LLM Agents as Computational Typologists

Anonymous ACL submission

Abstract

Linguistic typology relies on expert analysis of reference grammars across languages, making large-scale crosslinguistic comparison labor-intensive and unscalable. We introduce AUTOTYPOLOGIST, an LLM agent for evidence-grounded typological analysis over reference grammars. The agent is capable of retrieving relevant grammar sections, analyzing interlinear glossed text (IGT), and iteratively reasoning over typological hypotheses using a ReAct-style workflow. We evaluate the system on TYPOLOGICAL FEATURE CODING against expert annotations and TYPOLOGICAL HYPOTHESIS TESTING with typological universals using 25 open-source reference grammars. Operating under different information constraints in TYPOLOGICAL FEATURE CODING, the agent can synthesize information from reference grammar prose but still faces challenges with only IGTs in the target language. In TYPOLOGICAL HYPOTHESIS TESTING, the agent can synthesize crosslinguistic evidence and identify both supporting cases and counterexamples. These findings suggest that LLM agents can support scalable and inspectable typological analysis, while still requiring expert validation.

1 Introduction

Linguistic typology seeks to uncover linguistic universals of human language (Greenberg, 1963, 2005; Comrie, 1989). Typological knowledge is typically derived from comparing, re-analyzing, and synthesizing crosslinguistic data collected through linguistic fieldwork. Typological research leads to open-sourced typological linguistic databases, such as the World Atlas of Language Structures (WALS) (Dryer and Haspelmath, 2013), the Atlas of Pidgin and Creole Language Structures (APiCS) (Michaelis et al., 2013), Grambank (Skirgård et al., 2023). In these databases, grammar is decomposed into a few hundred individual features, where each language will be assigned a value for its grammatical structure. Traditionally, creating such a database takes years of effort from hundreds of trained linguists, which is not scalable.

Recent research suggests that LLMs are capable of performing meta-linguistic analyses over languages (Şahin et al., 2020; Bean et al., 2024; Begus et al., 2025; Yang et al., 2025b; Schneider and Anastasopoulos, 2026), making them potential general problem solvers in language documentation and linguistic typology.

Motivated by these pioneering works, we present AUTOTYPOLOGIST, an LLM agent that can actively query reference grammars to summarize and compare the grammatical structures of typologically diverse languages, with little human supervision. Our contributions are: First, we introduce the AUTOTYPOLOGIST, the first LLM agent for evidence-grounded typological analysis. Our research opens up new opportunities to large-scale typological analyses, as LLM agents can assist linguists in gathering evidence for hypothesis testing.

Second, we formalize two agentic typology tasks: TYPOLOGICAL FEATURE CODING, where the agent assigns typological feature values and is evaluated against expert annotations, and TYPOLOGICAL HYPOTHESIS TESTING, where it tests typological universals by synthesizing crosslinguistic evidence. AUTOTYPOLOGIST reproduces expert codings with high agreement when reference-grammar prose is available and identifies both supporting evidence and counterexamples in hypothesis testing.

Third, we release a typological dataset built from 25 open-source reference grammars across 14 language families, containing structured grammar chunks, extracted interlinear glossed text (IGT) examples, and metadata for evaluating evidence-grounded typological reasoning.

Our research departs from the established NLP approach of developing task-specific and language-specific models but instead pursues ab-

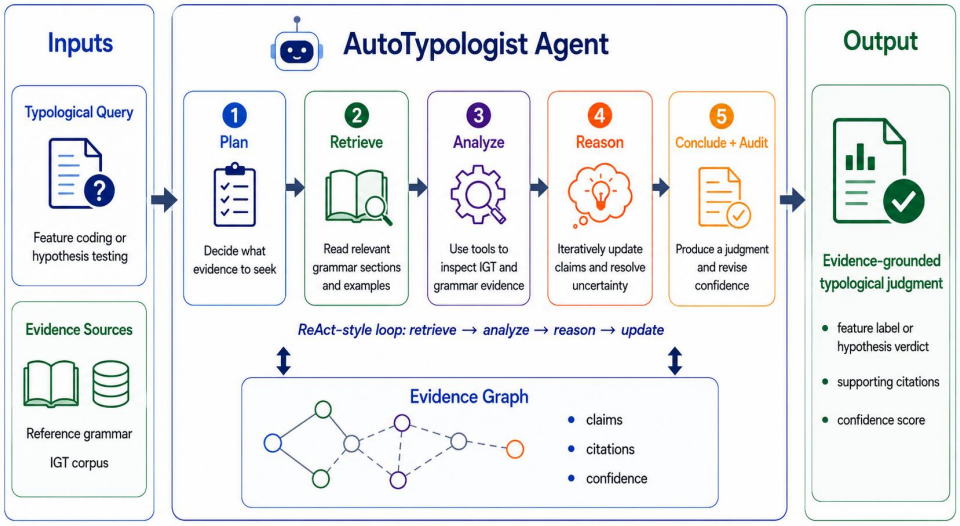

Figure 1: Simplified AutoTypologist pipeline. The agent converts a typological query into an evidence-grounded judgment by planning what evidence is needed, retrieving relevant grammar sections and IGT examples, analyzing the retrieved evidence with linguistically motivated tools, and iteratively updating its claims through a ReAct-style loop. A structured evidence graph records claims, citations, and confidence scores, supporting transparent feature coding or hypothesis testing.

stract language-agnostic meta-linguistic reasoning, where LLMs *"communicate explicit conceptualizations of language phenomena"* (Schneider and Anastasopoulos, 2026). We believe AUTOTYPOLOGIST points to a promising direction for human-AI collaboration in linguistic research.

2 Background

2.1 LLM agents for auto-research

In recent years, researchers are actively exploring using LLM agents to accelerate scientific discoveries across disciplines, including mathematics, physics, and biochemical sciences (Romera-Paredes et al., 2024; Merchant et al., 2023; Fawzi et al., 2022; Hayes et al., 2025; Zhang et al., 2024). Generally, LLMs exhibit potential to perform scientific research, or at least a part of the workflow (Schmidgall et al., 2025), but still struggle to learn scientific taste (Tong et al., 2026). In contrast, relatively few studies have explored how LLM agents can conduct research in the social sciences and humanities (Grossmann et al., 2023; Bail, 2024; Ziems et al., 2024), but the interest is rising rapidly, especially in sociology, management, and economics (Hao and Xie, 2025; Ji et al., 2025; Hu et al., 2026; Engzell and Wilmers, 2026; Korinek, 2025). However, despite the potential, LLM agents have yet to be applied to linguistic research, where low-resource and unseen languages can pose significant challenges. Our research aims to fill in the gap by developing a general LLM agent that can perform automated linguistic analysis.

2.2 LLMs for meta-linguistic reasoning

Some early works explore the automatic induction of grammatical rules from glossed texts (Zamaraeva, 2016; Howell et al., 2017; Chaudhary et al., 2020, 2021). These approaches were built on traditional machine learning approaches and often center on specific aspects of grammar.

Recently, studies find that incorporating meta-linguistic knowledge from reference grammars can aid machine translation in low-resource languages (Tanzer et al., 2024; Hus and Anastasopoulos, 2024), and LLMs do possess abilities to understand meta-linguistic knowledge (Şahin et al., 2020; Bean et al., 2024; Begus et al., 2025; Yang et al., 2025b; Schneider and Anastasopoulos, 2026), though their understanding might be limited (Aycock et al., 2025; Marmonier et al., 2025). Yet it has been observed that, at the moment, no language technology has been truly incorporated into language documentation or typological research workflow (Moeller et al., 2026; Schneider and Anastasopoulos, 2026). Rather than focusing on specific domains or languages, we develop a general LLM research agent that can perform autonomous typological research with meta-linguistic knowledge across languages and grammatical domains.

3 Tasks

We focus on two tasks essential to linguistic typology, with a formulation mirroring the actual research process.

TYPOLOGICAL FEATURE CODING This task mirrors the feature coding process in Grambank (Skirgård et al., 2023), the largest typological database covering 2,467 languages in 195 features.

The 195 features were curated by experienced linguists and include feature definitions, coding procedures, and illustrative evidence from source grammars. Feature GB020[1] below illustrates this format: feature values are assigned based on grammatical descriptions and examples from the relevant source materials (Examples that may be presented in interlinear glossed text (IGT) format. We explain the IGT format in §4.2). While the task instructions were originally developed for human linguists, they can also be adapted to LLM agents. Given the feature query and coding procedures, the agent is instructed to selectively read through a reference grammar to perform feature coding. The agent then produces both a predicted label and an evidence-grounded justification.

Feature GB020. *Are there definite or specific articles?*

Procedure.

1. Code **1** if there is a morpheme that can mark definiteness or specificity without also conveying a spatial deictic meaning.
2. Code **0** if no definite article is mentioned and none can be found in examples or texts in an otherwise comprehensive grammar.
3. Code **?** if the grammar lacks enough analysis to determine whether a definite article exists.

Output.

Language: Aiton
Coding label: **1**
Source: Morey (2004)

mə	nan	a māt	nɛ	wā
time	that	minister	DEF	say

'Then the minister said.'

[1] https://grambank.clld.org/parameters/GB020#2/21.0/151.7

TYPOLOGICAL HYPOTHESIS TESTING A central goal of linguistic typology is to formulate and evaluate hypotheses about crosslinguistic generalizations. Traditionally, human linguists test such hypotheses by comparing descriptions of many languages, identifying recurring structural patterns, and checking whether potential counterexamples exist. For example, the following classic typological hypothesis of word order is from Greenberg (1963).

Universal 1. Word Order: *In declarative sentences with nominal subject and object, the dominant order is almost always one in which the subject precedes the object.*

In this task, the agent will test a typological hypothesis by examining multiple reference grammars to gather evidence and formulate arguments. Hypothesis testing requires the agent to determine whether each language provides supporting evidence, counter-evidence, irrelevant evidence, or insufficient evidence, and then synthesize these language-level judgments into a crosslinguistic conclusion.

