Check for updates

Received: 6 August 2025 | Accepted: 7 August 2026

DOI: 10.1111/2041-210x.70403

Methods in Ecology and Evolution

RESEARCH ARTICLE

Context-engineering large language models for scalable structured metadata extraction in ecological literature reviews

Sruthi M. Krishna Moorthy[1,2,3,4] | Man Qi[2] | Alice Rosen[2] | Yadvinder Malhi[3,4] | Roberto Salguero-Gómez[1,2]

[1]Pembroke College, University of Oxford, Oxford, UK

[2]Department of Biology, University of Oxford, Oxford, UK

[3]Environmental Change Institute, School of Geography and the Environment, University of Oxford, Oxford, UK

[4]Leverhulme Centre for Nature Recovery, University of Oxford, Oxford, UK

Correspondence
Sruthi M. Krishna Moorthy
Email: sruthi.krishnamoorthyparvathi@pmb.ox.ac.uk

Funding information
National Centre for Earth Observation, Grant/Award Number: NE/X013766/1; Biotechnology and Biological Sciences Research Council, Grant/Award Number: BB/T008784/1; H2020 Marie Skłodowska-Curie Actions, Grant/Award Number: EP/Z00232X/1

Handling Editor: Russell Dinnage

Abstract

1. Hundreds of thousands of peer-reviewed articles and grey-literature reports appear each year in ecology and conservation, creating a pressing need for scalable synthesis tools. We demonstrate that transformer-based Large Language Models (LLMs), when embedded within a practical, expert-in-the-loop pipeline, offer a reliable path forward requiring no deep AI expertise.
2. In this study, we benchmark nine high ranking LLMs (five standard and four reasoning-enhanced) across four context-aware prompting strategies from Basic zero-shot to advanced Directive formats to extract structured fields (study location, biome, taxa) from abstracts.
3. Our results reveal that while certain models emerge as top performers across the tested fields, no single LLM dominates every task. However, context-rich prompting drives gains across all models. By mapping API latency, token costs and F_1 score, we show how researchers can select the optimal model–prompt pairing for their speed, budget or performance needs. We also provide actionable guidelines for effective context engineering in prompts and validation workflows that build trust in model outputs.
4. To lower the barrier to entry, we also introduce LitStructLLM, a cross-platform desktop application that lets users upload CSV or PDF inputs, tailor prompts and export reliable, schema-compliant outputs without writing code. This work underscores the synergy of human expertise and LLM capabilities, offering a robust, transparent pipeline for efficient, large-scale literature mining in ecology and conservation.

KEYWORDS
evidence synthesis, information extraction, large language models, literature review, prompt engineering, research gaps and trends

1 | INTRODUCTION

Ecology and conservation science are data-driven disciplines that rely on synthesizing research across diverse fields to shape management and policy decisions. Over the past decade, the volume of both peer-reviewed and grey literature in ecology has doubled. While this expansion supports better responses to challenges such as habitat loss and climate change, it also makes manual literature reviews increasingly unscalable. Literature synthesis requires identifying relevant studies, extracting study-level information and integrating findings. However, one critical bottleneck is converting unstructured text into structured metadata. Key information such as

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2026 The Author(s). *Methods in Ecology and Evolution* published by John Wiley & Sons Ltd on behalf of British Ecological Society.

country, biome, taxa and sensors used is embedded in heterogeneous texts and must be transformed into structured formats (e.g. CSV tables) for trend analysis, meta-analysis and evidence synthesis. This step remains labour-intensive and error-prone, motivating the need for scalable, automated extraction workflows.

Automatic retrieval of context-based information from unstructured ecological text has long been a core pursuit in ecology and conservation (Nundloll et al., 2022). Traditional Natural Language Processing (NLP) toolkits such as SpaCy (Vasiliev, 2020) provide off-the-shelf pipelines for named-entity recognition (NER) (e.g. location), and text classification (Qiu et al., 2020), but often struggle with domain-specific vocabulary or novel entity types, requiring task-specific annotation and fine-tuning (e.g. TaxoNERD via transfer learning; Le Guillarme & Thuiller, 2022). Earlier pre-transformer sequence models (e.g. RNN/LSTM architectures) can carry context forward, but had limited ability to capture long-range dependencies (Pascanu et al., 2012). Transformer architectures overcame these limitations by using self-attention to model both local and global context simultaneously (Vaswani et al., 2017). Transformer-based Large Language Models (LLMs), such as the GPT family, can therefore perform zero- or few-shot extraction of structured information without task-specific training data, enabling more robust and generalizable extraction across diverse unstructured scientific texts.

Since the release of GPT-2 in 2019, the LLM landscape has exploded with dozens of models from various providers such as OpenAI, Meta, Google, xAI, DeepSeek, etc. ChatGPT and its use in academia has been at the centre of an active debate regarding the ethics of its usage since its launch, in 2022. This debate has predominantly focused on its use in academic writing (Johnson et al., 2024), coding (Cooper et al., 2024), education (Extance, 2023) and its energy and water consumption (Ren et al., 2024). At the same time, LLMs are transforming scientific workflows, particularly in biomedicine, where they are widely used to extract structured information for evidence synthesis and meta-analysis (Balasubramanian et al., 2025; Dagdelen et al., 2024; Huang et al., 2024; Idnay et al., 2021; Sacoransky et al., 2024). Ecology and conservation science have also begun to see the potential of LLMs for evidence synthesis (Berger-Tal et al., 2024; Reynolds et al., 2024), and this landscape has expanded sharply over the last 2 years. Several studies now show that general-purpose LLMs can extract key ecological entities such as taxa and study location with high accuracy from unstructured texts, in many cases exceeding the performance of task-specific NER pipelines (Castro et al., 2024; Gougherty & Clipp, 2024) and can also infer ecological relationships among entities (Scheepens et al., 2024).

Despite growing interest, evidence for LLM-assisted metadata extraction in ecology remains fragmented. Recent studies have begun evaluating multiple models, prompting approaches and metadata fields within ecological extraction workflows (e.g. Lu et al., 2026). However, important questions remain about how model choice and prompting interact across extraction tasks that differ systematically in contextual ambiguity. Although LLMs differ on general benchmarks (Hendrycks et al., 2020; Ono & Morita, 2024), their outputs may reflect biases and domain gaps (Meyer et al., 2023), which is particularly relevant in ecology where target fields are often implicit and context-dependent (e.g. distinguishing focal taxa or overlapping biome categories). Prompt design is known to influence extraction quality in other domains (Lee et al., 2025; Sivarajkumar et al., 2024) and increasingly in ecology (Lu et al., 2026; Scheepens et al., 2024). Building on this work, we examine prompting across a broader set of models and explicitly test tasks spanning a gradient of ambiguity, including whether reasoning-oriented models reduce sensitivity to prompt design.

To address this gap, we develop an expert-in-the-loop pipeline combining systematic prompt refinement, evaluation across models with varying reasoning capacities selected from high-ranking models on Vals AI benchmarks (Vals AI, 2025; an independent third-party organization that publishes LLM benchmarks) and quantitative validation for structured extraction from ecological texts. Using 500 manually annotated abstracts, we compare nine LLMs from four providers (including reasoning and non-reasoning variants) under four prompting strategies, from Basic to Directive. We evaluate each model–prompt combination using F_1 score alongside API latency and token cost to quantify cost–time–performance trade-offs. This design allows us to provide evidence-backed answers to three applied questions: (1) how strongly model choice influences extraction performance, (2) how prompting strategy alters performance across tasks spanning a gradient of ambiguity and (3) which model–prompt combinations offer the best efficiency–performance balance for real review workflows. Finally, to support reproducible adoption, we release LitStructLLM, a cross-platform desktop application that enables prompt-based metadata extraction from unstructured text across models from different providers without coding (https://zenodo.org/records/18307181).

