomic names on leads that terminate at a taxon. At a terminal, the tool returns the taxon name and source description. The identification prompt instructs the agent to compare that description with the specimen and to abstain at unsupported ranks when submitting its report.

B.3 Trait QA–Report Pairing

Pairing and sample selection. Each record links a trait question to an identification report for the same evaluated model and specimen. We perform this pairing separately for Baseline and dKey-guided. Questions whose reference answer is Unknown are excluded. Figure 3 further selects records answered correctly in Trait QA. The selected records can therefore differ across models.

Semantic matching. DeepSeek-V4-Flash (DeepSeek-AI, 2026) processes each target trait in two stages. First, it extracts the relevant statement from the identification report using the target organ and attribute. The extraction retains the original report text, the stated trait value, and its certainty. The reference answer and the evaluated model's QA result are withheld at this stage. Second, the extracted statement is compared with the reference trait state. Matches require agreement on the target organ and attribute. The retained text spans support inspection of individual matches.

Report judgments and aggregation. A definite reported state is classified as correct when it agrees with the reference and incorrect when it contradicts the reference. The remaining records are assigned no definite trait judgment. For each model and setting, we divide the count in each category by the number of valid paired records answered correctly in Trait QA. The three proportions sum to 100% within each bar. The labels in Figure 3 give the percentage of correct report judgments. Appendix B.4 gives the formula.

B.4 Evaluation Metrics

Species identification. Let N be the number of evaluated specimens with a reference species, A the number receiving a species prediction, and C the number identified correctly to species. We compute

$$\text{Coverage} = \frac{A}{N}, \qquad P = \frac{C}{A}, \qquad R = \frac{C}{N}, \qquad F_1 = \frac{2PR}{P+R} = \frac{2C}{A+N}. \tag{9}$$

Abstentions remain in N and contribute to neither A nor C. Precision is reported as zero when $A = 0$, and F1 is zero when $C = 0$. Table values are expressed as percentages.

Hierarchical identification. hF1 measures overlap between predicted and reference taxonomic paths across family, genus, and species (Kiritchenko et al., 2006; Riehl et al., 2023). Let $\hat{T}_i$ and T_i contain the predicted and reference taxonomic nodes for specimen i. We pool node counts across specimens and compute

$$\text{hF1} = \frac{2\sum_{i=1}^{N} |\hat{T}_i \cap T_i|}{\sum_{i=1}^{N} |\hat{T}_i| + \sum_{i=1}^{N} |T_i|}. \tag{10}$$

This credits correct higher-rank predictions when species identification is incomplete or incorrect. An empty prediction contributes no predicted or matched nodes, while its reference nodes remain in the denominator.

Trait QA. Let Q contain the valid question–answer records for a model. Each record q has a reference answer y_q and a model answer $\hat{y}_q$. Define Q_{ans} as records with a concrete reference state and Q_{unk} as records whose reference answer is Unknown. For any of these sets, exact-match accuracy is

$$\text{Acc}(S) = \frac{1}{|S|} \sum_{q \in S} \mathbf{1}[\hat{y}_q = y_q], \qquad S \in \{Q_{\text{ans}}, Q_{\text{unk}}, Q\}. \tag{11}$$

These three values correspond to Ans., Unk., and All in Table 1. An Unknown response is correct only when the reference answer is also Unknown. Invalid or missing answers are excluded from the valid records.

Report judgment proportions. For model m and setting $s \in \{\text{Baseline}, \text{dKey-guided}\}$, let $M_{m,s}$ contain valid paired records whose reference trait is answerable and whose Trait QA answer is correct. Let $z_{q,s}$ be the report judgment assigned to record q. Let c denote correct, incorrect, or no definite trait judgment. Its proportion is

$$p_{m,s}(c) = \frac{1}{|M_{m,s}|} \sum_{q \in M_{m,s}} \mathbf{1}[z_{q,s} = c]. \tag{12}$$

The three proportions sum to one for each model and setting and are shown as percentages in Figure 3.

C Taxonomic Entry and Identification Scope

Using diagnostic knowledge also requires selecting an applicable key. We examine this requirement by supplying the correct family or genus entry (Table 3). Agents then judge the remaining conditions from the image and select their own branches.

Table 3: Species F1 with model-selected and oracle taxonomic entries. All settings enable dKeys and visual inspection. Oracle settings supply the correct entry and initialize its key. Scores are percentages; Δ is the Oracle Genus gain over dKey-guided in percentage points. The first two models are proprietary.