4 Typological Data

4.1 Reference grammars

Our grammar corpus[2] was constructed from *Grammar Library*[3], an online list of available reference grammars. We only selected open-sourced reference grammars with a permissive license and included in Grambank, mostly from published books and dissertations. In total, we included 25 reference grammars, representing 14 language families (see details in Appendix A).

4.2 Preprocessing

We digitized the collected reference grammars. First, each grammar was converted from PDF or LaTeX format into plain text. For PDF-based grammars, we primarily used `pdfplumber`. For grammars available in LaTeX, we directly extracted the textual content from the source files.

We also extracted interlinear glossed text (IGT) samples, linguistic examples in which a target language sentence is accompanied by morpheme-by-morpheme segmentation, grammatical glosses

[2] Our data is available: https://link/to/our/repo
[3] https://www.aaronbraver.com/resources/grammars/

(tags), and a free translation. Below is an IGT example from Kagayanen (Pebley and Payne, 2024)[4]:

Source:	Palimpyuan a din ta aparador.
Morpheme:	Pa-limpyo-an a din ta aparador.
Gloss:	T.R-clean-APL 1S.ABS 3S.ERG NABS cabinet
Translation:	'S/he cleaned the cabinet for me.'

IGT provides rich linguistic annotations for the agent to reason over grammatical categories of empirical language data. Candidate IGT examples were identified automatically by detecting numbered example labels. To ensure data quality, all digitized reference grammars and extracted IGT examples were manually reviewed by trained linguists to correct extraction errors.

4.3 Indexing

To address context-length limitation, the digitized reference grammars were converted into structured chunks with unique identifiers. Each grammar was organized by chapter, sections, subsections, and subsubsections based on the original book. To make the representation suitable for retrieval, chunks were kept relatively small, typically capped at 1,000 words with meta information. Each chunk also included a short 2-3 sentence summary generated with `GPT5.4-mini` (OpenAI, 2026), allowing the model to first access a brief overview before reading the full text of relevant sections. Each extracted IGT example was also stored as a structured entry with meta information.Example chunks and IGT are shown in Appendix F.

5 AutoTypologist Framework

5.1 Agent workflow

Our framework is a multi-stage LLM agent architecture for typological research[5]. The system combines structured grammar reading, retrieval, quantitative IGT analysis[6], evidence tracking, and iterative reasoning. The agentic workflow was built using the ReAct framework (Yao et al., 2023). As illustrated in Fig. 1, the workflow consists of the following stages:

Research Planning. Queries of typological phenomena can be supplied explicitly by the user[7]. The agent first reviews the reference grammar overviews[8] and aggregate IGT statistics, and then constructs a per-query investigation plan for execution.

Targeted Pre-retrieval. Before the iterative reasoning loop begins, the agent executes the planned retrieval in two passes. First, it reads the identified target grammar sections in full and extracts explicit author claims into the evidence graph. Second, it runs a set of planned IGT operations, including tag frequency analysis, construction search, and absence checks over the IGT corpus.

Iterative ReAct Investigation. The agent then performs a ReAct-style investigation loop (Yao et al., 2023). In our experiments, each language-level analysis was set up to 10 interaction rounds. At each round, the agent observes the current evidence state, reasons about what information is still missing or uncertain, and selects a tool call (§5.2) such as retrieving additional grammar chunks, or terminates the loop by issuing a *conclude* action.

Evidence Graph. Throughout the pre-retrieval and ReAct phases, the collected observations are incrementally organized into a structured evidence graph, where each node represents a one-sentence claim with meta information including evidence type, source, confidence score, and support/refute decision. The graph tracks supporting evidence, counter-evidence, uncertainty, and contradiction relations, while preserving links to the corresponding grammar chunks and IGT examples.

Synthesis and Reporting. Once the ReAct loop terminates, the agent produces an evidence-grounded typological analysis. The output includes the predicted feature value or typological conclusion, a justification grounded in the retrieved grammar chunks and IGT examples. Finally, the model pass re-examines the synthesized conclusion against the collected evidence. The auditor can uphold, weaken, or overturn the conclusion, and assigns a revised confidence score.

5.2 Tools

We design a set of linguistically motivated tools to constrain the agent action space, each corresponding to a specific evidential operation a human typologist might perform in the research workflow. As summarized in Table 1, the tools fall into three

[4] T.R-CLEAN-APL = transitive realis 'clean' with applicative; 1S.ABS = first-person singular absolutive, 3S.ERG = third-person singular ergative, and NABS = non-absolutive.

[5] Our code is available: `https://link/to/our/repo`

[6] To help the model interpret gloss tags, we also provide the gloss abbreviation list from each grammar as input.

[7] The system also supports automatic summarization based on the grammar structure; we describe this function in Appendix C.

[8] The grammar overview consists of the table of contents reconstructed from chapter, section, subsection, and subsubsection titles, together with summaries associated with each structured chunk.

Tool	Type	Function
read_full_section	Grammar	Read a complete grammar section.
follow_cross_references	Grammar	Follow internal references to related sections.
extract_author_claims	Grammar	Extract explicit analytical claims from prose.
search_text	Grammar	Retrieve relevant grammar passages by hybrid search.
analyse_tag	IGT	Profile frequency, position, and co-occurrence of a tag.
analyse_construction	IGT	Find ordered gloss-tag sequences.
analyse_absence	IGT	Check whether a category is unattested or rare.
compare_tags	IGT	Compare two tags distributionally.
get_section_igt	IGT	Retrieve IGT examples from matching sections.
search_translations	IGT	Search translation lines for semantic evidence.
get_triline_examples	IGT	Retrieve aligned form, gloss, and translation examples.
get_tag_inventory	IGT-only	List all gloss tags by frequency.
get_construction_inventory	IGT-only	List frequent gloss n-gram patterns.
find_tag_cluster	IGT-only	Find tags that co-occur with a seed tag.
analyse_semantic_context	IGT-only	Inspect translation contexts for a tag.
analyse_morpheme_position	IGT-only	Infer prefix/suffix/stem position of a tag.
get_morpheme_forms	IGT-only	List surface forms associated with a tag.
analyse_tag_usage	IGT-only	Infer a tag's function from examples.
parse_example_structure	IGT-only	Analyze clause structure in examples.

Table 1: Tools available to the agent during ReAct-style typological reasoning.

broad groups:

- Grammar-reading tools Grammar-reading tools operate over structured grammar chunks. These tools are intended to approximate the way a typologist consults a reference grammar: first locating the relevant descriptive section, then checking whether the author makes an explicit analytical claim about the target phenomenon.

- IGT-analysis tools IGT-analysis tools operate over the extracted interlinear glossed examples to provide both aggregate and example-level evidence. These tools allow the agent to test whether grammatical categories described in grammar prose are also reflected in the observed examples, and to infer patterns directly from IGT data when prose descriptions are incomplete.

- IGT-only exploratory tools In the IGT-only condition, the agent must reason with only IGT examples. We therefore provide additional exploratory tools designed to support grammatical induction from annotated data, while making the evidence used for each inference explicit and inspectable.

Tool availability is determined by the information sources accessible to the agent: grammar-reading tools are only available when reference grammar text is provided, IGT-only exploratory tools are only available when no reference grammar is present, and IGT-analysis tools are available whenever IGT examples are provided.

Observability and literature grounding To enhance agent observability, we keep all agent traces during the multistep inferences. In each inference step, we force the agent to ground their decision by citing the unique identifier of original grammar chunks and IGTs from the corpus. This practice allows end users to inspect the agent decisions with reference to the sources to understand how agent decisions evolve.

6 Experiments

Models We deployed our agents with a range of LLMs, including open-source models: `Qwen2.5-7B` (Hui et al., 2024), `Qwen3.5-9B` (Yang et al., 2025a), `Qwen3.6-27B` (Yang et al., 2025a), and `Gemma4-31B` (Google DeepMind, 2026), and a closed-source model: `GPT5.4-mini`. All open-sourced local models were deployed with `vllm` for efficient inference (Kwon et al., 2023) in `bfloat16` precision. Full details are available in Appendix G.

Design In both TYPOLOGICAL FEATURE CODING and TYPOLOGICAL HYPOTHESIS TESTING, we contrast three different information conditions in which the agent has access to different sources of linguistic evidence:

Condition	Grammar chunks	IGT entries
IGT	✗	✓
Grammar	✓	✗
Grammar+IGT	✓	✓

The **IGT** condition presents a more challenging but realistic scenario in which only annotated linguistic examples are available, requiring the agent to infer grammatical features from raw data without access to author explanations. In the **Gram-**

mar condition, the agent reasons over structured grammar chunks. The **Grammar+IGT** condition provides the richest setting, allowing the agent to combine explicit grammatical descriptions with example-based evidence from IGT entries.

For TYPOLOGICAL FEATURE CODING, we use all 25 languages coded in Grambank. For TYPOLOGICAL HYPOTHESIS TESTING, we select 10 languages per hypothesis-testing run: Aguaruna, Chácobo, Chakali, Moloko, Ik, Eastern Geshiza, Papuan Malay, Komnzo, Kalamang, and Pite Saami. We choose these languages to maximize coverage across language families. We selected 10 relatively general universals from Greenberg's original set of 45 universals, prioritizing hypotheses that could be evaluated across a broad range of reference grammars. We also manually created adversarial hypotheses by permuting or negating the original hypotheses, making them unlikely to be correct or to have appeared verbatim in pretraining data. The full list of selected universals and their adversarial counterparts is provided in Appendix B. We evaluated three models for this task: `Qwen3.6-27B`, `Gemma4-31B`, and `GPT5.4-mini`. For each information condition, we ran the agent 3 times to account for variability. We report results averaged across these three runs.

Baselines Specifically for TYPOLOGICAL FEATURE CODING, we include two baselines for comparison. The first is a **zero-shot prompting** baseline, in which the model directly predicts feature codings without any additional contextual or grammatical guidance. This baseline evaluates whether LLMs can infer typological features of low-resource languages from pretrained knowledge alone. The second is a **majority-vote** baseline derived from Grambank, where each feature is assigned the value that occurs most frequently across languages (Details are shown in Appendix E). Although simple, this baseline is difficult to beat: because Grambank feature-value distributions are highly imbalanced, majority voting benefits from strong dataset-level typological biases and can achieve competitive performance by favoring frequent values. Thus, improvements over the majority-vote baseline provide evidence that the model is capturing grammatical information beyond the majority-class prior.