2 | METHODS

2.1 | Building an LLM-integrated pipeline for extracting data from literature

A typical literature review pipeline in ecology or conservation follows four main steps: (1) defining the research question; (2) searching/collecting publications (e.g. Web of Science, Scopus, or Elicit; Whitfield & Hofmann, 2023); (3) screening relevant studies; and (4) extracting structured information (e.g. location, ecosystem type, outcomes). While LLMs can assist with search and collection (Berger-Tal et al., 2024), we focus here on structured data extraction from scientific text. This step is crucial and challenging, requiring contextual interpretation, consistent categorization, and machine-readable outputs for downstream analyses. Extracted data can then support qualitative reviews or quantitative meta-analyses, informing research, policy and management. LLMs can be integrated into this step to convert unstructured text into structured outputs (e.g. CSV) while keeping domain experts in control (Reynolds et al., 2024). In our pipeline (Figure 1), experts define controlled vocabularies, design prompts for schema-compliant outputs and validate performance

2041210x, 2026, 10, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210x.70403 by NICE, National Institute for Health and Care Excellence, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

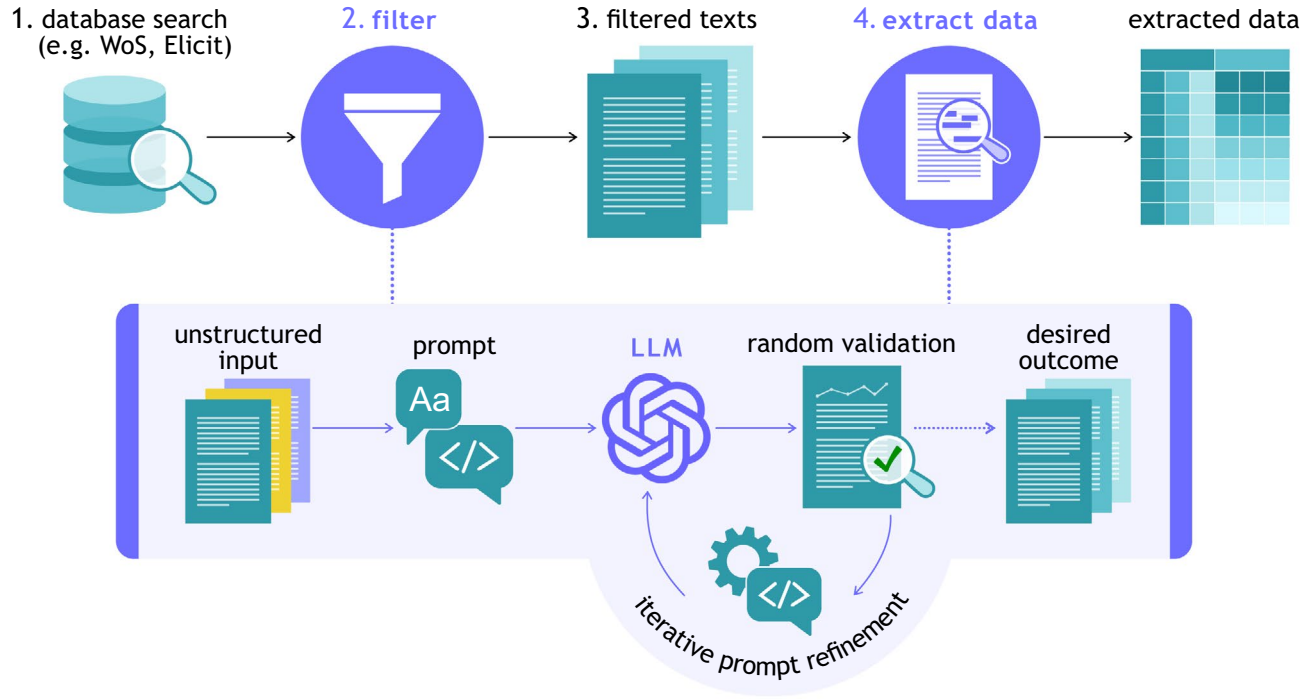


FIGURE 1 Overview of our Large Language Model (LLM)-powered literature review pipeline. After collecting publications via conventional or AI-based databases (e.g. Web of Science), relevant articles are filtered (with or without LLM) before the LLM extracts key information (e.g. study location). An iterative feedback loop with random validation checks and prompt refinements ensures consistent performance and results in a structured final dataset for further analyses.

against a manually annotated reference set. Although we focus on metadata extraction, these structured fields can also support screening and filtering (e.g. excluding non-target taxa, locations or platforms; Figure 1).

2.1.1 | Overview of the pipeline

Our workflow recommends two stages for the data extraction pipeline:

1. *Prompt development*: Use a small, randomly sampled subset to iteratively refine prompts, diagnose systematic errors and stabilize output formatting.
2. *Quantitative validation*: Evaluate the frozen prompt(s) on a larger, independently annotated subset and report accuracy metrics (e.g. F_1 score).

This separation avoids the common confusion between prompt development and model evaluation. Prompt development is used to improve prompts, whereas validation is a quantitative, held-out performance assessment. The prompt development subset should be large enough to identify recurring error patterns, whereas the validation subset should be sufficiently large to provide stable estimates of extraction performance. The exact allocation depends on corpus size, task complexity and observed error patterns.

2.1.2 | Case study dataset ($n = 500$ abstracts)

We evaluated the pipeline on a random sample of 500 abstracts from the autonomous ecosystem monitoring literature (Qi et al., 2025). We used 100 abstracts (20%) for prompt development and the rest of the 400 abstracts (80%) for validation. This allocation reflects our primary aim of assessing LLM performance under a fixed prompt configuration and therefore prioritizes using as much data as possible for held-out evaluation once prompt development is complete. Each abstract was manually annotated for seven fields spanning a gradient from explicit to context-dependent extraction:

1. *Country* of study
2. *Biome* (e.g. Forests, Savannas, Marine, Coastal)
3. *Taxonomic kingdom* monitored (e.g. Animalia, Plantae, Fungi)
4. *Platform* deployed (e.g. drone, ground robot, glider)
5. *Sensor* onboard the autonomous system (e.g. RGB, multispectral, LiDAR, sampler)
6. *Study type* (Methodology vs Ecology)
7. *Application* (Monitoring vs Sampling)

The first three fields were used for the comparative benchmarking of models and prompting strategies, whereas the remaining four fields were reserved for evaluating the transferability of the selected model–prompt configuration. Several of these fields require whole-abstract interpretation rather than keyword matching. For example, sampling is

2041210x, 2026, 10, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210x.70403 by NICE, National Institute for Health and Care Excellence, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

used inconsistently in abstracts and may refer either to physical specimen collection or to non-contact acquisition (e.g. imagery-based 'sampling' of an area), so disambiguation depends on surrounding context. Similarly, sensor mentions may include satellites or in situ devices. Our target label requires identifying the sensor mounted on the autonomous platform rather than merely mentioned comparators.

2.1.3 | Annotation procedure

To reduce subjectivity and potential bias in the reference labels, one researcher completed initial annotations, and a second researcher independently reviewed a subset and flagged disagreements for resolution (e.g. whether coral reefs should be coastal or marine). During prompt development, reviewing LLM's justifications occasionally revealed genuine human annotation mistakes or inconsistencies, which were corrected according to the predefined labeling rules. For example, one UAV disturbance study assessed wildlife responses by comparing field-observed behavioural reactions during ground-based and UAV surveys, independent of the drone's onboard sensors (Krause et al., 2021). Our initial manual label mistakenly implied imaging-based sensing, but the LLM's response has no mention of onboard cameras, and we corrected the annotation to NA. We did not systematically track the number of such corrections, and all adjustments were restricted to the 100 abstracts used for prompt development; the 400 evaluation abstracts were not modified after prompt finalization.

2.1.4 | Prompting via context engineering

We compared four prompting styles spanning a gradient from minimal context specification to highly specified contexts and strong extraction constraints. We refer to this as context engineering, because the key variation is the amount and structure of context provided to the model. We evaluated the effect of increasing contextual guidance on three fields that span a natural difficulty gradient: Country where the study was conducted (highly constrained and explicit), Kingdom of the taxa monitored by the autonomous system (limited to a fixed set of labels (e.g. Animalia, Plantae), yet context-based inferences require careful prompts, such as distinguishing the organism of primary interest from the organism actually monitored by drones/robots.), and Biome where the autonomous system was deployed (semantically ambiguous with overlapping categories, e.g. tropical forest, tropical rainforest). The remaining four manually annotated fields (Platform, Sensor, Study type and Application) were reserved for a later extended end-to-end validation using the selected best-performing model–context configuration (Section 2.2.4).

1. *Basic (Zero-Shot)*: A minimal, one-sentence instruction (e.g. 'Extract the country of study from this abstract'.), representing the simplest out-of-the-box use case most practitioners might try first (Supplementary Section S1).
2. *Categorical*: We constrain the model by listing acceptable output labels (e.g. ISO3 country codes; Kingdom = {Animalia, Plantae, Fungi, …}; Biome = {Forest, Savanna, Marine, Coastal, …}), reducing label-set ambiguity (Supplementary Section S1).
3. *Chain-of-Thought*: We prompt the model to think aloud by laying out the reasoning steps it should follow and by providing one or more worked examples of how to apply the rules, testing whether explicit internal reasoning steps improve extraction fidelity (Supplementary Section S1).
4. *Directive (Iteratively Refined)*: Our most detailed style, built by iteratively refining the prompt, combining clear task definitions, output-format enforcement (clear set of expected labels) and disambiguation rules (Supplementary Section S1).