Model	dKey-guided	Oracle Family	Oracle Genus	Δ
GPT-5.6-Sol	54.44	48.33	65.43	+10.99
Qwen3.8-Max	58.41	60.35	67.97	+9.56
Kimi-K3	52.88	51.85	64.91	+12.03
Qwen3.5-27B	14.20	26.88	51.86	+37.66
Qwen3.5-9B	5.75	21.24	34.91	+29.16
Qwen3.5-35B-A3B	10.85	25.04	41.00	+30.15
Gemma4-31B	4.09	28.53	57.64	+53.55
Gemma4-26B-A4B	10.27	22.19	48.54	+38.27
Gemma4-E4B-IT	0.00	17.71	38.41	+38.41

Correct genus entries substantially improve identification. Oracle Genus improves species F1 over dKey-guided for all nine models, with gains of 9.56–53.55 percentage points. Supplying the genus directs the agent to the relevant key, narrows the taxonomic scope, and shortens the remaining identification path. Together, these changes make the professional knowledge easier to apply within the appropriate group.

The depth of entry guidance matters. Oracle Genus outperforms Oracle Family for every model, whereas family-level guidance alone has mixed effects relative to dKey-guided. A family entry retains the need to distinguish among genera. It can also start higher in the hierarchy than an agent's own genus entry under dKey-guided. Genus guidance focuses the task on distinctions within one genus, producing consistently stronger results. Errors and abstentions still occur within the correct genus, where agents must resolve the remaining trait judgments and classification decisions.

D Adaptive Beam Search with Incomplete Trait Evidence

Motivation. A specimen image may not reveal the trait required at a key step. Our single-path policy stops when no branch has sufficient visual support. We test whether retaining multiple branches allows later observations to resolve such cases. This follows taxonomic key guidance to try both choices when uncertain and verify the resulting identification (Saupe, 2009).

Adaptive beam strategy. The agent rates each branch as *match*, *not-excluded*, *unclear*, *not-visible*, or *contradicts*, scored as 1.0, 0.2, 0, -0.2, and -1.0, respectively. Paths are ranked by their mean

branch and terminal-verification scores. The candidate budget is two paths, temporarily expanded to four when the second- and third-ranked scores differ by at most 0.25. This extension lasts up to two additional rounds. Active paths and terminal candidates share the budget. Destination taxon names are hidden during branch scoring, and terminal candidates require separate verification. Contradiction judgments require recorded crop evidence meeting a confidence threshold; terminal matches have the same requirement. If ambiguity persists, the agent reports the jointly supported higher rank.

We evaluate three models on 851 specimens under model-selected, correct family, and correct genus entries (Table 4).

Table 4: Single-path and adaptive-beam identification on 851 specimens. Scores are percentages. Bold indicates the higher species F1 within each pair.

Model	Strategy	Coverage	Precision	Recall	Sp. F1	hF1
Model-selected entry (dKey-guided)						
Gemma4-31B	Single path	15.04	15.62	2.35	4.09	18.66
	Beam	39.37	14.93	5.88	**8.43**	22.85
Qwen3.5-27B	Single path	34.08	27.93	9.52	**14.20**	31.40
	Beam	35.14	17.73	6.23	9.22	28.53
Kimi-K3	Single path	61.34	69.54	42.66	**52.88**	67.31
	Beam	51.23	74.08	37.96	50.19	66.86
Correct family entry						
Gemma4-31B	Single path	55.70	39.87	22.21	**28.53**	61.92
	Beam	47.00	44.00	20.68	28.14	62.79
Qwen3.5-27B	Single path	56.52	37.21	21.03	**26.88**	60.62
	Beam	47.00	35.75	16.80	22.86	61.69
Kimi-K3	Single path	74.97	60.50	45.36	**51.85**	77.07
	Beam	79.44	58.43	46.42	51.74	77.18
Correct genus entry						
Gemma4-31B	Single path	76.97	66.26	51.00	**57.64**	81.67
	Beam	80.26	60.61	48.65	53.98	80.85
Qwen3.5-27B	Single path	76.73	59.72	45.83	**51.86**	81.43
	Beam	96.24	47.99	46.18	47.07	79.65
Kimi-K3	Single path	91.89	67.77	62.28	**64.91**	88.20
	Beam	72.50	74.07	53.70	62.26	87.58

Broader exploration has different effects on answering behavior. With model-selected entries, Gemma4-31B improves F1 from 4.09% to 8.43%. Coverage rises from 15.04% to 39.37% and recall from 2.35% to 5.88%, while precision remains near 15%. Qwen3.5-27B instead loses 4.98 F1 points: coverage changes little, but precision falls from 27.93% to 17.73%. Kimi-K3 shows a different trade-off. Precision rises from 69.54% to 74.08%, but coverage falls from 61.34% to 51.23%, reducing F1 by 2.69 points.

Correct genus entries do not ensure gains from branching. With the correct genus supplied, F1 falls by 3.66, 4.79, and 2.65 points for Gemma4-31B, Qwen3.5-27B, and Kimi-K3, respectively. For Qwen3.5-27B, coverage rises by 19.51 points, but recall improves by only 0.35 points and precision falls by 11.73 points. Most of the net increase in species predictions is therefore incorrect. Retaining alternative paths does not consistently improve identification, even within the correct taxonomic scope.