6.1 Metrics

For TYPOLOGICAL FEATURE CODING, we evaluate whether the model's predicted feature values are consistent with the corresponding human annotations in Grambank. Following the Grambank coding scheme, most features are coded with three labels, {0, 1, ?}, where 0 and 1 indicate the absence or presence of a feature and ? marks uncertainty or insufficient information. Only 6 features additionally allow multi-valued labels such as 2 or 3 (Haynie et al., 2023). Since typological feature labels are often highly imbalanced across languages, we report both **Weighted F1** and **Macro F1**. Weighted F1 accounts for label frequency, while Macro F1 gives equal weight to each label.

For TYPOLOGICAL HYPOTHESIS TESTING, the task is hypothesis-driven typological analysis rather than categorical feature prediction. We report the average number of languages assigned to each category across runs. These language-level evidence counts determine whether the hypothesis is ultimately confirmed, refuted, or left inconclusive.

7 Main Findings

7.1 Typological feature coding

Table 2 presents the agent's performance on TYPOLOGICAL FEATURE CODING. Detailed predictions are shown in Appendix D.

Performance under different information conditions Overall, providing access to reference grammars substantially improves typological feature coding performance across nearly all models. Compared with the zero-shot baseline, the Grammar condition consistently yields the strongest gains, particularly for larger models such as `Qwen3.6-27B` and `Gemma4-31B`. This suggests that current LLMs are able to effectively utilize explicit grammatical explanations and descriptive prose from reference grammars for typological reasoning. In contrast, the IGT-only condition remains considerably more challenging, although all models still outperform the corresponding zero-shot baseline. This indicates that LLMs can induce non-trivial grammatical generalizations directly from glossed examples, even without access to meta-linguistic explanations.

Interestingly, the Grammar+IGT condition does not consistently outperform the Grammar-only condition. For most models, adding IGT examples on top of the full grammar leads to only marginal improvements or even slight performance degradation. This suggests that current LLM agents may still rely more heavily on structured grammatical prose than on bottom-up grammatical induction from raw

Condition	Majority Vote		Qwen2.5-7b		Qwen3-9b		Qwen3.6-27b		Gemma4-31B		GPT5.4-mini	
	Wgt.	Mac.	Wgt.	Mac.	Wgt.	Mac.	Wgt.	Mac.	Wgt.	Mac.	Wgt.	Mac.
Zero-shot	62.5	41.5	42.3	26.7	40.9	31.1	58.1	40.2	56.8	43.4	52.2	38.3
IGT	62.5	41.5	52.7	34.0	54.4	38.1	63.8	45.4	64.5	50.5	58.5	42.8
Grammar	62.5	41.5	55.9	41.3	60.1	47.3	**70.7**	**59.2**	69.6	54.6	69.1	56.4
Grammar+IGT	62.5	41.5	50.6	40.3	52.8	41.7	67.8	53.1	67.5	52.3	65.6	54.4

Table 2: Weighted and Macro F1 scores (%) of TYPOLOGICAL FEATURE CODING under different conditions.

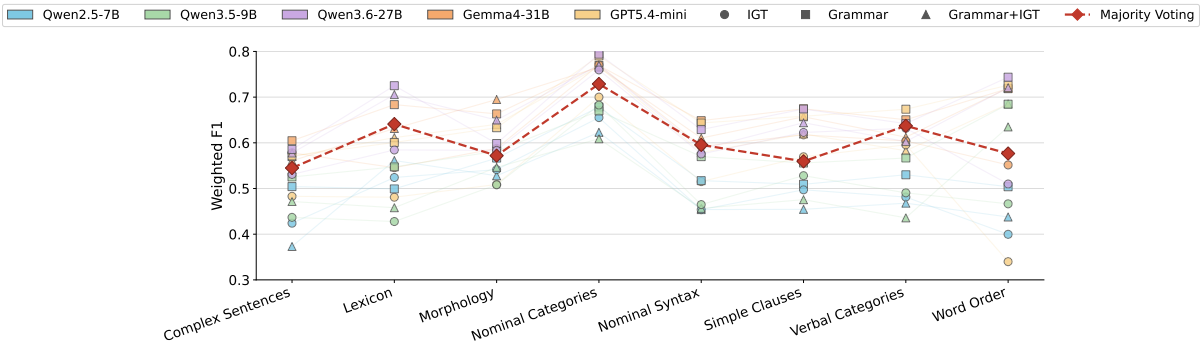

Figure 2: Weighted F1 scores of TYPOLOGICAL FEATURE CODING across domains (all settings).

IGT evidence. One possible explanation is that the additional IGT evidence introduces retrieval noise or increases the complexity of evidence integration during the ReAct process.

Performance across grammatical domains We categorize the evaluated features into 8 grammatical domains, following the broad feature-domain organization used in WALS (Dryer and Haspelmath, 2013). This domain-level grouping allows us to evaluate whether model performance varies across different areas of grammar, such as morphology, syntax, and nominal or verbal categories. Details about the feature grouping is shown in Appendix H. We depict our weighted F1 score in Fig. 2. In general, LLMs are more consistent with human coding in *Nominal Categories*, but face more challenges in *Complex Sentences* and *Verbal Categories*. The performance fluctuations across domains are highly aligned with the majority-voting baseline performance, which seems to suggest that prior background knowledge still plays a role.

Token usage We found that token usage does not necessarily reflect performance, confirming the findings from (Bai et al., 2026). As shown in Fig. 3, more token usage seems to be associated with lower performance. This is because small models generally consume more tokens than large models, and the IGT-only condition costs less token usage than the other two conditions.

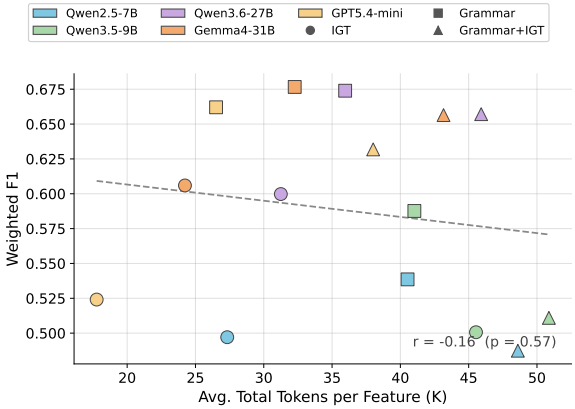

Figure 3: Weighted F1 with respect to token usage in TYPOLOGICAL FEATURE CODING. More token usage does not necessarily lead to higher performance.

7.2 Typological hypotheses testing

Table 3 illustrates that AUTOTYPOLOGIST is capable of performing TYPOLOGICAL HYPOTHESIS TESTING across languages independently. Even for adversarial hypotheses, which are deliberately crafted to be false for most languages, AUTOTYPOLOGIST can still successfully falsify those hypotheses based on available evidence from reference grammars.

Generally speaking, LLMs tend to be less accurate in confirming or refuting hypotheses when only IGTs are given, again highlighting the challenges of inducing grammatical structures from raw data. LLMs are still better at synthesizing information

Model	Condition	Standard				Adversarial			
		# For	# Against	# N/A	# Insuf.	# For	# Against	# N/A	# Insuf.
Qwen3.6-27b	IGT	4.47	0.13	1.73	3.67	0.63	4.67	1.80	2.90
	Grammar	3.57	0.10	1.83	4.50	0.07	3.80	1.70	4.43
	Grammar+IGT	5.73	0.10	2.13	2.03	0.03	5.27	2.20	2.50
Gemma4-31b	IGT	4.77	0.60	2.37	2.27	1.13	4.73	2.10	2.03
	Grammar	5.53	0.23	3.00	1.23	0.07	5.73	2.63	1.57
	Grammar+IGT	6.33	0.07	2.70	0.90	0.20	5.73	2.63	1.43
GPT5.4-mini	IGT	1.93	0.03	1.07	6.97	0.13	1.40	0.90	7.57
	Grammar	1.33	0.00	0.67	8.00	0.03	0.87	0.73	8.37
	Grammar+IGT	1.47	0.00	0.67	7.87	0.07	0.83	0.63	8.47

Table 3: Average evidence counts for standard and adversarial universals in TYPOLOGICAL HYPOTHESIS TESTING, averaged over 3 runs and 10 universals. # For + # Against. + # N/A + # Insuf. = 10 (languages per run).

from well-structured English grammatical descriptions, but struggle at reasoning over IGTs from rare or unseen languages.

LLMs also exhibit different behaviors during the agentic analysis. `GPT5.4-mini` is the most uncertain LLM, as it concludes insufficient evidence for most languages in a batch. In contrast, both `Gemma4-31b` and `Qwen3.6-27b` are more willing to draw conclusions from individual languages. We speculate that such differences might arise from the post-training stage.

8 Error Analysis

We further analyze the error patterns of our strongest setting. We check the predictions from `Qwen3.6-27B` with grammar access. For TYPOLOGICAL FEATURE CODING, as shown in the table below: The model sometimes treats absent features as present, or misses evidence for features that are actually present. Interestingly, gold ? cases are often forced into categorical labels, and tend to predict as absent.

Gold	pred 0	pred 1	pred ?	Others
0	—	56.0%	44.0%	—
1	66.4%	—	31.4%	2.2%
?	67.2%	30.8%	—	2.0%

We further inspect one GB049 prediction trace for Japhug (Appendix I). The agent incorrectly predicts `0` because it interprets the feature too narrowly: it looks for a dedicated affix exclusively deriving object nouns from verbs, rather than considering broader productive morphological patterns. At the same time, retrieval is not well aligned with the terminology used in the grammar. The agent searches around nominal derivations and Object–Verb compounds, but misses the relevant section on object participles. This case shows that the model can fail when a Grambank feature and a reference grammar describe the same phenomenon using different categories or terms.