We explicitly included an NA category only in the structured prompting styles (Categorical, Chain-of-Thought and Directive), where we specify a constrained label set. The Basic zero-shot prompt was intentionally minimal and did not enforce an NA option.

2.1.5 | How was prompt refinement performed?

During refinement on the prompt development subset (100 out of 500 abstracts), we required the model to provide a brief justification for each extracted label. Domain experts reviewed these justifications to identify recurring failure modes (e.g. ambiguous biome boundaries; studied vs merely mentioned taxa; distinguishing onboard sensors from satellites or in situ devices) and revised the prompts by adding more context accordingly. Refinement stopped when successive prompt revisions no longer reduced recurring error patterns on the development subset. Prompt refinement was performed using a single, consistent reference model (gpt 4o-mini) to keep the development process tractable. The finalized prompts were then applied unchanged across all benchmarked models. We therefore treat the refined prompts as a standardized test instrument rather than a per-model optimization strategy. We acknowledge that additional gains may be achievable through model-specific prompt or context engineering, since models differ in capabilities, instruction-following behaviour and sensitivity to contextual framing. However, such tuning was outside the scope of this study, which instead evaluates the cross-model robustness of a shared prompting framework.

2.2 | Model selection and evaluation

2.2.1 | Model selection and settings

In this study, we focus on high-ranking models, including both open-weight and closed models, accessed consistently via model-provider APIs to ensure comparable deployment conditions. We acknowledge that including locally deployable open-weight models would strengthen the reproducibility dimension of this study; however, we do not restrict our evaluation to open-weight models for two reasons.

2041210x, 2026, 10, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210x.70403 by NICE, National Institute for Health and Care Excellence, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

First, the landscape is evolving rapidly, with some previously proprietary models later released as open-weight (e.g. Grok-1, initially released in November 2023 and made available as open weights in March 2024). Second, the most widely used models in practice are typically the highest-performing ones, regardless of openness and our goal is to demonstrate their use for literature review tasks under realistic conditions. By evaluating both open-weight and closed models through a common API interface, we isolate the role of prompting and context engineering rather than deployment differences. Our aim is not to promote specific model types, but to show that structured prompting and validation can improve extraction performance across top-performing models, with principles expected to transfer to other model families.

We applied the four prompting styles across nine high-ranking and reasonably priced reasoning and non-reasoning LLMs from top providers, selected based on their ranks on the MMLU-PRO (Wang et al., 2024) leaderboard reported on vals.ai (Vals AI, 2025) as of 23 May 2025. Among these, two models (DeepSeek variants) belong to open-weight families, while the remainder are closed models. We excluded OpenAI's o3 because it was approximately an order of magnitude more expensive than the next most expensive model considered. The models were:

- *Non-reasoning variants (5)*: OpenAI's GPT-4o mini, GPT-4o, GPT-4.1; Google's Gemini-2.0-flash-001; deepseek-chat
- *Reasoning-optimized variants (4)*: OpenAI's GPT-4o mini-Reasoning, Google's Gemini-2.5-pro-preview-03-25, xAI's grok-3-mini-fast, deepseek-reasoner

This full factorial (9 models ×4 prompt styles) lets us pinpoint (1) how much the choice of model matters, (2) whether native reasoning-enabled architectures perform better than their non-reasoning counterparts and (3) which prompting strategies are most robust across architectures.

Across providers, we controlled the top-level system instruction, keeping it constant across all runs: 'You are an ecological research assistant focused on extracting information from scientific text. Do not fabricate data'. This ensured consistent behaviour across all prompting styles. Where supported, we set temperature = 0 to minimize variability (Windisch et al., 2024), although this was not available for some models (e.g. OpenAI reasoning models and Google Gemini), where defaults were used. Residual non-determinism can persist (Atil et al., 2024); in repeated runs (Supplementary Section S3), variability was low under constrained prompting (~3/50 abstracts) and higher under Basic prompting (~8/50), with differences limited to boundary cases or non-semantic phrasing. Outputs were returned in machine-readable format (e.g. JSON) and exported to CSV for analysis.

2.2.2 | Evaluation metrics, post-processing and cost–time measurements

We treated each field as a multi-label classification task (an abstract may map to multiple biomes or multiple kingdoms). We evaluated performance using the weighted F_1 score (the harmonic mean of precision and recall, weighted by class support) (Hu & Zhou, 2024), which reduces sensitivity to rare classes and strong label imbalance typical of ecological corpora. When prompts did not constrain outputs (only Basic prompting), models produced many semantically similar but non-identical strings (e.g. 'tropical rainforest', 'Amazon forest', 'woodland'). To compute label-based metrics, we normalized these free-form responses to our controlled vocabulary using a two-stage fuzzy-matching pipeline implemented in Python. First, each predicted label was lowercased and stripped, then mapped through a manually curated synonym dictionary that links common variants to canonical biome labels (e.g. 'rainforest', 'tropical forest' to Forests; 'estuary', 'salt marsh' to Coastal; 'reef', 'coral' to Coral reef). If no exact synonym match was found, we applied approximate string matching using *fuzzywuzzy* python package with the partial ratio scorer to compare the candidate string against the list of canonical biome labels. A prediction was assigned to the best-scoring canonical label if its similarity score exceeded a fixed threshold (80); otherwise, it was mapped to NA. For multi-label outputs, we split comma-separated predictions into individual items, mapped each item independently, dropped items mapped to NA and then deduplicated the remaining mapped labels while preserving order. If all items failed to match, the final mapped output was [NA]. This normalization was applied only for the basic prompting style to enable fair computation of F_1. Other prompting strategies explicitly constrain the label set and hence remove the need for post-hoc mapping. For each model–prompt pairing we recorded: (i) API latency (end-to-end response time per abstract under our calling conditions) and (ii) token costs per abstract, enabling a cost–time–performance comparison across all evaluated configurations.

2.2.3 | Statistical analysis

To test whether extraction performance differed across context styles, we treated each model as a paired replicate ($N = 9$) and compared F_1 scores across the four context styles (Basic, Categorical, Chain-of-Thought, Directive) using the Friedman test (Friedman, 1937), a non-parametric repeated-measures analogue of one-way ANOVA. Besides, we conducted planned paired post-hoc comparisons between adjacent context styles (Basic to Categorical, Categorical to Chain-of-Thought, Chain-of-Thought to Directive) using Wilcoxon signed-rank tests (Wilcoxon, 1945). To control the family-wise error rate across the set of post hoc tests within each field, *p*-values were adjusted using the Holm step-down procedure (Holm, 1979).

2.2.4 | End-to-end pipeline validation on additional fields

After identifying the best-performing model–context combination from the three-field benchmark (Country, Biome, Kingdom), we applied that configuration to the remaining four fields (Platform, Sensor,

2041210x, 2026, 10, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210x.70403 by NICE, National Institute for Health and Care Excellence, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Study type, Application) on the full set of 500 abstracts. This also serves as an additional held-out validation where the four remaining fields were manually annotated (one annotator, independently reviewed by a second), with the first 100 abstracts used for prompt development and the remaining 400 used for quantitative validation. These fields are deliberately included because they require contextual interpretation (e.g. distinguishing onboard sensors from satellites; disambiguating 'sampling' versus 'monitoring'; and separating methodological contributions from ecological findings) and therefore test whether the pipeline generalizes for more nuanced fields. The full prompts used for all fields in the directive style are provided in the Supplementary Materials (Section S2).

2.3 | Software implementation

Our methods are implemented in LitStructLLM, a user-friendly, open-access, cross-platform desktop application (Krishna Moorthy et al., 2026). LitStructLLM allows users to process directories of PDFs or CSV files of abstracts, customize extraction prompts via an intuitive graphical interface and export structured outputs directly to CSV for analysis. Compared to commercial tools, LitStructLLM offers two main advantages. First, scalability and cost-effectiveness: it is model-agnostic and supports any API-accessible model from four providers (OpenAI, Google, DeepSeek and xAI), with throughput limited only by API quotas rather than fixed publication caps or extraction limits. This includes both closed models and models from open-weight families (e.g. DeepSeek and xAI's Grok-1), where available via these providers. Second, transparency and reproducibility: all user-defined and system-generated prompts are explicitly recorded, enabling auditability, customization and reproducible data extraction without requiring programming expertise.

