Using large language models to identify reporting challenges for Target 6 implementation: a short validation against human assessment

Quentin J. Groom* [1], Katelyn Faulkner [2,3], Sabrina Kumschick [4], Aileen Mill [5], Hanno Seebens [6], Tanara Renard Truong [6,7], and Sonia Vanderhoeven [8,9]

[1]Meise Botanic Garden, Meise, Belgium
[2]South African National Biodiversity Institute, Kirstenbosch Research Centre, Cape Town, South Africa
[3]University of Pretoria, Pretoria, South Africa
[4]Centre for Invasion Biology, Stellenbosch University, Stellenbosch, South Africa
[5]School of Natural and Environmental Sciences, Newcastle University, Newcastle upon Tyne, UK
[6]Justus Liebig University, Institute for Bioinformatics, Giessen, Germany
[7]Senckenberg Society for Nature Research, Frankfurt am Main, Germany, Senckenberg Biodiversity and Climate Research Centre
[8]Belgian Biodiversity Platform, Belgium
[9]Université Libre de Bruxelles, Belgium

Preprint – not peer reviewed

Abstract

National reports and related biodiversity policy documents contain important information on implementation barriers, but extracting comparable evidence across many countries is time-consuming. We tested whether large language models (LLMs) could provide a reliable first-pass synthesis of reported challenges relevant to Target 6 of the Kunming–Montreal Global Biodiversity Framework. Human assessor scores for 50 countries were compared with structured outputs from three LLMs: Claude, ChatGPT and Gemini. The aim was not to test exact score reproduction, but to determine whether LLMs captured the same relative patterns in reported challenge categories. Claude showed the strongest alignment with human assessment, especially at the level of challenge-category ranking. Claude-derived scores were therefore used to summarise average challenge patterns across 126 reports. The results suggest that LLM scoring is useful for identifying broad thematic patterns in reporting challenges, but should not be used for precise country-level ranking.

Keywords: biodiversity reporting; Target 6; invasive alien species; large language models; validation; policy synthesis; Global Biodiversity Framework

1 Introduction

The Kunming–Montreal Global Biodiversity Framework includes Target 6 on the management of invasive alien species and their impacts (Convention on Biological Diversity, 2022). National reports and related policy documents contain valuable information on implementation barriers, but extracting comparable evidence across many countries is time-consuming and difficult to standardise.

*Corresponding author: quentin.groom@plantentuinmeise.be

Large language models (LLMs) offer a potential way to support rapid synthesis of such documents, particularly where the aim is to identify broad patterns rather than produce legally or scientifically definitive country assessments.

This study sits within a rapidly developing literature on LLM-assisted text annotation, information extraction and evidence synthesis. In general text-annotation settings, LLMs have shown the potential to reproduce or exceed some forms of human labelling performance (Gilardi, Alizadeh, and Kubli, 2023). However, methodological work also stresses that automated annotation is task-dependent and must be validated against human-generated labels rather than assumed to be reliable (Pangakis, Wolken, and Fasching, 2023; Törnberg, 2024). In conservation and environmental evidence synthesis, AI tools are increasingly discussed as a way to accelerate review workflows, while retaining transparency, reproducibility and human oversight (Berger-Tal et al., 2024). More directly related to the present study, Gougherty and Clipp (2024) compared LLM-based extraction of ecological information with human extraction, finding strong performance for discrete and categorical ecological data but weaker performance for some quantitative fields. Similarly, Spillias et al. (2025) evaluated generative AI tools for qualitative data extraction in community-based fisheries management literature and concluded that they can support, but not fully automate, human-led evidence synthesis.

However, the use of LLMs for biodiversity-policy synthesis requires validation. A key question is whether LLM-derived assessments are sufficiently aligned with human judgement to support exploratory reporting and comparative analysis. In this study, we compared structured LLM assessments of Target 6 reporting challenges with human assessor scores. The purpose was not to determine whether LLM scores exactly reproduced human scores, but to test whether they captured the same relative signal: which challenge categories were generally more or less prominent across national reporting.