For TYPOLOGICAL HYPOTHESIS TESTING, we analyze the agent trace when testing Greenberg Universal rule (U30). As shown in Appendix J, the model incorrectly refutes the standard Greenberg universal by treating Ik as a counterexample. This error also appears to come from an overly narrow interpretation of the consequent: the model equates "tense-mode categories" with a dedicated or obligatory tense paradigm. The planning step is also suboptimal, since the selected sections for Ik focus on phonology and morphophonology rather than verbal morphology or tense, aspect, and modality.

These errors highlights the gap between human linguists and AUTOTYPOLOGIST in interpreting linguistic data meta-linguistically. Human expertise is still necessary in validating the outputs.

9 Conclusions

In this study, we propose AUTOTYPOLOGIST, an LLM agent that can conduct typological research across languages autonomously. Empirical evaluations show that AUTOTYPOLOGIST can complete TYPOLOGICAL FEATURE CODING and TYPOLOGICAL HYPOTHESIS TESTING with strong performance. Beyond task performance, our findings highlight the importance of decomposing typological research into structured tools and steps with traceable evidence.

By automating the otherwise time-consuming work of querying evidence across reference grammars, AUTOTYPOLOGIST can enable linguists to focus on deeper analysis. Future research can extend this framework to more languages, more modalities (speech and signs), and incorporate interactive feedback from human linguists.

Limitations

Our research agent is validated against existing typological databases carefully curated by expert linguists. The values of WALS-like typological databases (Grambank and APiCS) are also not without controversies within the linguistic community from theoretical and empirical grounds (e.g., Davis et al., 2014; Haspelmath, 2014). A small percentage of coded features in some languages are also found to be inaccurate (Haspelmath, 2014). For comparability, such databases reduce linguistic diversity into a limited set of discrete codes, though such discrete codes applied uniformly to languages are needed for cross-linguistic comparison.

Secondly, while the agent can generate and test new hypotheses, extensive human validation is still needed to validate the scientific value and the correctness of these hypotheses. AUTOTYPOLOGIST brings a validation asymmetry where providing answers is way easier than validating these answers. Also, AUTOTYPOLOGIST cannot replace human field linguists who gather linguistic data from fieldwork. Many universal hypotheses in TYPOLOGICAL HYPOTHESIS TESTING still need more linguistic data to validate.

Thirdly, language is inherently multi-modal. Our current research only focuses on the textual aspects, due to the constraints of available data and LLMs. But the speech and the visual (signing) modalities of language are also essential components of language, which deserve further research.

Languages are living and fluid, constantly changing with the linguistic communities. AUTOTYPOLOGIST relies on reference grammars, which are only static snapshots of the living languages. LLM agents cannot replace field linguists who are doing fieldwork to interact with diverse communities.

Ethical considerations

Our research builds on the collective effort from the field linguists and various linguistic communities. Our purpose is not to replace human linguists but to provide tools that facilitate linguistic research. We acknowledge the importance of linguists and fieldwork, not just in collecting language data, but also in building bonds with those linguistic communities.

We only selected reference grammars with permissive licenses like Creative Commons licenses. All reference grammars are cited and will only be used for research purposes. We also implement an opt-out mechanism where the authors and communities can ask to remove their grammars from the agentic system.

Statements on LLM usage We disclose that LLMs had been used to assist with grammar correction and text polishing. LLMs were also used for general coding assistance under human oversight.

References

Seth Aycock, David Stap, Di Wu, Christof Monz, and Khalil Simaan. 2025. Can llms really learn to translate a low-resource language from one grammar book? In *International Conference on Learning Representations*, volume 2025, pages 12334–12357.

Longju Bai, Zhemin Huang, Xingyao Wang, Jiao Sun, Rada Mihalcea, Erik Brynjolfsson, Alex Pentland, and Jiaxin Pei. 2026. How do ai agents spend your money? analyzing and predicting token consumption in agentic coding tasks. *arXiv preprint arXiv:2604.22750*.

Christopher A Bail. 2024. Can generative ai improve social science? *Proceedings of the National Academy of Sciences*, 121(21):e2314021121.

Andrew Bean, Simi Hellsten, Harry Mayne, Jabez Magomere, Ryan Chi, Scott A Hale, Hannah R Kirk, and 1 others. 2024. Lingoly: A benchmark of olympiad-level linguistic reasoning puzzles in low resource and extinct languages. *Advances in Neural Information Processing Systems*, 37:26224–26237.

Gasper Begus, Maksymilian Dabkowski, and Ryan Rhodes. 2025. Large linguistic models: Investigating llms' metalinguistic abilities. *IEEE Transactions on Artificial Intelligence*.

Aditi Chaudhary, Antonios Anastasopoulos, Adithya Pratapa, David R. Mortensen, Zaid Sheikh, Yulia Tsvetkov, and Graham Neubig. 2020. Automatic extraction of rules governing morphological agreement. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5212–5236, Online. Association for Computational Linguistics.

Aditi Chaudhary, Kayo Yin, Antonios Anastasopoulos, and Graham Neubig. 2021. When is wall a pared and when a muro?: Extracting rules governing lexical selection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6911–6929, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bernard Comrie. 1989. *Language universals and linguistic typology: Syntax and morphology*. University of Chicago press.

Henry Davis, Carrie Gillon, and Lisa Matthewson. 2014. How to investigate linguistic diversity: Lessons from the pacific northwest. *Language*, 90(4):e180–e226.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online (v2020.4)*. Zenodo.

Per Engzell and Nathan Wilmers. 2026. The paper factory.

Alhussein Fawzi, Matej Balog, Aja Huang, Thomas Hubert, Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Francisco J R Ruiz, Julian Schrittwieser, Grzegorz Swirszcz, and 1 others. 2022. Discovering faster matrix multiplication algorithms with reinforcement learning. *Nature*, 610(7930):47–53.

Google DeepMind. 2026. Gemma 4 Model Card. https://ai.google.dev/gemma/docs/core/model_card_4. Last updated 2026-04-17. Accessed: 2026-05-26.

Joseph H Greenberg. 1963. Some universals of grammar with particular reference to the order of meaningful elements. *Universals of language*, 2(1):73–113.

Joseph H Greenberg. 2005. *Language universals: With special reference to feature hierarchies*, volume 59. Walter de Gruyter.

Igor Grossmann, Matthew Feinberg, Dawn C Parker, Nicholas A Christakis, Philip E Tetlock, and William A Cunningham. 2023. Ai and the transformation of social science research. *Science*, 380(6650):1108–1109.

Yuzhi Hao and Danyang Xie. 2025. A multi-llm-agent-based framework for economic and public policy analysis. *arXiv preprint arXiv:2502.16879*.

Martin Haspelmath. 2014. Descriptive hypothesis testing is distinct from comparative hypothesis testing: Commentary on davis, gillon, and matthewson. *Language*, 90(4):e250–e257.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, and 1 others. 2025. Simulating 500 million years of evolution with a language model. *Science*, page eads0018.

Hannah J. Haynie, Damián Blasi, Hedvig Skirgård, Simon J. Greenhill, Quentin D. Atkinson, and Russell D. Gray. 2023. Grambank's typological advances support computational research on diverse languages. In *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 147–149, Dubrovnik, Croatia. Association for Computational Linguistics.

Kristen Howell, Emily M. Bender, Michel Lockwood, Fei Xia, and Olga Zamaraeva. 2017. Inferring case systems from IGT: Enriching the enrichment. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 67–75, Honolulu. Association for Computational Linguistics.

Saiquan Hu, Yushan Wei, and Shuyang Sun. 2026. Soclitgen: An llm-assisted framework for automated literature review in social sciences. *Information Processing & Management*, 63(7):104885.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, and 1 others. 2024. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*.

Jonathan Hus and Antonios Anastasopoulos. 2024. Back to school: Translation using grammar books. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20207–20219, Miami, Florida, USA. Association for Computational Linguistics.

Xinyu Ji, Jingshen Xue, Siyi Chen, and Tingshao Zhu. 2025. Automating social science: llms vs. human experts in variable relationship identification. *Collabra: Psychology*, 11(1):144442.

Anton Korinek. 2025. Ai agents for economic research. Technical report, National Bureau of Economic Research.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pages 611–626.

Malik Marmonier, Rachel Bawden, and Benoît Sagot. 2025. Explicit learning and the LLM in machine translation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31372–31422, Suzhou, China. Association for Computational Linguistics.

Amil Merchant, Simon Batzner, Samuel S Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. 2023. Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85.

Susanne Maria Michaelis, Philippe Maurer, Martin Haspelmath, and Magnus Huber. 2013. *The atlas of pidgin and creole language structures*. Oxford University Press.

Sarah Moeller, Godfred Agyapong, Jarrod Cruz, Alexis Palmer, and Mans Hulden. 2026. Computational methods for language documentation and description. *Annual Review of Linguistics*, 12(1):147–170.

Stephen Morey. 2004. The tai languages of assam1. In *The Tai-Kadai languages*, pages 207–253. Routledge.

OpenAI. 2026. GPT-5 is Here. https://openai.com/gpt-5/. Accessed: 2026-05-26.

Carol J Pebley and Thomas E Payne. 2024. *A grammar of Kagayanen*, volume 8. Language Science Press.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M Pawan Kumar, Emilien Dupont, Francisco JR Ruiz, Jordan S Ellenberg, Pengming Wang, Omar Fawzi, and 1 others. 2024. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475.

Gözde Gül Şahin, Yova Kementchedjhieva, Phillip Rust, and Iryna Gurevych. 2020. PuzzLing Machines: A Challenge on Learning From Small Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1241–1254, Online. Association for Computational Linguistics.

Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Michael Moor, Zicheng Liu, and Emad Barsoum. 2025. Agent laboratory: Using LLM agents as research assistants. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5977–6043, Suzhou, China. Association for Computational Linguistics.

Nathan Schneider and Antonios Anastasopoulos. 2026. Speaking of language: Reflections on metalanguage research in nlp. *arXiv preprint arXiv:2604.02645*.