3 | RESULTS

Figure 2 summarizes weighted F_1 scores for each model across the four context styles (Basic, Categorical, Chain-of-Thought, Directive) and all three extraction fields (Country, Kingdom, Biome) and is referenced throughout the Results.

3.1 | How critical is the choice of model for structured data extraction?

Model choice affects the final F_1 score, but differences among the strongest models are modest and there is no single clear winner (Figure 2 and Supplementary Table S1). Across all nine LLMs, the

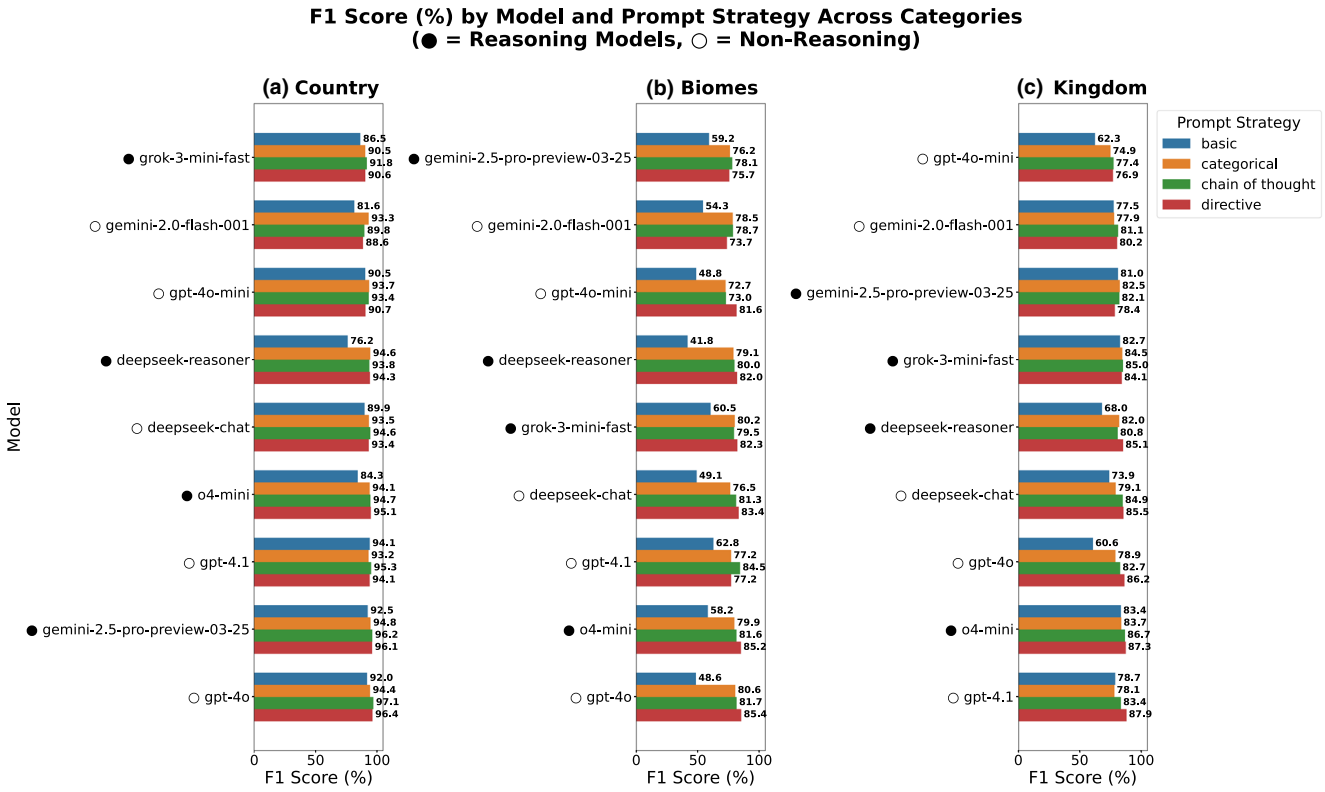


FIGURE 2 Illustrating the model performance across extraction fields. Weighted F_1 scores (%) for nine pay-per-token, API-accessible LLMs evaluated under four context styles (basic, categorical, chain-of-thought, directive) on three manually annotated fields: (a) Country, (b) Biomes and (c) Kingdom. Bars show F_1 for each model–context pairing within a field. Filled circles denote reasoning-optimized variants and open circles denote non-reasoning variants.

2041210x, 2026, 10, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210x.70403 by NICE, National Institute for Health and Care Excellence, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Directive context generally produced the highest or near-highest F_1 score for each field when performance is aggregated across models (Figure 2), reflecting the benefits of explicit label constraints and disambiguation rules. Some models nonetheless achieved their peak F_1 score under the categorical or Chain-of-Thought context, consistent with field- and model-specific sensitivity to how contextual guidance is presented. To provide a consistent, high-performing basis for model comparison, we therefore summarize model performance below under the Directive context and refer to Figure 2 (and Supplementary Table S1) for the full model × context-style results.

Under the Directive context, Country extraction was consistently high across models (range 88%–97%), while the more context-dependent fields exhibited comparatively lower F_1 score range Kingdom (77%–90%) and Biome (77%–86%). The top-performing models were closely clustered. For example, the best three models differed by ~2% F_1 for Biome and ~1% for Kingdom, indicating that several models deliver comparable performance once the context is well specified. GPT-4o achieved the highest Directive F_1 for Country (96%) and Biome (86%), while GPT-4.1 was highest for Kingdom (90%). The weakest performers overall (e.g. DeepSeek-chat, Gemini-2.0-flash and GPT-4o-mini under the Directive context as in Figure 2) lagged primarily on Biome and Kingdom, whereas Country performance was comparatively high across all models. Biome had the lowest best-case F_1 across contexts, with errors concentrated in boundary categories with the most frequent confusions involving the Coastal, Marine and Coral reef labels (see Supplementary Figure S1). In practice, our results indicate that testing a small set of general high-performing models across a few context styles, with output validation, can help identify a high-performing model–prompt configuration for a given task.

3.2 | To what extent do different context engineering strategies impact extraction performance?

Prompt design via context specification emerged as a major driver of extraction F_1 score across all nine LLMs (Figures 2 and 3 and Supplementary Figure S2). When using Basic (zero-shot) prompts without constraining outputs, the raw model responses require substantial post-processing before evaluation. Basic prompting yielded 1673 and 47 unique labels for the biome and kingdom fields, respectively, across 500 abstracts, far exceeding the predefined categories of nine and six labels (see Supplementary Section S6 for the full list of returned labels). To match these responses with our categorical labels, we applied a fuzzy-matching algorithm that mapped each free-form response to the closest biome term in our controlled vocabulary before computing F_1 scores. In contrast, for the Categorical, Chain-of-Thought and Directive contexts, no fuzzy matching or post-processing was performed, and F_1 scores were evaluated directly on the model outputs. Across all validation runs, responses adhered strictly to the predefined label set (including an explicit NA option), with zero labels outside the permitted categories.

We quantified the effect of context style by treating models as paired replicates ($N = 9$) and applying a Friedman test across the four styles. Context style significantly affected performance for all three fields: Country ($\chi^2 = 16.7$, df = 3, $p < 0.005$), Kingdom ($\chi^2 = 16.3$, df = 3, $p < 0.005$) and Biome ($\chi^2 = 19$, df = 3, $p < 0.005$). Planned paired post hoc comparisons used Wilcoxon signed-rank tests with Holm correction (Figure 3). The largest and most consistent improvement came from adding output constraints (Basic to Categorical), which significantly increased F_1 for Country ($p_adj < 0.05$) and Biome ($p_adj < 0.05$). For Biome, adding additional contextual guidance beyond categorical constraints also provided a measurable gain: Categorical to Chain-of-Thought was weakly significant ($p_adj < 0.1$), whereas Chain-of-Thought to Directive was not ($p_adj > 0.1$), indicating diminishing returns once the label space and key disambiguation rules are established. For Kingdom, stepwise improvements from Categorical through Directive were visually consistent across models but only weakly significant ($p_adj < 0.1$) and non-significant for Basic to Categorical ($p_adj > 0.1$), suggesting that gains are more variable across models and may depend on specific inference failure modes rather than formatting alone.

Overall, these results show that structured context specification, especially constraining the output label set and explicitly defining ambiguous categories, accounts for a substantial fraction of achievable performance gains, often exceeding differences attributable to model choice alone. Even with advanced reasoning-enabled model architectures, context engineering remains essential for structured data extraction in ecological texts.