We focused on eight challenge categories: data gaps, indicator and methodological problems, technical capacity limitations, financial constraints, institutional coordination barriers, policy and legal limitations, monitoring or information-system weaknesses, and contextual or logistical constraints. We also assessed reporting clarity on challenges. The analysis was designed as a pragmatic validation exercise to determine whether LLM-derived scores could be considered reasonably trustworthy for identifying broad patterns in Target 6 implementation challenges.

2 Methods

2.1 Human assessment

Human assessments were collected using a structured coding template. Each country was assessed by one designated human assessor. Assessors scored the presence or severity of reporting challenges using the ordinal scoring scale in the template. Because only one assessor was available per country, inter-assessor reliability could not be estimated. The human scores were therefore treated as country-level reference judgements rather than as an absolute gold standard.

The human dataset contained 50 country-level assessments. Scores were cleaned and converted into long format, with one row per country and criterion. Missing values in the initial file were subsequently completed. The exported spreadsheet field for total challenge burden was not used because it contained zeros for all countries; instead, total challenge burden was recalculated as the sum of the eight challenge-category scores.

2.2 LLM assessment workflow

Three LLMs were used to score the same challenge categories: Claude (`claude-haiku-4-5`), ChatGPT/OpenAI (`gpt-5-nano`) and Gemini (`gemini-2.5-flash`). These are reported as the API model identifiers used in the extraction scripts at the time of analysis. Model outputs were converted into the same long-table structure as the human assessments, with fields for country, model, criterion and LLM score. Human and LLM scores were joined by country and criterion.

The reporting of the LLM workflow was designed to follow emerging guidance for AI use in environmental evidence synthesis, including justification of the AI system, documentation of where and how it was used, validation against human reviewers, and disclosure of limitations and reproducibility materials (Macura et al., 2025).

The validation dataset contained 450 human–LLM criterion-level comparisons for Claude and ChatGPT, corresponding to 50 countries $\times$ 9 criteria. Gemini contained 441 comparisons because one country, Japan, was missing from the Gemini output.

2.3 Statistical comparison

Because the purpose of the analysis was to evaluate relative trustworthiness rather than exact score reproduction, several complementary measures were used.

At the criterion-score level, we calculated exact agreement, within-one-point agreement, major disagreement rate, mean absolute difference, root mean square error and mean signed difference. Exact agreement measured whether the LLM gave the same score as the human assessor. Within-one agreement measured whether the LLM score differed from the human score by no more than one point. Major disagreement was defined as a difference of two or more points. Mean signed difference was used to assess whether models tended to score higher or lower than human assessors.

To assess whether LLMs preserved relative patterns, we calculated Spearman rank correlations between human and LLM scores at two levels: mean scores by challenge criterion, and country-level total challenge burden. Criterion-level correlations tested whether models identified the same challenge types as generally high or low. Country-level correlations tested whether models ranked countries similarly to human assessors.

After validation, Claude was selected for the main descriptive analysis because it showed the strongest overall alignment with human assessment. Mean Claude scores were calculated for each challenge category across 126 reports. For each mean, we calculated the standard error and a 95% confidence interval using the t distribution:

2.4 Computational workflow and public repository

The computational workflow is available in the public GitHub repository `GBF_IAS_Challenges` (Agentschap Plantentuin Meise, 2026). The repository contains scripts for reviewing CBD Seventh National Reports with a focus on GBF Target 6, scoring reports against a predefined challenge framework, comparing outputs across multiple LLMs, visualising patterns in reporting progress and challenges, and supporting validation using human assessors (Agentschap Plantentuin Meise, 2026). The repository is organised around model-specific extraction and scoring scripts, including `extract.py`, `extract_claude.py` and `extract_gemini.py`, and downstream plotting and summary code in textttfigures.py. The model identifiers were specified directly in these scripts as `gpt-5-nano`, `claude-haiku-4-5` and `gemini-2.5-flash`, allowing the model choice to be inspected from the public workflow.