Hedvig Skirgård, Hannah J. Haynie, Damián E. Blasi, Harald Hammarström, Jeremy Collins, Jay J. Latarche, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Sam Passmore, Angela Chira, Luke Maurits, Russell Dinnage, Michael Dunn, Ger Reesink, Ruth Singer, Claire Bowern, Patience Epps, Jane Hill, and 86 others. 2023. Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances*, 9(16).

Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. 2024. A benchmark for learning to translate a new language from one grammar book. In *International Conference on Learning Representations*, volume 2024, pages 18955–18985.

Jingqi Tong, Mingzhe Li, Hangcheng Li, Yongzhuo Yang, Yurong Mou, Weijie Ma, Zhiheng Xi, Hongji Chen, Xiaoran Liu, Qinyuan Cheng, and 1 others. 2026. Ai can learn scientific taste. *arXiv preprint arXiv:2603.14473*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Changbing Yang, Franklin Ma, Freda Shi, and Jian Zhu. 2025b. LingGym: How far are LLMs from thinking like field linguists? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1314–1340, Suzhou, China. Association for Computational Linguistics.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Olga Zamaraeva. 2016. Inferring morphotactics from interlinear glossed text: Combining clustering and precision grammars. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 141–150, Berlin, Germany. Association for Computational Linguistics.

Qiang Zhang, Keyang Ding, Tianwen Lyv, Xinda Wang, Qingyu Yin, Yiwen Zhang, Jing Yu, Yuhao Wang, Xiaotong Li, Zhuoyi Xiang, and 1 others. 2024. Scientific large language models: A survey on biological & chemical domains. *arXiv preprint arXiv:2401.14656*.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

A Reference Grammars Details

Table 4 and 5 shows detailed information about the reference grammar languages and sources we used in the paper.

Language	Family	Chunk #	IGT #
Moloko	Afro-Asiatic	219	488
Chakali	Atlantic-Congo	97	405
Kam	Atlantic-Congo	262	698
Paunaka	Arawakan	345	1604
Belep	Austronesian	303	1009
Kagayanen	Austronesian	296	570
Lele	Austronesian	168	494
Nese	Austronesian	205	611
Paluai	Austronesian	343	603
Papuan Malay	Austronesian	450	1155
Aguaruna	Chicham	523	512
Ulwa	Keram	354	1502
Ik	Kuliak	128	103
Hup	Naduhup	425	2074
Bargam	Nuclear Trans New Guinea	241	823
Mauwake	Nuclear Trans New Guinea	276	1646
Mian	Nuclear Trans New Guinea	601	1013
Chácobo	Pano-Tacanan	452	1302
Eastern Geshiza	Sino-Tibetan	580	1079
Japhug	Sino-Tibetan	1232	5517
Karbi	Sino-Tibetan	384	681
Yakkha	Sino-Tibetan	355	1312
Pite Saami	Uralic	222	179
Kalamang	West Bomberai	337	630
Komnzo	Yam	223	565

Table 4: Languages included in the analysis, together with the number of extracted chunks and IGT instances.

B Greenberg Universals

Table 6 shows both the standard Greenberg universals and adversarial counterparts used in the TYPOLOGICAL HYPOTHESIS TESTING.

C Automatic Summarization

Beyond the two evaluated tasks, AUTOTYPOLOGIST supports an automatic summarization mode in which the agent autonomously generates a structured typological profile of a language without any user-supplied queries.

In this mode, the agent is first presented with the grammar table of contents and aggregate IGT statistics, from which it identifies major typological domains and generates a set of candidate feature questions for each domain. It then executes the full ReAct investigation pipeline for each candidate question, producing an evidence-grounded answer. The resulting feature-level answers are finally aggregated into a structured typological summary covering the language's major grammatical domains, including word order, morphological complexity, tense-aspect-modality, nominal categories, and argument structure, among others.

This mode is intended to support rapid typological documentation of under-described languages, where no specific research question is known in advance. The output is a structured JSON report containing, for each investigated feature, the predicted value, a prose justification grounded in grammar sections and IGT examples, quantitative IGT evidence, and a confidence score. We did not conduct a systematic evaluation of this mode in the present study, as it would require a different evaluation framework from the feature-by-feature comparison against Grambank. We leave a full evaluation of automatic summarization quality to future work.

D Label Prediction Details

Tables 7, 8, 9, 10, and 11 show detailed prediction labels for various languages, models, and settings.

E Majority Voting Labels

Table 12 shows details about the majority voting labels for all Grambank features.

F Grammar Chunk and IGT entry example

Figures 4 and 5 show the example of grammar chunk and IGT entry, respectively.

G Hyperparameters

Table 13 describes different types of system hyperparameters.

H Grambank Domain Types and Distributions

Table 14 shows the detailed information about Grambank feature domain distribution according to WALS.

I Sample TYPOLOGICAL FEATURE CODINGAgent Traces

Figure 6 shows a complete agent trace for coding Grambank feature GB049 in Japhug, including the retrieved grammar sections, intermediate evidence, final label prediction, and audit result.

J Sample TYPOLOGICAL HYPOTHESIS TESTINGAgent Traces

Figure 7 shows the agent trace results for testing Greenberg Universal U30 with Qwen3.6-27B under Grammar setting.

K Agent Prompts

Figures 8, 9, 10, and 11 illustrate representative prompt templates used at different stages of the system workflow.

Example of a structured grammar chunk	
language	Kagayanen
chunk_id	chunk_0087_p0
chapter	Verb structure and inflection
section	Overall verb structure
subsection	
subsubsection	
text	A verbal predicate in Kagayanen consists minimally of a stem, plus one and only one inflectional affix. Verbal predicate = Infl-Stem. A stem consists minimally of a root, and may contain one or more stem-forming processes: Stem = (SF)-Root-(SF) ...
summary	In Kagayanen, a verbal predicate minimally consists of a stem plus exactly one inflectional affix, with optional adverbial elements appearing before or after this core structure. The stem itself minimally contains a root and may also include one or more stem-forming processes.

Figure 4: Example of a structured grammar chunk. Each chunk has a unique identifier and stores the original text together with hierarchical metadata and a short summary.

Example of a structured IGT entry	
language	Kagayanen
example_id	ex:thecabinet
source	Palimpyuan a din ta aparador.
morpheme	Pa-limpyo-an a din ta aparador.
gloss	T.R-clean-APL 1S.ABS 3S.ERG NABS cabinet
translation	S/he cleaned the cabinet for me. (Meaning I will not have to clean it.)
gloss_tags	[T.R, clean, APL], [1S.ABS], [3S.ERG], [NABS], [cabinet]
chapter	Voice
section	The Kagayanen choir
subsection	Applicative constructions
subsubsection	

Figure 5: Example of a structured IGT entry. Each IGT example has a unique identifier and stores the source sentence, morpheme-segmented line, gloss line, free translation, token-level gloss tags, and the corresponding structural metadata from the original grammar.

Language	Family	Grambank Source	Our Source*
Moloko	Afro-Asiatic	Friesen & Mamalis. 2004. *The Moloko Verb Phrase*. SIL.	Friesen. 2017. *A grammar of Moloko*. Language Science Press.
Paunaka	Arawakan	Danielsen & Terhart. 2014. Paunaka. In *Oriente*.	Terhart. 2024. *A grammar of Paunaka*. Language Science Press.
Chakali	Atlantic-Congo	Brindle. 2010. *Aspects of the Chakali language*. NTNU dissertation.	Brindle. 2017. *A dictionary and grammatical outline of Chakali*. Language Science Press.
Kam	Atlantic-Congo	Lesage. 2020. *A grammar and lexicon of Kam*. INALCO dissertation.	
Belep	Austronesian	McCracken. 2013. *A grammar of Belep*. Rice University dissertation.	
Kagayanen	Austronesian	Pebley & MacGregor. 1999. *Studies in Kagayanen*. SIL.	Pebley & Payne. 2024. *A grammar of Kagayanen*. Language Science Press.
Lele	Austronesian	Boettger. 2015. *Topics in the grammar of Lele*. James Cook University dissertation.	
Nese	Austronesian	Takau. 2016. *A grammar of Nese*. University of Newcastle dissertation.	
Paluai	Austronesian	Schokkin. 2014. *A grammar of Paluai*. James Cook University dissertation.	
Papuan Malay	Austronesian	Kluge. 2017. *A grammar of Papuan Malay*. Language Science Press.	
Aguaruna	Chicham	Overall. 2007. *A Grammar of Aguaruna*. LaTrobe University dissertation.	
Ulwa (PNG)	Keram	Barlow. 2018. *A grammar of Ulwa*. University of Hawai'i dissertation.	Barlow. 2023. *A grammar of Ulwa*. Language Science Press.
Ik	Kuliak	Schrock. 2014. *A grammar of Ik*. Leiden University dissertation.	Schrock. 2017. *The Ik language*. Language Science Press.
Hup	Naduhup	Epps. 2008. *A grammar of Hup*. Mouton de Gruyter.	
Bargam	Nuclear Trans New Guinea	Hepner. 2006. *Bargam Grammar Sketch*. SIL.	
Mauwake	Nuclear Trans New Guinea	Berghäll. 2010. *Mauwake reference grammar*. University of Helsinki dissertation.	Berghäll. 2015. *A grammar of Mauwake*. Language Science Press.
Mian	Nuclear Trans New Guinea	Fedden. 2007. *A Grammar of Mian*. University of Melbourne dissertation.	
Chácobo	Pano-Tacanan	Tallman. 2018. *A Grammar of Chácobo*. University of Texas dissertation.	
Eastern Geshiza	Sino-Tibetan	Honkasalo. 2019. *A grammar of Eastern Geshiza*. University of Helsinki dissertation.	
Japhug	Sino-Tibetan	Jacques. 2004/2017. *Phonologie et Morphologie du Japhug*; Japhug. In *Sino-Tibetan Languages*.	Jacques. 2021. *A grammar of Japhug*. Language Science Press.
Karbi	Sino-Tibetan	Grüßner. 1978. *Arleng Alam Die Sprache der Mikir*. Franz Steiner.	Philippova, Nailya. 2021. A Grammar of Amri Karbi. Diss. University of Helsinki.
Yakkha	Sino-Tibetan	Schackow. 2015. *A grammar of Yakkha*. Language Science Press.	
Pite Saami	Uralic	Wilbur, J. (p.c.)	
Kalamang	West Bomberai	Visser. 2016. *A grammar sketch of Kalamang*. University of Oslo MA thesis.	Visser. 2022. *A grammar of Kalamang*. Language Science Press.
Komnzo	Yam	Döhler. 2016. *Komnzo*. ANU dissertation.	Döhler. 2018. *A grammar of Komnzo*. Language Science Press.