3.3 | What are the cost–time–performance trade-offs practitioners should consider?

Our measurements of end-to-end response time (elapsed time from sending an API request to receiving a model's answer) are averaged over 3–4 days of testing at varying times of day, providing a realistic estimate of per-abstract latency. We pair this with token-based cost estimates to map each model–prompt combination on a time–cost–performance spectrum. We illustrate these trade-offs using Kingdom extraction under the Directive context (Figure 4).

3.3.1 | Latency

Under Directive prompting, response time varied widely across models (≈1.6–101.9 s per abstract; Figure 4). For example, Gemini-2.0-flash-001 averaged 1.58 s/abstract, whereas deepseek-reasoner averaged 101.92 s/abstract. For large-scale screening (e.g. 10,000 abstracts), a difference of a few seconds per abstract translates into hours versus days of clock time depending on parallelization.

2041210x, 2026, 10, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210x.70403 by NICE, National Institute for Health and Care Excellence, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

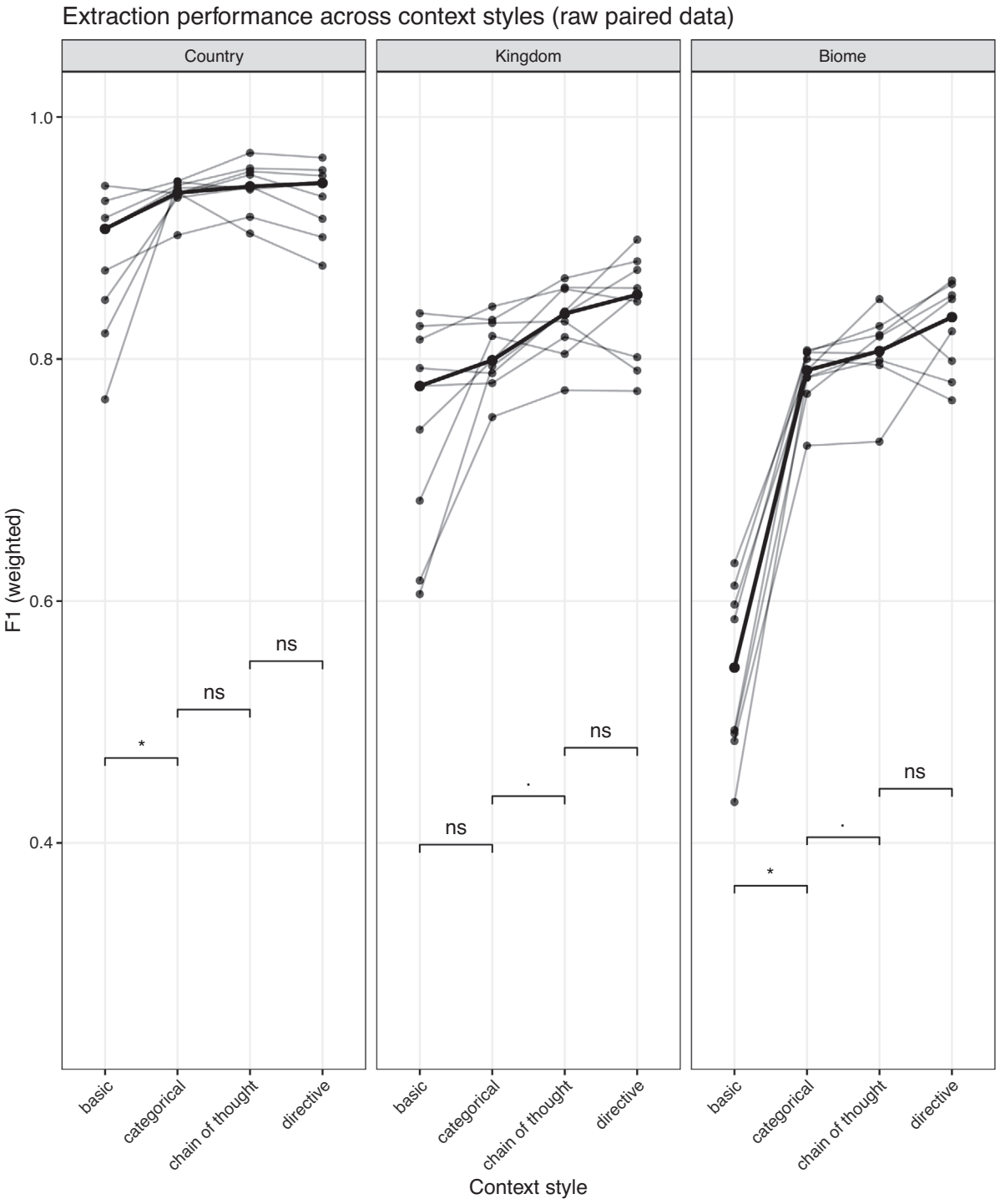


FIGURE 3 Effect of context engineering on extraction performance. Thin grey lines show per-model F_1 scores across context styles (Basic, Categorical, Chain-of-Thought, Directive), thick black points/line show the median across models. Brackets indicate planned paired Wilcoxon signed-rank tests between adjacent context styles with Holm correction (ns, $p \geq 0.1$; ., $p < 0.1$; *, $p < 0.05$).

3.3.2 | Cost

At the high-cost end, GPT-4o not only delivers the best performance under Directive prompts but also incurs the highest token costs (≈$0.0035 per 500-word abstract). In contrast, deepseek-chat sits at the low-cost end (≈$0.0004 per 500-word abstract) and with Directive prompting achieves an average F_1 score of 88% across all three categories.

3.3.3 | Overall balance

When jointly considering time, cost and performance, o4-mini emerges as the winner, offering near–top-tier F_1 scores under Directive prompting style (95% on Country, 86% on Biome, 88% on Kingdom) with moderate latency (7 s) and mid-range costs ($0.002 per 500-word abstract).

2041210x, 2026, 10, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210x.70403 by NICE, National Institute for Health and Care Excellence, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

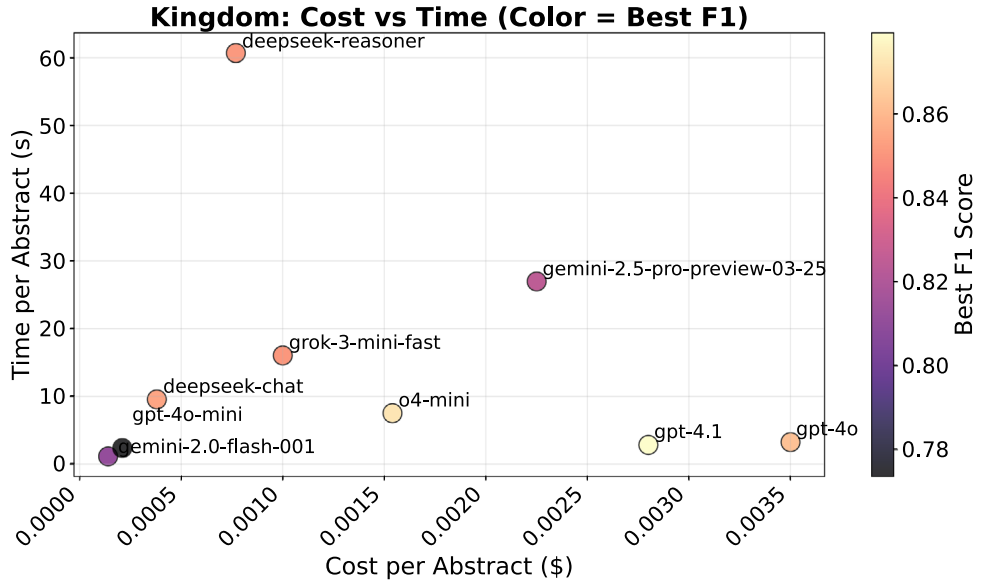


FIGURE 4 Scatterplots for Kingdom extraction with token-based cost (USD per abstract) on the x-axis and end-to-end latency (seconds per abstract) for each LLM on the y-axis. Point colour encodes the model's F_1 score (under directive context), with black indicating the lowest F_1 score and bright yellow indicating the highest. This integrated view highlights how o4-mini achieves top-tier F_1 scores (harmonic mean of precision and recall) while maintaining moderate latency and cost, illustrating its favourable balance of speed, expense and extraction performance.

These trade-offs highlight that the choice of an LLM depends on a project's priorities with respect to speed, budget or maximum performance and underscore the value of our cost–time–performance framework for guiding ecologists in selecting the optimal model–prompt combination for their specific needs.