The code implements a four-stage workflow. First, the extraction scripts start from the CBD online reporting tool page for Seventh National Reports, identify report-page URLs and associated

document identifiers, and retrieve the underlying report JSON through the CBD document API. For each report, the scripts preserve the source JSON, extracted report text and metadata, including country, document identifier, report-page URL, document API URL and the number of extracted characters.

Second, report text is cleaned and prepared for model scoring. The code recursively extracts text from nested JSON objects and removes repeated or empty text fragments. In the Claude workflow, the report text is additionally filtered to retain passages containing terms relevant to Target 6 and implementation challenges, such as invasive alien species, biosecurity, pathway, monitoring, indicator, data, capacity, finance, coordination, law, policy, surveillance, early warning, baseline, inventory, database, methodology and reporting. This step was used to keep the prompt focused on relevant evidence and within model input limits.

Third, each model is prompted with the same challenge-focused coding framework. The prompt instructs the model to score each report from 0 to 3 for eight challenge categories: data gaps, indicator and methodological problems, technical capacity limitations, financial constraints, institutional coordination barriers, policy and legal limitations, monitoring or information-system weaknesses, and contextual or logistical constraints. It also asks for reporting clarity, a short supporting quote or near-quote, and brief notes justifying the score. The prompt instructs models to focus on challenges relevant to reporting or producing Target 6 indicator data, to score only what is supported by the report text, and to ensure that total challenge burden equals the sum of the eight challenge-category scores.

Fourth, model outputs are written as machine-readable CSV files and used for downstream validation and analysis. The validation code standardises country names and criterion labels, reshapes human and LLM scores into a common long format, joins scores by country and criterion, computes agreement statistics and rank correlations, and identifies major disagreements. The plotting workflow then summarises challenge-category means and uncertainty intervals and produces figures for reporting. This structure allowed the same scoring framework to be applied consistently across models and made the human–LLM comparison reproducible.

The challenge categories were specified in the prompt as broad labels rather than as a fully operationalised codebook with definitions, inclusion and exclusion rules, or worked examples. This design was chosen to keep the scoring task simple and transferable across models, but it also means that category boundaries were not fully constrained. Several categories, particularly data gaps, indicator or methodological problems, and monitoring or information-system weaknesses, are conceptually related and may be difficult to distinguish consistently in some reports.

2.5 Use of large language models

Large language models (LLMs) were used both as part of the analytical workflow described in this study and as assistive tools during the preparation of the manuscript. LLMs were used to support the development and debugging of analysis code and to assist with editing, restructuring, and improving the clarity of manuscript text.

All LLM-generated code, analyses, and text were critically reviewed, tested, and, where necessary, modified by the authors. The human authors made all substantive scientific judgments, including decisions concerning study design, analytical methods, interpretation of results, and conclusions, and take full responsibility for the final content of the manuscript. LLM outputs were not treated as authoritative sources of scientific evidence.

3 Results

3.1 Agreement between LLMs and human assessors

Claude showed the strongest overall agreement with human assessor scores. Across 450 criterion-level comparisons, Claude had exact agreement of 26.2%, within-one agreement of 70.0%, and a mean absolute difference of 1.076. ChatGPT had exact agreement of 31.3%, within-one agreement of 58.4%, and a mean absolute difference of 1.287. Gemini had exact agreement of 31.1%, within-one agreement of 56.9%, and a mean absolute difference of 1.306 (Table 1).

All three models tended to score challenges higher than the human assessors. The mean signed difference was +0.422 for Claude, +1.016 for ChatGPT and +0.975 for Gemini. This indicates that LLMs, especially ChatGPT and Gemini, were generally more likely to infer or emphasise implementation challenges than the human assessors.

Table 1: Agreement between LLM scores and human assessor scores.