Table 5: Reference grammars used in this study, grouped by language family. *Where our source differs from the Grambank source (typically a more recent published version of the same grammar), it is listed in the "Our Source" column. Grammars highlighted in read were not used in this study.

ID	Logic	Domain	Standard universal	Adversarial hypothesis
U1	absolute	Word order	In declarative sentences with nominal subject and object, the dominant order is almost always one in which the subject precedes the object.	In declarative sentences with nominal subject and object, the dominant order is almost always one in which the object precedes the subject.
U8	implication	Information structure	When a yes-no question is marked by a single morpheme, it is always differentiated from the corresponding declarative by intonation or by a characteristic position of the morpheme at the beginning or end of the sentence.	When a yes-no question is marked by a single morpheme, it is not differentiated from the corresponding declarative by intonation or by a characteristic position of the morpheme at the beginning or end of the sentence.
U14	absolute	Word order	In conditional statements, the conditional clause precedes the conclusion as the normal order in all languages.	In conditional statements, the conclusion precedes the conditional clause as the normal order in all languages.
U18	correlation	Noun phrase structure	When the descriptive adjective precedes the noun, the demonstrative and the numeral, with overwhelmingly more than chance frequency, do likewise.	When the descriptive adjective precedes the noun, the demonstrative and the numeral, with overwhelmingly more than chance frequency, follow the noun.
U25	implication	Word order	If the pronominal object follows the verb, so does the nominal object.	If the pronominal object follows the verb, the nominal object precedes the verb.
U29	implication	Morphological complexity	If a language has inflection, it always has derivation.	If a language has inflection, it does not have derivation.
U30	implication	Tense/aspect/modality	If the verb has categories of person-number or of gender, it always has tense-mode categories.	If the verb has categories of person-number or of gender, it does not have tense-mode categories.
U34b	implication	Morphological complexity	No language has a dual unless it has some nonsingular number category.	A language may have a dual even if it has no other nonsingular number category.
U36	implication	Agreement	If a language has the category of gender, it always has the category of number.	If a language has the category of gender, it does not have the category of number.
U42	absolute	Morphological complexity	All languages have pronominal categories involving at least three persons and two numbers.	All languages have pronominal categories involving fewer than three persons or fewer than two numbers.

Table 6: Greenberg universals and their adversarial counterparts used in the hypothesis-testing experiments.

Language	IGT					Grammar					Grammar+IGT				
	0	1	2	3	?	0	1	2	3	?	0	1	2	3	?
Aguaruna	0.61	0.42	0.00	0.00	0.12	0.66	0.56	0.50	1.00	0.00	0.58	0.48	0.33	1.00	0.08
Bargam	0.58	0.36	0.00	*NA*	0.08	0.60	0.42	0.33	*NA*	0.08	0.49	0.37	0.00	*NA*	0.05
Belep	0.65	0.35	0.40	*NA*	0.09	0.68	0.51	0.57	*NA*	0.00	0.58	0.43	0.29	*NA*	0.08
Chakali	0.68	0.33	0.40	*NA*	0.00	0.74	0.44	0.57	*NA*	0.00	0.66	0.45	0.33	*NA*	0.00
Chacobo	0.67	0.56	0.00	0.00	0.00	0.60	0.57	0.00	0.67	0.14	0.57	0.54	1.00	0.67	0.08
Eastern Geshiza	0.66	0.50	0.00	*NA*	0.00	0.68	0.58	0.40	*NA*	0.00	0.68	0.58	0.00	*NA*	0.00
Hup	0.62	0.43	0.00	*NA*	0.00	0.58	0.51	0.00	*NA*	0.09	0.39	0.44	0.50	*NA*	0.00
Ik	0.72	0.26	0.57	0.00	0.00	0.65	0.46	1.00	0.00	0.00	0.70	0.47	0.91	0.00	0.22
Japhug	0.59	0.45	0.40	0.00	0.00	0.58	0.57	0.50	0.00	0.00	0.45	0.47	0.00	0.00	0.00
Kagayanen	0.63	0.43	1.00	0.00	0.12	0.68	0.43	0.67	0.00	0.13	0.54	0.40	0.50	0.50	0.07
Kalamang	0.53	0.18	0.50	*NA*	0.08	0.45	0.19	0.33	*NA*	0.06	0.46	0.21	0.67	*NA*	0.06
Kam	0.75	0.40	0.67	*NA*	0.00	0.73	0.53	0.75	*NA*	0.00	0.66	0.53	0.67	*NA*	0.00
Karbi	0.59	0.33	0.00	*NA*	0.04	0.51	0.38	0.50	*NA*	0.12	0.42	0.38	0.00	*NA*	0.08
Komnzo	0.65	0.46	*NA*	*NA*	0.00	0.68	0.55	*NA*	*NA*	0.00	0.57	0.49	*NA*	*NA*	0.29
Lele	0.64	0.36	0.50	0.00	0.05	0.68	0.48	0.75	0.00	0.07	0.65	0.48	0.67	1.00	0.06
Mauwake	0.58	0.46	*NA*	*NA*	0.00	0.65	0.62	*NA*	*NA*	0.17	0.64	0.63	*NA*	*NA*	0.22
Mian	0.66	0.51	0.67	*NA*	0.00	0.65	0.65	0.00	*NA*	0.00	0.59	0.60	0.00	*NA*	0.50
Moloko	0.53	0.35	0.80	*NA*	0.09	0.58	0.29	0.40	*NA*	0.09	0.50	0.26	0.00	*NA*	0.04
Nese	0.73	0.38	0.40	0.00	0.00	0.75	0.48	0.00	0.00	0.00	0.67	0.35	0.57	1.00	0.00
Paluai	0.63	0.34	0.00	*NA*	0.25	0.66	0.48	0.57	*NA*	0.29	0.58	0.48	0.75	*NA*	0.00
Papuan Malay	0.77	0.40	0.50	0.00	0.00	0.76	0.55	0.00	1.00	0.00	0.69	0.42	0.00	0.50	0.00
Paunaka	0.70	0.56	0.00	*NA*	0.00	0.68	0.51	1.00	*NA*	0.00	0.56	0.58	1.00	*NA*	0.00
Pite Saami	0.74	0.31	*NA*	*NA*	*NA*	0.70	0.50	*NA*	*NA*	*NA*	0.66	0.49	*NA*	*NA*	*NA*
Ulwa	0.72	0.43	0.33	0.00	0.00	0.73	0.49	0.75	0.00	0.00	0.61	0.41	0.57	0.00	0.22
Yakkha	0.67	0.41	*NA*	*NA*	0.22	0.71	0.53	*NA*	*NA*	0.14	0.66	0.54	*NA*	*NA*	0.12

Table 7: Per-class F1 scores for `Qwen2.5-7B`. *NA* indicates the class does not appear in the gold labels for that language.

Language	IGT					Grammar					Grammar+IGT				
	0	1	2	3	?	0	1	2	3	?	0	1	2	3	?
Aguaruna	0.66	0.51	0.00	0.00	0.11	0.69	0.61	0.67	1.00	0.29	0.50	0.58	0.67	0.50	0.43
Bargam	0.56	0.49	0.00	*NA*	0.27	0.62	0.62	0.67	*NA*	0.37	0.45	0.54	0.86	*NA*	0.33
Belep	0.63	0.41	0.57	*NA*	0.20	0.65	0.51	0.33	*NA*	0.18	0.48	0.50	0.50	*NA*	0.15
Chakali	0.68	0.41	0.67	*NA*	0.06	0.52	0.44	0.40	*NA*	0.09	0.36	0.51	0.00	*NA*	0.10
Chacobo	0.53	0.57	1.00	0.00	0.14	0.52	0.62	0.00	0.50	0.17	0.47	0.62	0.00	0.00	0.29
Eastern Geshiza	0.65	0.57	0.00	*NA*	0.07	0.58	0.62	0.50	*NA*	0.11	0.50	0.73	0.00	*NA*	0.21
Hup	0.59	0.54	0.67	*NA*	0.06	0.61	0.64	0.50	*NA*	0.18	0.43	0.54	0.67	*NA*	0.13
Ik	0.72	0.34	0.67	0.00	0.09	0.73	0.67	0.80	0.00	0.14	0.63	0.63	0.89	0.67	0.12
Japhug	0.66	0.60	0.00	0.00	0.08	0.69	0.70	0.80	0.00	0.17	0.58	0.66	0.40	0.50	0.17
Kagayanen	0.59	0.38	0.00	0.00	0.29	0.61	0.51	0.67	0.00	0.23	0.53	0.48	0.67	0.80	0.30
Kalamang	0.53	0.26	0.00	*NA*	0.20	0.59	0.29	0.33	*NA*	0.31	0.48	0.30	0.33	*NA*	0.37
Kam	0.67	0.42	0.00	*NA*	0.00	0.71	0.58	0.80	*NA*	0.00	0.55	0.57	0.67	*NA*	0.00
Karbi	0.64	0.39	0.50	*NA*	0.26	0.61	0.50	0.00	*NA*	0.36	0.44	0.43	0.00	*NA*	0.30
Komnzo	0.62	0.47	*NA*	*NA*	0.00	0.63	0.63	*NA*	*NA*	0.04	0.61	0.65	*NA*	*NA*	0.14
Lele	0.66	0.38	0.57	0.00	0.19	0.65	0.61	0.67	0.00	0.32	0.54	0.54	0.80	0.00	0.29
Mauwake	0.59	0.54	*NA*	*NA*	0.00	0.68	0.76	*NA*	*NA*	0.08	0.62	0.72	*NA*	*NA*	0.08
Mian	0.57	0.43	0.67	*NA*	0.11	0.49	0.55	0.67	*NA*	0.07	0.43	0.67	0.00	*NA*	0.00
Moloko	0.51	0.30	0.40	*NA*	0.21	0.50	0.46	0.00	*NA*	0.33	0.47	0.35	0.00	*NA*	0.37
Nese	0.65	0.24	0.33	0.00	0.00	0.69	0.59	0.67	1.00	0.04	0.51	0.47	0.00	0.50	0.06
Paluai	0.64	0.43	0.67	*NA*	0.00	0.74	0.61	0.86	*NA*	0.12	0.62	0.54	0.33	*NA*	0.04
Papuan Malay	0.82	0.45	0.50	0.00	0.00	0.76	0.73	0.50	0.67	0.11	0.76	0.70	0.50	0.50	0.11
Paunaka	0.69	0.57	0.00	*NA*	0.00	0.73	0.62	0.00	*NA*	0.00	0.59	0.55	0.00	*NA*	0.00
Pite Saami	0.72	0.50	*NA*	*NA*	*NA*	0.76	0.68	*NA*	*NA*	*NA*	0.61	0.60	*NA*	*NA*	*NA*
Ulwa	0.70	0.45	0.80	1.00	0.13	0.75	0.59	0.86	0.00	0.15	0.66	0.59	0.75	0.00	0.09
Yakkha	0.67	0.56	*NA*	*NA*	0.24	0.73	0.64	*NA*	*NA*	0.11	0.62	0.59	*NA*	*NA*	0.10