3.4 | Transfer of the best model–prompt configuration to additional extraction tasks

Following the benchmark on the three primary fields, the recommended model–prompt configuration, selected based on the overall cost–time–performance trade-off (Directive-context OpenAI o4-mini; Figure 4; Supplementary Section S2 for the full prompt) to extract the remaining four context-dependent fields. Study Type (methodological vs ecological) yielded an F_1 of 94%, Application (monitoring vs sampling) 99%, Sensor Used 96% and Platform Deployed 98%. These strong results, even for tasks requiring disambiguation of terms like sampling or identifying which sensor is truly onboard the focal autonomous system, demonstrate that carefully engineered, context-aware prompts can result in reliable extraction of nuanced metadata from unstructured abstracts.

4 | CONTEXT ENGINEERING GUIDELINES

The perfect prompt does not exist. Instead, domain experts (e.g. ecologists, conservation biologists) can improve extraction performance at scale by supplying richer context and refining it iteratively based on observed errors, a practice we refer to as context engineering. Drawing on published prompt engineering best practices (Chen et al., 2025; Lee et al., 2025) and our own systematic experiments, we summarize a set of context engineering guidelines that consistently improved extraction performance from scientific abstracts. These principles focus on structuring the model instructions, specifying acceptable outputs and demanding justification from the model, helping the end user to build trust in the LLM-assisted workflow.

1. Fix the top-level instruction: Use a consistent user-specified system-role instruction (or equivalent 'instructions' parameter) to define the model's role and constraints that explicitly constrains the model to be non-creative (e.g. 'extract only what is supported by the text; do not fabricate'), and hold it constant while varying downstream prompts so differences reflect context design rather than changes in task framing.
2. Control variability when possible: Where supported by the provider, set temperature = 0 (and other determinism-related parameters) to reduce run-to-run stochasticity.
3. Be explicit: Create clear and structured prompts with rich context by using delimiters, breaking down complex requests into manageable steps. Avoid ambiguity by asking for specific output formats or predefined labels for categorical data where possible. Provide explicit instructions. Ask the model to explicitly check for relevant information before generating a response from a given source. This helps in mitigating or even eliminating the issue of model hallucinations (Barkley & van der Merwe, 2024).

2041210x, 2026, 10, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210x.70403 by NICE, National Institute for Health and Care Excellence, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

4. Iteratively refine prompts with context: Test different context configurations and continue to refine the prompts iteratively until the desired outcome is obtained.
5. Demand justification: Request justification for the model's responses along with the prompt, including evidence to support its claims (e.g. Provide justification for each extracted entity, citing the relevant phrase(s) from the text where possible). Also helps in keeping model hallucinations under check (Barkley & van der Merwe, 2024).
6. Request structured output: Whenever possible for easier post-processing of the model's responses, ask for structured output (e.g. key-value formats like JSON, delimiter-separated rows like CSV). We chose JSON (JavaScript Object Notation) as the primary output format due to its robustness for machine parsing and its straightforward conversion to other representations (e.g. CSV) during post-processing.

5 | DISCUSSION

Large language models (LLMs) present an unprecedented opportunity to scale ecological literature reviews enabling automated, context-aware extraction of metadata and insights across thousands of documents at a speed and breadth previously unattainable. A recent study has demonstrated the feasibility of LLM-based metadata extraction in ecological contexts, including across multiple fields and often with post-processing pipelines (Lu et al., 2026). Nevertheless, significant gaps persist. Transparent, consensus-driven workflows for prompt design and output validation remain scarce, and generalist models often overlook niche domain nuances without expert-in-the-loop refinement. Here, we show that transformer-based LLMs make large-scale, context-based literature mining in ecology both feasible and affordable, if workflows prioritize effective prompt design, and that high extraction performance can be achieved with minimal post-processing under structured prompting. While OpenAI's models often lead, no single model dominantly outperforms others, so practitioners should test several top-ranked LLMs (e.g. via vals.ai), including open-weight options, and select the one that best balances cost, speed and performance. Crucially, advanced prompting styles (Chain-of-Thought and Directive format) boost F_1 scores by 10%–27% over basic prompts and are essential for nuanced, context-dependent categories (e.g. focal taxa). To support adoption, we release LitStructLLM, a cross-platform desktop application that enables prompt-based metadata extraction across API-accessible LLMs from multiple providers without coding.

The ability of LLMs to mine thousands of publications at once unlocks several transformative applications for ecology and conservation biology. First, researchers can now map gaps and trends in subfields such as our case study on autonomous ecosystem monitoring at a scale previously unattainable, revealing shifts in methods, focal taxa or geographical regions over time (Qi et al., 2025). Second, by extracting categorical data en masse, LLMs can highlight geographical (Lin et al., 2024) and taxonomic disparities (Gaulke et al., 2019) in research effort, guiding funders and policymakers to understudied regions or species and helping to realign conservation priorities more equitably. Third, LLMs hold promise for evidence synthesis. While our current study focused on abstracts and thus did not attempt to pull quantitative metrics, early work in clinical domains shows LLMs can automate extraction of effect sizes and statistical outcomes from full texts (Gartlehner et al., 2024). More broadly, LLMs may also assist with other stages of literature reviews, such as linking concepts across studies, synthesizing findings and identifying research gaps, once reliable structured extraction and validation workflows are in place.

LLM-based extraction workflows can substantially lower technical barriers to large-scale literature mining by enabling domain researchers to run standardized pipelines without building dedicated NLP models or maintaining ML infrastructure. At the same time, AI can risk reinforcing inequities by concentrating resources and displacing local expertise (Reynolds et al., 2024). To address this, our workflow keeps domain experts in the loop through transparent prompting, justification and validation, ensuring that local knowledge remains central. LLMs also facilitate multilingual extraction, enabling better integration of local-language literature and indigenous knowledge, while emerging multimodal models promise to extend extraction to figures and maps in full-text articles (OpenAI et al., 2023).

As LLMs evolve rapidly, our benchmark should be viewed as a time-stamped snapshot rather than a lasting ranking. The main contribution of our study is the general workflow of context engineering and expert-in-the-loop validation, alongside practical guidelines that consistently improved format compliance and extraction performance. Even reasoning-optimized models showed limited gains under minimal prompting but improved substantially when guided by detailed, justification-driven instructions, underscoring that context engineering remains the key skill for reliable LLM-based extraction in ecology (Chen et al., 2025; Lee et al., 2025; Sivarajkumar et al., 2024). Although hallucinations remain a concern, prompt-guided, text-grounded extraction is less prone to invention, and no such cases were observed under explicit constraints in our study (Mittelstadt et al., 2023).

Context-aware prompt crafting is task-specific, and our guidelines are optimized for API-based workflows rather than web chat interfaces such as ChatGPT. API calls are typically stateless, allowing each abstract to be processed independently with a fixed context, whereas persistent chat sessions may introduce unintended context drift. APIs also allow consistent system role instructions and fixed parameters (e.g. temperature=0) to reduce output variability. Future work could explore reinforcement-style approaches for prompt refinement, iteratively providing examples of correct and incorrect extractions to further calibrate model outputs (Ouyang et al., 2022). If careful prompt engineering remains insufficient, this likely reflects domain-specific complexity that generalist LLMs struggle to capture. In such cases, users may either adopt more capable (and costly) models or fine-tune pretrained transformers on task-specific annotated data, trading lower inference cost and higher accuracy against added annotation and training effort (Anisuzzaman et al., 2025; Lu et al., 2025).

Even expert annotators are prone to error, a point underscored when LLM-generated justifications led us to revise some of our

2041210x, 2026, 10, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210x.70403 by NICE, National Institute for Health and Care Excellence, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

own labels during prompt refinement. Manual extraction from scientific texts is known to have nontrivial error rates, particularly as fatigue accumulates (Gartlehner et al., 2024). In contrast, our best model–prompt pairing processed abstracts in ~7 s, versus 1–1.5 min for human annotation, shifting the expert's role from extractor to verifier. By focusing on spot-checking justifications and correcting edge cases, domain specialists can reduce overall error burden while substantially increasing efficiency. We framed extraction as a multi-label classification task and evaluated performance using the F_1 score (Hu & Zhou, 2024). For tasks involving continuous values, metrics such as MAE or RMSE may be more appropriate. Ultimately, metric choice should reflect the real-world costs of errors in the intended workflow (Anisuzzaman et al., 2025).