Model	Comparisons	Exact agreement	Within-one agreement	Major disagreement	Mean absolute difference
Claude	450	26.2%	70.0%	30.0%	1.076
ChatGPT	450	31.3%	58.4%	41.6%	1.287
Gemini	441	31.1%	56.9%	43.1%	1.306

3.2 Relative agreement by challenge category

The strongest validation result was at the level of challenge categories. Spearman rank correlations between human and LLM mean scores by criterion were 0.810 for Claude, 0.765 for Gemini and 0.636 for ChatGPT (Table 2). This suggests that the LLMs, particularly Claude, broadly reproduced the human-assessed relative ordering of challenge types.

In contrast, agreement was much weaker at the country-total level. Spearman rank correlations between human and LLM country-level challenge-burden totals were 0.284 for Claude, 0.057 for ChatGPT and 0.040 for Gemini. This indicates that the LLMs were not reliable for precise ranking of countries by challenge burden.

The validation therefore supports the use of LLM scores for identifying broad patterns in the relative prominence of challenge categories, but not for definitive country-level comparison.

Table 2: Spearman rank correlations between human and LLM scores at criterion and country-total levels.

Model	Criterion-level rank correlation	Country-total rank correlation
Claude	0.810	0.284
Gemini	0.765	0.040
ChatGPT	0.636	0.057

3.3 Average challenge scores from Claude

Because Claude showed the strongest alignment with human assessment, Claude scores were used for the descriptive analysis of average challenge severity across the full set of 126 reports. The highest average challenge score was for data gaps, followed by monitoring or information-system

weakness. Technical capacity limitations, institutional coordination barriers and contextual or logistical constraints also received relatively high average scores. Financial constraints and indicator or methodological problems were intermediate. Policy or legal limitations had the lowest average score (Table 3; Figure 1).

Table 3: Average Claude challenge scores with standard errors and 95% confidence intervals.

Challenge	Mean	SE	95% CI
Data gaps	1.90	0.07	1.77–2.03
Monitoring / information system weakness	1.83	0.06	1.70–1.95
Technical capacity limitations	1.66	0.07	1.52–1.79
Institutional coordination barriers	1.59	0.07	1.45–1.73
Context / logistical constraints	1.50	0.07	1.36–1.64
Financial constraints	1.32	0.11	1.11–1.53
Indicator / method problems	1.17	0.06	1.05–1.30
Policy / legal limitations	0.82	0.07	0.68–0.96

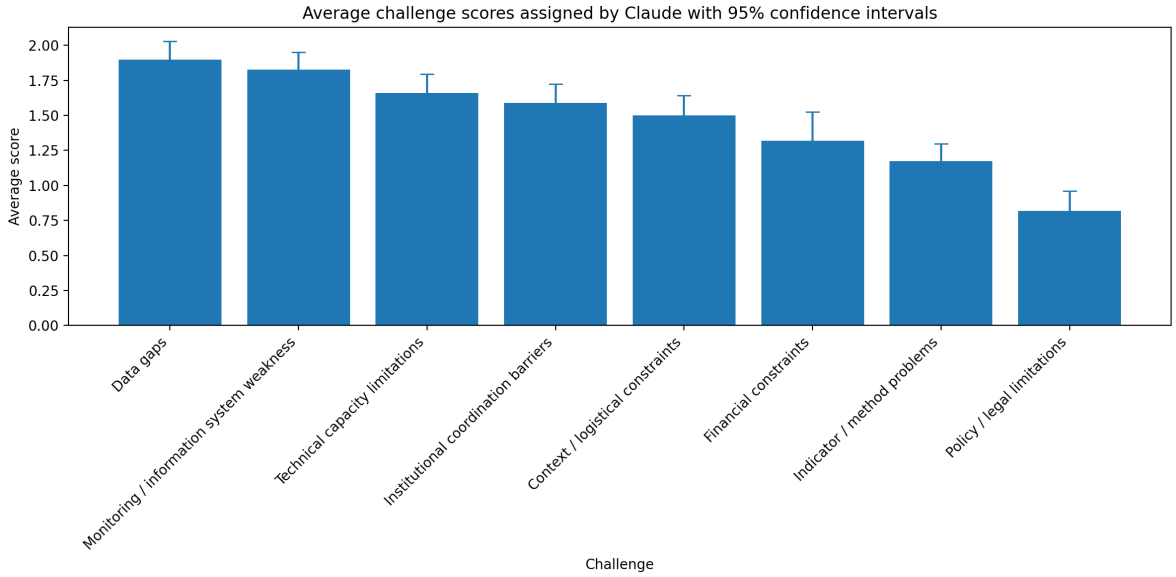

Figure 1: Average challenge scores assigned by Claude across 126 reports, with 95% confidence intervals. Challenge categories are ordered from highest to lowest mean score.