Table 8: Per-class F1 scores for `Qwen3.5-9B`. *NA* indicates the class does not appear in the gold labels for that language.

Language	IGT					Grammar					Grammar+IGT				
	0	**1**	**2**	**3**	**?**	**0**	**1**	**2**	**3**	**?**	**0**	**1**	**2**	**3**	**?**
Aguaruna	0.78	0.63	0.00	0.00	0.17	0.80	0.75	1.00	1.00	0.45	0.76	0.65	0.00	0.67	0.33
Bargam	0.70	0.44	0.00	*NA*	0.18	0.70	0.59	0.80	*NA*	0.30	0.67	0.60	0.00	*NA*	0.31
Belep	0.80	0.47	0.40	*NA*	0.12	0.80	0.69	0.00	*NA*	0.13	0.80	0.72	0.67	*NA*	0.32
Chakali	0.86	0.55	0.00	*NA*	0.13	0.77	0.53	0.86	*NA*	0.12	0.67	0.44	0.40	*NA*	0.13
Chacobo	0.70	0.57	0.00	0.00	0.05	0.76	0.56	1.00	0.67	0.19	0.71	0.59	0.00	0.40	0.27
Eastern Geshiza	0.82	0.64	0.50	*NA*	0.00	0.80	0.66	1.00	*NA*	0.23	0.74	0.69	0.50	*NA*	0.05
Hup	0.70	0.50	0.67	*NA*	0.15	0.79	0.58	0.80	*NA*	0.14	0.69	0.56	0.50	*NA*	0.26
Ik	0.78	0.22	0.50	0.00	0.13	0.85	0.69	1.00	0.00	0.29	0.82	0.72	0.91	0.00	0.25
Japhug	0.76	0.67	0.00	0.00	0.10	0.79	0.72	0.00	0.67	0.15	0.75	0.69	1.00	1.00	0.11
Kagayanen	0.78	0.51	0.00	0.67	0.23	0.80	0.55	0.67	0.00	0.00	0.74	0.50	0.67	0.50	0.23
Kalamang	0.61	0.41	0.40	*NA*	0.28	0.64	0.41	0.67	*NA*	0.17	0.63	0.34	0.57	*NA*	0.21
Kam	0.84	0.42	0.89	*NA*	0.00	0.83	0.61	0.80	*NA*	0.00	0.81	0.66	1.00	*NA*	0.00
Karbi	0.75	0.52	0.50	*NA*	0.16	0.74	0.48	0.00	*NA*	0.15	0.68	0.39	0.00	*NA*	0.25
Komnzo	0.76	0.47	*NA*	*NA*	0.14	0.87	0.63	*NA*	*NA*	0.00	0.83	0.62	*NA*	*NA*	0.00
Lele	0.71	0.41	0.33	0.00	0.13	0.78	0.68	1.00	0.00	0.21	0.70	0.66	0.57	0.00	0.18
Mauwake	0.68	0.55	*NA*	*NA*	0.00	0.86	0.81	*NA*	*NA*	0.00	0.82	0.78	*NA*	*NA*	0.00
Mian	0.72	0.58	1.00	*NA*	0.00	0.92	0.82	1.00	*NA*	0.36	0.76	0.63	1.00	*NA*	0.00
Moloko	0.62	0.44	0.40	*NA*	0.28	0.70	0.44	0.33	*NA*	0.21	0.67	0.53	0.57	*NA*	0.21
Nese	0.82	0.51	0.86	0.67	0.00	0.85	0.64	0.86	0.00	0.00	0.81	0.59	0.86	1.00	0.11
Paluai	0.78	0.42	0.86	*NA*	0.00	0.81	0.61	0.86	*NA*	0.00	0.80	0.65	0.67	*NA*	0.11
Papuan Malay	0.89	0.62	0.80	0.67	0.09	0.89	0.71	1.00	1.00	0.00	0.89	0.66	1.00	0.67	0.13
Paunaka	0.77	0.55	0.00	*NA*	0.00	0.85	0.56	0.67	*NA*	0.00	0.81	0.71	0.00	*NA*	0.50
Pite Saami	0.83	0.59	*NA*	*NA*	*NA*	0.87	0.74	*NA*	*NA*	*NA*	0.86	0.73	*NA*	*NA*	*NA*
Ulwa	0.85	0.55	0.80	1.00	0.08	0.87	0.65	1.00	0.00	0.14	0.83	0.63	0.80	0.00	0.11
Yakkha	0.80	0.57	*NA*	*NA*	0.17	0.86	0.67	*NA*	*NA*	0.27	0.84	0.73	*NA*	*NA*	0.19

Table 9: Per-class F1 scores for `Qwen3.6-27B`. *NA* indicates the class does not appear in the gold labels for that language.

Language	IGT					Grammar					Grammar+IGT				
	0	**1**	**2**	**3**	**?**	**0**	**1**	**2**	**3**	**?**	**0**	**1**	**2**	**3**	**?**
Aguaruna	0.76	0.62	1.00	0.00	0.06	0.80	0.70	1.00	0.67	0.41	0.78	0.74	0.00	0.50	0.22
Bargam	0.73	0.51	0.86	*NA*	0.26	0.74	0.65	1.00	*NA*	0.27	0.70	0.49	1.00	*NA*	0.15
Belep	0.76	0.49	1.00	*NA*	0.27	0.83	0.67	0.86	*NA*	0.17	0.77	0.71	0.80	*NA*	0.46
Chakali	0.80	0.42	0.80	*NA*	0.09	0.82	0.64	0.80	*NA*	0.20	0.76	0.62	1.00	*NA*	0.08
Chacobo	0.69	0.65	0.00	0.00	0.24	0.71	0.69	0.00	0.67	0.17	0.67	0.71	0.00	0.57	0.23
Eastern Geshiza	0.82	0.66	0.80	*NA*	0.00	0.86	0.77	0.80	*NA*	0.00	0.79	0.73	0.80	*NA*	0.00
Hup	0.69	0.51	0.50	*NA*	0.19	0.74	0.68	1.00	*NA*	0.00	0.66	0.70	1.00	*NA*	0.11
Ik	0.74	0.41	0.50	0.00	0.17	0.85	0.73	0.91	0.00	0.10	0.82	0.68	0.89	0.00	0.14
Japhug	0.73	0.63	1.00	0.00	0.12	0.78	0.74	0.50	0.00	0.00	0.72	0.71	0.50	0.00	0.14
Kagayanen	0.79	0.62	0.00	0.00	0.37	0.76	0.58	0.00	0.50	0.08	0.76	0.50	0.00	0.00	0.07
Kalamang	0.58	0.42	0.40	*NA*	0.23	0.64	0.42	0.80	*NA*	0.11	0.63	0.35	0.40	*NA*	0.11
Kam	0.82	0.53	0.86	*NA*	0.00	0.82	0.65	0.80	*NA*	0.00	0.78	0.67	0.75	*NA*	0.00
Karbi	0.67	0.50	0.50	*NA*	0.13	0.69	0.56	0.00	*NA*	0.00	0.62	0.50	0.50	*NA*	0.11
Komnzo	0.72	0.55	*NA*	*NA*	0.08	0.81	0.68	*NA*	*NA*	0.00	0.80	0.71	*NA*	*NA*	0.11
Lele	0.74	0.51	0.75	0.00	0.18	0.79	0.70	0.75	0.00	0.22	0.80	0.71	0.89	0.00	0.31
Mauwake	0.75	0.60	*NA*	*NA*	0.08	0.83	0.85	*NA*	*NA*	0.00	0.77	0.81	*NA*	*NA*	0.00
Mian	0.67	0.59	1.00	*NA*	0.00	0.74	0.79	1.00	*NA*	0.22	0.75	0.75	1.00	*NA*	0.00
Moloko	0.65	0.49	0.57	*NA*	0.24	0.63	0.45	0.40	*NA*	0.17	0.60	0.44	0.40	*NA*	0.02
Nese	0.82	0.47	0.80	0.00	0.08	0.85	0.68	0.33	0.00	0.00	0.82	0.70	1.00	0.00	0.00
Paluai	0.79	0.44	0.86	*NA*	0.00	0.83	0.67	0.67	*NA*	0.00	0.87	0.77	0.67	*NA*	0.00
Papuan Malay	0.92	0.64	1.00	0.00	0.14	0.87	0.68	0.67	0.00	0.00	0.90	0.73	0.80	0.00	0.12
Paunaka	0.78	0.64	1.00	*NA*	0.20	0.79	0.59	0.00	*NA*	0.25	0.73	0.63	0.67	*NA*	0.00
Pite Saami	0.84	0.58	*NA*	*NA*	*NA*	0.86	0.72	*NA*	*NA*	*NA*	0.87	0.76	*NA*	*NA*	*NA*
Ulwa	0.85	0.62	0.67	0.00	0.15	0.84	0.69	0.86	0.00	0.12	0.77	0.64	0.80	0.00	0.00
Yakkha	0.77	0.61	*NA*	*NA*	0.14	0.83	0.70	*NA*	*NA*	0.40	0.83	0.74	*NA*	*NA*	0.14

Table 10: Per-class F1 scores for `Gemma4-31B`. *NA* indicates the class does not appear in the gold labels for that language.