Traditional rule-based or ML-driven NLP pipelines remain effective for straightforward tasks. Examples include listing every taxon name mentioned (Le Guillarme & Thuiller, 2022) or geocoding explicit place names (Vasiliev, 2020), where context is minimal and well-defined lexicons exist. In contrast, when distinction between 'mentioned' versus 'studied' taxa matters, or when extracting specialized fields like sensor types or methodological details for which no off-the-shelf models exist, LLMs' contextual reasoning and zero-/few-shot adaptability make them far more robust (Zhao et al., 2023). As ecological research continues to diversify in techniques and metadata requirements, LLMs will increasingly fill gaps left by conventional tools, enabling flexible, domain-agnostic extraction of complex, context-dependent information.

Among the high-ranking models evaluated here, only two are open-weight, with the remainder being closed models, which limits openness and reproducibility. Closed models are typically accessed through model-specific providers, whereas open-weight models may also be accessed via third-party platforms (e.g. groq), reducing single-vendor reliance. However, these platforms still control model availability, with older versions being deprecated and replaced by newer ones. Consequently, API-based access, whether for open-weight or closed models, remains externally controlled and does not ensure reproducibility. Full reproducibility requires local deployment, where model weights and execution environments can be fixed (Farrell et al., 2024; Wihl et al., 2025). Accordingly, the absence of locally deployed models in our evaluation represents a limitation for reproducible benchmarking. Smaller, locally deployable models are therefore an important direction for improving reproducibility, although evidence on their performance remains mixed (Sinha et al., 2025). While LLMs enable rapid, large-scale metadata extraction, they carry substantial environmental costs (Alzoubi & Mishra, 2024), highlighting the value of smaller, domain-specific models developed with AI specialists (Reynolds et al., 2024). Simple, context-aware prompting can deliver accurate results, while future efforts can focus on lightweight, task-specific open models (e.g. BERT-based geocoder like Mordecai) for efficient extraction of features such as locations and biomes (Halterman, 2023). Ultimately, combining domain expertise with open LLM-based workflows can increase the rigour, throughput and reproducibility of ecological literature reviews.

AUTHOR CONTRIBUTIONS

Sruthi M. Krishna Moorthy, Yadvinder Malhi and Roberto Salguero-Gómez conceived the ideas for the study; Sruthi M. Krishna Moorthy and Man Qi collected and processed the data; Sruthi M. Krishna Moorthy analysed the data; Sruthi M. Krishna Moorthy and Alice Rosen designed the figures; Alice Rosen tested the codes and software; Sruthi M. Krishna Moorthy led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

ACKNOWLEDGEMENTS

Sruthi M. Krishna Moorthy was funded by the joint Smithsonian-Pembroke research fellowship. Man Qi was supported by a MSCA fellowship (EP/Z00232X/1) hosted by Roberto Salguero-Gomez. Alice Rosen was supported by the Oxford BBSRC Doctoral Training Partnership (BB/T008784/1). Yadvinder Malhi was supported by the Jackson Foundation. Roberto Salguero-Gómez was supported by a NERC Pushing the Frontiers grant (NE/X013766/1). While OpenAI's GPT-4o mini, a large language model, is used for analysing the data, it was not used for writing or generating the codes used in the manuscript.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/2041-210x.70403.

DATA AVAILABILITY STATEMENT

All the data, codes, the cross-platform desktop application together with its source code are made available via the following zenodo repository: https://zenodo.org/records/18317108 (Krishna Moorthy et al., 2026).

ORCID

Sruthi M. Krishna Moorthy https://orcid.org/0000-0002-6838-2880
Man Qi https://orcid.org/0000-0001-8437-9959
Alice Rosen https://orcid.org/0000-0002-4808-8318
Roberto Salguero-Gómez https://orcid.org/0000-0002-6085-4433

REFERENCES

Alzoubi, Y. I., & Mishra, A. (2024). Green artificial intelligence initiatives: Potentials and challenges. *Journal of Cleaner Production*, *468*, 143090.

Anisuzzaman, D. M., Malins, J. G., Friedman, P. A., & Attia, Z. I. (2025). Fine-tuning large language models for specialized use cases. *Mayo Clinic Proceedings. Digital Health*, *3*(1), 100184.

Atil, B., Aykent, S., Chittams, A., Fu, L., Passonneau, R. J., Radcliffe, E., Rajagopal, G. R., Sloan, A., Tudrej, T., Ture, F., Wu, Z., Xu, L., & Baldwin, B. (2024). Non-determinism of 'deterministic' LLM settings. *arXiv [cs.CL]*. http://arxiv.org/abs/2408.04667

2041210x, 2026, 10, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210x.70403 by NICE, National Institute for Health and Care Excellence, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Balasubramanian, J. B., Adams, D., Roxanis, I., de Gonzalez, A. B., Coulson, P., Almeida, J. S., & García-Closas, M. (2025). Leveraging large language models for structured information extraction from pathology reports. *Journal of Pathology Informatics*, *19*, 100521.

Barkley, L., & van der Merwe, B. (2024). Investigating the role of prompting and external tools in hallucination rates of large language models. *arXiv [cs.CL]*. http://arxiv.org/abs/2410.19385

Berger-Tal, O., Wong, B. B. M., Adams, C. A., Blumstein, D. T., Candolin, U., Gibson, M. J., Greggor, A. L., Lagisz, M., Macura, B., Price, C. J., Putman, B. J., Snijders, L., & Nakagawa, S. (2024). Leveraging AI to improve evidence synthesis in conservation. *Trends in Ecology & Evolution*, *39*(6), 548–557.

Castro, A., Pinto, J., Reino, L., Pipek, P., & Capinha, C. (2024). Large language models overcome the challenges of unstructured text data in ecology. *Ecological Informatics*, *82*, 102742.

Chen, B., Zhang, Z., Langrené, N., & Zhu, S. (2025). Unleashing the potential of prompt engineering for large language models. *Patterns (New York, N.Y.)*, *6*(6), 101260.

Cooper, N., Clark, A. T., Lecomte, N., Qiao, H., & Ellison, A. M. (2024). Harnessing large language models for coding, teaching and inclusion to empower research in ecology and evolution. *Methods in Ecology and Evolution*, *15*(10), 1757–1763.

Dagdelen, J., Dunn, A., Lee, S., Walker, N., Rosen, A. S., Ceder, G., Persson, K. A., & Jain, A. (2024). Structured information extraction from scientific text with large language models. *Nature Communications*, *15*(1), 1418.

Extance, A. (2023). ChatGPT has entered the classroom: How LLMs could transform education. *Nature*, *623*(7987), 474–477.

Farrell, M. J., Le Guillarme, N., Brierley, L., Hunter, B., Scheepens, D., Willoughby, A., Yates, A., & Mideo, N. (2024). The changing landscape of text mining: A review of approaches for ecology and evolution. *Proceedings of the Royal Society B: Biological Sciences*, *291*(2027), 20240423.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, *32*(200), 675–701.

Gartlehner, G., Kahwati, L., Hilscher, R., Thomas, I., Kugley, S., Crotty, K., Viswanathan, M., Nussbaumer-Streit, B., Booth, G., Erskine, N., Konet, A., & Chew, R. (2024). Data extraction for evidence synthesis using a large language model: A proof-of-concept study. *Research Synthesis Methods*, *15*(4), 576–589.

Gaulke, S., Martelli, E., Johnson, L., Letelier, C. G., Dawson, N., & Nelson, C. R. (2019). Threatened and endangered mammals of Chile: Does research align with conservation information needs? *Conservation Science and Practice*, *1*(9), e99.

Gougherty, A. V., & Clipp, H. L. (2024). Testing the reliability of an AI-based large language model to extract ecological information from the scientific literature. *npj Biodiversity*, *3*(1), 13.

Halterman, A. (2023). Mordecai 3: A neural geoparser and event geocoder. *arXiv [cs.CL]*. http://arxiv.org/abs/2303.13675

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2020). Measuring massive multitask language understanding. *arXiv [cs.CY]*. http://arxiv.org/abs/2009.03300

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70.

Hu, T., & Zhou, X.-H. (2024). Unveiling LLM evaluation focused on metrics: Challenges and solutions. *arXiv [cs.CL]*. http://arxiv.org/abs/2404.09135

Huang, J., Yang, D. M., Rong, R., Nezafati, K., Treager, C., Chi, Z., Wang, S., Cheng, X., Guo, Y., Klesse, L. J., Xiao, G., Peterson, E. D., Zhan, X., & Xie, Y. (2024). A critical assessment of using ChatGPT for extracting structured data from clinical notes. *npj Digital Medicine*, *7*(1), 106.