The 95% confidence intervals around the mean Claude scores were relatively narrow, indicating that the estimated average pattern was stable across the analysed reports. The resulting pattern suggests that Target 6 reporting challenges are most strongly associated with evidence, monitoring and capacity limitations, rather than primarily with formal policy or legal barriers.

4 Discussion

This short validation exercise suggests that LLM-based scoring can be useful for rapid synthesis of Target 6 reporting challenges, provided that the results are interpreted at the appropriate level.

This is consistent with the broader evidence-synthesis literature, which suggests that LLMs can assist with structured information extraction and qualitative coding, but that reliability varies by task and requires human validation (Gougherty and Clipp, 2024; Spillias et al., 2025; Pangakis, Wolken, and Fasching, 2023). The main strength of the LLM approach was not exact agreement with individual human scores, but the ability to reproduce broad relative patterns across challenge categories. Claude, in particular, showed strong rank agreement with human assessors at the criterion level.

The results support the use of Claude-derived scores for identifying common types of implementation and reporting challenges. The highest-scoring categories—data gaps, monitoring or information-system weakness and technical capacity limitations—are consistent with the kinds of barriers expected in biodiversity reporting, where implementation often depends on long-term surveillance, taxonomic and ecological expertise, data infrastructure and institutional capacity. The relatively low score for policy or legal limitations suggests that formal policy absence may be less commonly reported than practical barriers to evidence generation, monitoring and implementation.

However, the analysis also shows clear limitations. First, LLMs tended to score challenges higher than human assessors, indicating a tendency toward over-inference. This may occur when models interpret indirect statements, general implementation difficulties or contextual information as evidence of a challenge. Second, agreement was weak for country-level total challenge burden. LLM outputs should therefore not be used to rank countries definitively or to make strong claims that one country faces greater Target 6 challenges than another. Broad groupings or thematic summaries are more defensible than precise country rankings.

A further methodological limitation concerns the design of the scoring prompt. The prompt used broad challenge-category labels and a 0–3 scoring scale, but did not provide detailed definitions, decision rules, or examples for each category. This may have allowed both human assessors and LLMs to interpret category boundaries differently. In particular, evidence of limited monitoring capacity could plausibly be coded as a data gap, a monitoring or information-system weakness, an indicator or methodological problem, or a technical capacity limitation, depending on the assessor's interpretation. Such overlap may have contributed to differences between human and LLM scores and may also affect the interpretation of summed challenge-burden totals. Future applications should use a more fully specified coding manual, including definitions, inclusion and exclusion criteria, examples, counterexamples, and guidance for handling evidence that could fall under more than one challenge category.

A further limitation is the human validation design. Each country was assessed by only one human assessor, so assessor effects and country effects could not be separated. Differences between human and LLM scores may therefore reflect model error, assessor interpretation, ambiguity in the report, or variation in how strictly the scoring scale was applied. This limitation reinforces the need, highlighted in LLM-annotation guidance, to treat human comparison as an explicit validation step rather than as a one-time guarantee of model reliability (Törnberg, 2024; Macura et al., 2025). Future work could improve the validation by including multiple assessors for a subset of countries, adjudicating disagreements, and testing whether LLM performance differs by report length, language, region or document type.

Despite these limitations, the results indicate that LLMs can support biodiversity-policy synthesis when used cautiously. For Target 6 reporting, Claude-derived scores appear reasonably trustworthy for summarising the relative prominence of broad challenge categories. They should be treated as structured, reproducible indicators of reported challenge signals, not as definitive measures of national implementation performance.