Idnay, B., Dreisbach, C., Weng, C., & Schnall, R. (2021). A systematic review on natural language processing Systems for Eligibility Prescreening in clinical research. *Journal of the American Medical Informatics Association: JAMIA*, *29*(1), 197–206.

Johnson, T. F., Simmons, B. I., Millard, J., Strydom, T., Danet, A., Sweeny, A. R., & Evans, L. C. (2024). Pressure to publish introduces large-language model risks. *Methods in Ecology and Evolution*, *15*(10), 1771–1773.

Krause, D. J., Hinke, J. T., Goebel, M. E., & Perryman, W. L. (2021). Drones minimize Antarctic predator responses relative to ground survey methods: an appeal for context in policy advice. *Frontiers in Marine Science*, *8*, 648772.

Krishna Moorthy, S. M., Qi, M., Rosen, A., Malhi, Y., & Salguero-Gómez, R. (2026). LitStructLLM: A GUI-based multi-platform application for data extracting using off-the-shelf LLMs [computer software]. *Zenodo*. https://doi.org/10.5281/zenodo.18317108

Le Guillarme, N., & Thuiller, W. (2022). TaxoNERD: Deep neural models for the recognition of taxonomic entities in the ecological and evolutionary literature. *Methods in Ecology and Evolution*, *13*(3), 625–641.

Lee, Y., Oh, J. H., Lee, D., Kang, M., & Lee, S. (2025). Prompt engineering in ChatGPT for literature review: Practical guide exemplified with studies on white phosphors. *Scientific Reports*, *15*(1), 15310.

Lin, Z., Chen, A., Wang, X., Liu, Z., & Piao, S. (2024). Large language models reveal big disparities in current wildfire research. *Communications Earth & Environment*, *5*(1), 1–6.

Lu, W., Luu, R. K., & Buehler, M. J. (2025). Fine-tuning large language models for domain adaptation: Exploration of training strategies, scaling, model merging and synergistic capabilities. *npj Computational Materials*, *11*(1), 1–43.

Lu, Z., van der Plas, T. L., Parinaz Rashidi, W., Kissling, D., & Athanasiadis, I. N. (2026). Flexible metadata harvesting for ecology using large language models. In W.-T. Balke, K. Golub, Y. Manolopoulos, K. Stefanidis, Z. Zhang, T. Aalberg, & P. Manghi (Eds.), *New trends in theory and practice of digital libraries*. Communications in computer and information science (pp. 338–352). Springer Nature Switzerland.

Meyer, J. G., Urbanowicz, R. J., Martin, P. C. N., O'Connor, K., Li, R., Peng, P.-C., Bright, T. J., Tatonetti, N., Won, K. J., Gonzalez-Hernandez, G., & Moore, J. H. (2023). ChatGPT and large language models in academia: Opportunities and challenges. *Biodata Mining*, *16*(1), 20.

Mittelstadt, B., Wachter, S., & Russell, C. (2023). To protect science, we must use LLMs as zero-shot translators. *Nature Human Behaviour*, *7*(11), 1830–1832.

Nundloll, V., Smail, R., Stevens, C., & Blair, G. (2022). Automating the extraction of information from a historical text and building a linked data model for the domain of ecology and conservation science. *Heliyon*, *8*(10), e10710.

Ono, K., & Morita, A. (2024). Evaluating Large Language Models: ChatGPT-4, Mistral 8x7B, and Google Gemini Benchmarked against MMLU. *TechRxiv*. https://doi.org/10.36227/techrxiv.170956672.21573677/v

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., … Zoph, B. (2023). GPT-4 technical report. *arXiv [cs.CL]*. http://arxiv.org/abs/2303.08774

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *arXiv [cs.CL]*. http://arxiv.org/abs/2203.02155

Pascanu, R., Mikolov, T., & Bengio, Y. (2012). On the difficulty of training recurrent neural networks. *arXiv [cs.LG]*. http://arxiv.org/abs/1211.5063

Qi, M., Gadd, M., De Martini, D., Davis, K. J., Xiong, B., Rosen, A., Krishna Moorthy, S. M., Hawes, N., & Salguero-Gómez, R. (2025). Biodiversity research requires more Motors in air, water and on land. *Methods in Ecology and Evolution*, *17*, 668–682. https://doi.org/10.1111/2041-210x.70217

Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., & Huang, X. (2020). Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, *63*, 1872–1897. https://doi.org/10.1007/s11431-020-1647-3

2041210x, 2026, 10, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210x.70403 by NICE, National Institute for Health and Care Excellence, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Ren, S., Tomlinson, B., Black, R. W., & Torrance, A. W. (2024). Reconciling the contrasting narratives on the environmental impact of large language models. *Scientific Reports*, *14*(1), 26310.

Reynolds, S. A., Beery, S., Burgess, N., Burgman, M., Butchart, S. H. M., Cooke, S. J., Coomes, D., Danielsen, F., Di Minin, E., 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., … Sutherland, W. J. (2024). The potential for AI to revolutionize conservation: A horizon scan. *Trends in Ecology & Evolution*, *40*, 191–207.

Sacoransky, E., Kwan, B. Y. M., & Soboleski, D. (2024). ChatGPT and assistive AI in structured radiology reporting: A systematic review. *Current Problems in Diagnostic Radiology*, *53*(6), 728–737.

Scheepens, D., Millard, J., Farrell, M., & Newbold, T. (2024). Large language models help facilitate the automated synthesis of information on potential Pest controllers. *Methods in Ecology and Evolution*, *15*(7), 1261–1273.

Sinha, N., Jain, V., & Chadha, A. (2025). Are small language models ready to compete with large language models for practical applications? In *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)* (pp. 365–398). Association for Computational Linguistics.

Sivarajkumar, S., Kelley, M., Samolyk-Mazzanti, A., Visweswaran, S., & Wang, Y. (2024). An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing: Algorithm development and validation study. *JMIR Medical Informatics*, *12*, e55318.

Vals AI. (2025). Vals AI. https://www.vals.ai/benchmarks

Vasiliev, Y. (2020). *Natural language processing with python and spaCy: A practical introduction*. No Starch Press.

Vaswani, A., Shazeer, N., & Parmar, N. (2017). *Attention is all you need*. Paper presented NIPS 2017. Advances in Neural Information Processing Systems 30.

Wang, Y., Ma, X., Zhang, G., Ni, Y., Chandra, A., Guo, S., Ren, W., Arulraj, A., He, X., Jiang, Z., Li, T., Ku, M., Wang, K., Zhuang, A., Fan, R., Yue, X., & Chen, W. (2024). MMLU-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv [cs.CL]*. http://arxiv.org/abs/2406.01574

Whitfield, S., & Hofmann, M. A. (2023). Elicit: AI literature review research assistant. *Public Services Quarterly*, *19*(3), 201–207.

Wihl, J., Rosenkranz, E., Schramm, S., Berberich, C., Griessmair, M., Woźnicki, P., Pinto, F., Ziegelmayer, S., Adams, L. C., Bressem, K. K., Kirschke, J. S., Zimmer, C., Wiestler, B., Hedderich, D., & Kim, S. H. (2025). Data extraction from free-text stroke CT reports using GPT-4o and Llama-3.3-70B: The impact of annotation guidelines. *European Radiology Experimental*, *9*(1), 61.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, *1*(6), 80.

Windisch, P., Dennstädt, F., Koechli, C., Schröder, C., Aebersold, D. M., Förster, R., & Zwahlen, D. R. (2024). The impact of temperature on extracting information from clinical trial publications using large language models. *Cureus*, *16*(12), e75748.

Zhao, X., Ouyang, S., Yu, Z., Wu, M., & Li, L. (2023). Pre-trained language models can be fully zero-shot learners. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 15590–15606). ACL.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Table S1. Model performance across context styles and extraction fields.

Figure S1. Illustrating the per-class error profile for biome extraction.

Figure S2. Panels show the mean F_1 score (harmonic mean of precision and recall) averaged across the three extraction categories (Country, Biome, Kingdom) for each of the nine LLMs under four prompting styles: (a) Basic, (b) Categorical, (c) Chain-of-Thought and (d) Directive.

How to cite this article: Krishna Moorthy, S. M., Qi, M., Rosen, A., Malhi, Y., & Salguero-Gómez, R. (2026). Context-engineering large language models for scalable structured metadata extraction in ecological literature reviews. *Methods in Ecology and Evolution*, *17*, 3006–3018. https://doi.org/10.1111/2041-210x.70403

2041210x, 2026, 10, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210x.70403 by NICE, National Institute for Health and Care Excellence, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License