5 Conclusion

Comparison with human assessor scores suggests that Claude-derived assessments are sufficiently reliable for identifying broad relative patterns in Target 6 reporting challenges under the broad-label coding framework used here. The strongest support is for analysis at the level of challenge categories, where Claude closely reproduced the human-assessed ranking of common challenges. The evidence does not support precise country-level ranking based on LLM scores alone. Used with appropriate caution, LLM scoring provides a scalable method for exploratory synthesis of biodiversity reporting challenges, especially where the analytical aim is to identify common themes across many national reports.

Acknowledgements

This work was funded by the European Union's Horizon Europe Research and Innovation Programme, B3 (Biodiversity Building Blocks for policy), under grant agreement No. 101059592, and OneSTOP (OneBiosecurity Systems and Technology for People, Places and Pathways), under grant agreement No. 101180559. Views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them. KTF thanks the South African Department of Forestry, Fisheries and the Environment (DFFE) for funding, noting that this publication does not necessarily represent the views or opinions of DFFE or its employees.

Competing interests

The authors declare no competing interests.

Use of generative artificial intelligence

Large language models were used as part of the research methodology as described in the Methods. The authors take full responsibility for the analysis, interpretation, and final manuscript.

Data and code availability

The code used for report extraction, LLM scoring, model comparison and figure generation is publicly available in the GBF_IAS_Challenges repository (Agentschap Plantentuin Meise, 2026). The analysis used a structured Python workflow in which human and LLM score files were cleaned, reshaped into long format, joined by country and criterion, and analysed using agreement statistics, rank correlations and descriptive summaries. Claude average challenge scores were plotted with 95% confidence intervals. The main outputs were a human–LLM comparison workbook, joined comparison tables, model and criterion summaries, a major-disagreement table, and a bar chart of Claude average challenge scores with confidence intervals.

References

Agentschap Plantentuin Meise (2026). *GBF_IAS_Challenges: Scripts to query CBD reports with LLMs*. GitHub repository. Accessed 6 June 2026. URL: https://github.com/AgentschapPlantentuinMeise/GBF_IAS_Challenges.

Berger-Tal, Oded et al. (2024). "Leveraging AI to improve evidence synthesis in conservation". In: *Trends in Ecology & Evolution* 39.6, pp. 548–557. DOI: 10.1016/j.tree.2024.04.007.

Convention on Biological Diversity (2022). *Decision adopted by the Conference of the Parties to the Convention on Biological Diversity: 15/4. Kunming-Montreal Global Biodiversity Framework*. CBD/COP/DEC/15/4.

Gilardi, Fabrizio, Meysam Alizadeh, and Maël Kubli (2023). "ChatGPT outperforms crowd workers for text-annotation tasks". In: *Proceedings of the National Academy of Sciences of the United States of America* 120.30, e2305016120. DOI: 10.1073/pnas.2305016120.

Gougherty, Andrew V. and Hannah L. Clipp (2024). "Testing the reliability of an AI-based large language model to extract ecological information from the scientific literature". In: *npj Biodiversity* 3.13. DOI: 10.1038/s44185-024-00043-9.

Macura, Biljana et al. (2025). *Reporting guidance for the use of artificial intelligence in environmental evidence synthesis*. Collaboration for Environmental Evidence. Accessed 31 May 2026.

Pangakis, Nicholas, Samuel Wolken, and Neil Fasching (2023). *Automated Annotation with Generative AI Requires Validation*. DOI: 10.48550/arXiv.2306.00176.

Spillias, S. et al. (2025). "Evaluating generative AI for qualitative data extraction in community-based fisheries management literature". In: *Environmental Evidence* 14.9. DOI: 10.1186/s13750-025-00362-9.

Törnberg, Petter (2024). "Best Practices for Text Annotation with Large Language Models". In: *Sociologica* 18.2, pp. 67–85. DOI: 10.6092/issn.1971-8853/19461.