Language	IGT					Grammar					Grammar+IGT				
	0	**1**	**2**	**3**	**?**	**0**	**1**	**2**	**3**	**?**	**0**	**1**	**2**	**3**	**?**
Aguaruna	0.77	0.58	0.67	0.00	0.41	0.75	0.67	1.00	0.67	0.33	0.70	0.67	1.00	0.50	0.30
Bargam	0.63	0.41	0.86	*NA*	0.43	0.76	0.66	0.57	*NA*	0.13	0.64	0.60	0.86	*NA*	0.31
Belep	0.64	0.41	0.67	*NA*	0.17	0.79	0.70	0.75	*NA*	0.32	0.76	0.70	0.86	*NA*	0.48
Chakali	0.76	0.42	0.86	*NA*	0.10	0.77	0.55	1.00	*NA*	0.13	0.69	0.44	0.86	*NA*	0.12
Chacobo	0.65	0.50	0.00	0.00	0.13	0.68	0.63	1.00	0.67	0.25	0.72	0.68	0.00	0.33	0.30
Eastern Geshiza	0.70	0.52	0.67	*NA*	0.07	0.85	0.73	0.80	*NA*	0.16	0.72	0.69	0.80	*NA*	0.18
Hup	0.71	0.50	0.00	*NA*	0.21	0.77	0.69	1.00	*NA*	0.14	0.67	0.62	0.40	*NA*	0.34
Ik	0.68	0.23	0.80	0.00	0.11	0.85	0.69	0.89	0.00	0.19	0.81	0.72	0.83	0.00	0.14
Japhug	0.66	0.54	0.40	0.00	0.12	0.80	0.77	0.67	0.00	0.15	0.64	0.67	0.00	0.50	0.18
Kagayanen	0.65	0.35	0.40	0.00	0.22	0.74	0.48	0.67	0.00	0.04	0.71	0.53	0.67	0.67	0.08
Kalamang	0.61	0.29	0.57	*NA*	0.36	0.64	0.34	0.57	*NA*	0.14	0.63	0.34	0.67	*NA*	0.24
Kam	0.73	0.39	0.75	*NA*	0.03	0.83	0.64	0.89	*NA*	0.00	0.80	0.66	0.86	*NA*	0.00
Karbi	0.68	0.44	0.00	*NA*	0.31	0.71	0.53	0.00	*NA*	0.33	0.60	0.49	0.67	*NA*	0.33
Komnzo	0.63	0.44	*NA*	*NA*	0.09	0.81	0.65	*NA*	*NA*	0.00	0.79	0.65	*NA*	*NA*	0.18
Lele	0.72	0.37	0.57	0.00	0.30	0.78	0.66	0.89	0.00	0.13	0.75	0.67	0.89	0.00	0.33
Mauwake	0.59	0.45	*NA*	*NA*	0.00	0.80	0.77	*NA*	*NA*	0.00	0.78	0.80	*NA*	*NA*	0.13
Mian	0.68	0.68	0.67	*NA*	0.00	0.80	0.73	1.00	*NA*	0.00	0.70	0.68	1.00	*NA*	0.00
Moloko	0.58	0.43	0.57	*NA*	0.39	0.69	0.52	0.57	*NA*	0.10	0.56	0.45	0.57	*NA*	0.09
Nese	0.74	0.27	0.86	0.00	0.06	0.83	0.67	1.00	0.00	0.29	0.78	0.62	0.80	0.00	0.07
Paluai	0.68	0.37	0.57	*NA*	0.03	0.80	0.64	0.67	*NA*	0.11	0.77	0.63	1.00	*NA*	0.08
Papuan Malay	0.84	0.51	1.00	0.00	0.00	0.89	0.79	0.80	0.67	0.27	0.88	0.68	0.80	0.67	0.12
Paunaka	0.74	0.52	0.00	*NA*	0.11	0.76	0.53	0.00	*NA*	0.14	0.69	0.60	0.67	*NA*	0.22
Pite Saami	0.77	0.58	*NA*	*NA*	*NA*	0.90	0.76	*NA*	*NA*	*NA*	0.84	0.74	*NA*	*NA*	*NA*
Ulwa	0.73	0.38	0.67	0.00	0.12	0.87	0.63	0.86	0.00	0.09	0.79	0.62	0.00	0.00	0.25
Yakkha	0.70	0.55	*NA*	*NA*	0.00	0.81	0.70	*NA*	*NA*	0.21	0.75	0.63	*NA*	*NA*	0.20

Table 11: Per-class F1 scores for `GPT5.4-mini`. *NA* indicates the class does not appear in the gold labels for that language.

ID	Lbl	Count	Pr.	ID	Lbl	Count	Pr.	ID	Lbl	Count	Pr.
GB020	0	1374/2403	0.57	GB113	1	1316/2386	0.55	GB263	0	1065/2233	0.48
GB021	0	1920/2425	0.79	GB114	0	1101/2397	0.46	GB264	0	1561/2231	0.70
GB022	0	1644/2370	0.69	GB115	1	1135/2405	0.47	GB265	?	906/2189	0.41
GB023	0	1687/2376	0.71	GB116	0	1846/2310	0.80	GB266	?	934/2186	0.43
GB024	2	1087/2415	0.45	GB117	1	1152/2407	0.48	GB270	0	1039/2180	0.48
GB025	2	1066/2425	0.44	GB118	1	840/2410	0.35	GB273	?	950/2189	0.43
GB026	0	1660/2340	0.71	GB119	0	1356/2195	0.62	GB275	0	1051/2188	0.48
GB027	1	1065/2421	0.44	GB120	0	1301/2200	0.59	GB276	?	888/2188	0.41
GB028	1	1243/2451	0.51	GB121	0	1522/2216	0.69	GB285	0	1689/2198	0.77
GB030	0	1804/2438	0.74	GB122	0	1103/2405	0.46	GB286	0	1460/2247	0.65
GB031	0	1718/2394	0.72	GB123	?	1124/2388	0.47	GB291	0	1635/2158	0.76
GB035	0	1111/2428	0.46	GB124	0	1390/2384	0.58	GB296	?	1164/2137	0.54
GB036	0	2005/2395	0.84	GB126	1	1191/2366	0.50	GB297	0	1513/2141	0.71
GB037	0	1922/2389	0.80	GB127	0	1109/2339	0.47	GB298	0	1795/2189	0.82
GB038	0	2080/2365	0.88	GB129	0	2061/2361	0.87	GB299	1	1342/2194	0.61
GB039	0	1625/2366	0.69	GB130	1	1827/2415	0.76	GB300	0	1152/2126	0.54
GB041	0	1600/2352	0.68	GB131	0	1878/2448	0.77	GB301	?	997/2127	0.47
GB042	0	1916/2442	0.78	GB132	0	1216/2433	0.50	GB302	0	1439/2129	0.68
GB043	0	2117/2398	0.88	GB133	0	1322/2446	0.54	GB303	0	1524/2129	0.72
GB044	1	1282/2398	0.53	GB134	1	1644/2361	0.70	GB304	0	797/2131	0.37
GB046	0	1148/2328	0.49	GB135	1	1422/2360	0.60	GB305	0	855/2187	0.39
GB047	1	1182/2418	0.49	GB136	1	1245/2425	0.51	GB306	0	1170/2131	0.55
GB048	1	1090/2417	0.45	GB137	0	1364/2382	0.57	GB309	0	1325/2172	0.61
GB049	1	1022/2414	0.42	GB138	0	1299/2377	0.55	GB312	1	1344/2133	0.63
GB051	0	1796/2383	0.75	GB139	1	1517/2442	0.62	GB313	0	1233/2131	0.58
GB052	0	2073/2428	0.85	GB140	0	796/2376	0.34	GB314	0	1721/2130	0.81
GB053	0	1667/2381	0.70	GB146	?	983/2351	0.42	GB315	0	1690/2131	0.79
GB054	0	2036/2408	0.85	GB147	0	1114/2431	0.46	GB316	0	1908/2131	0.90
GB057	0	1655/2435	0.68	GB148	0	1649/2407	0.69	GB317	0	1929/2131	0.91
GB058	0	2013/2428	0.83	GB149	0	1869/2352	0.79	GB318	0	1676/2181	0.77
GB059	0	1070/2387	0.45	GB150	0	1075/2395	0.45	GB319	0	2016/2132	0.95
GB065	1	1040/2420	0.43	GB151	0	1485/2401	0.62	GB320	0	1993/2132	0.93
GB068	0	1002/2436	0.41	GB152	0	1126/2405	0.47	GB321	0	1745/2132	0.82
GB069	0	1454/2381	0.61	GB155	1	1494/2425	0.62	GB322	0	1204/2185	0.55
GB070	0	1485/2377	0.62	GB156	0	1740/2370	0.73	GB323	0	1000/2186	0.46
GB071	0	1209/2373	0.51	GB158	1	1387/2385	0.58	GB324	0	1622/2131	0.76
GB072	0	1184/2375	0.50	GB159	1	866/2371	0.37	GB325	0	1099/2128	0.52
GB073	0	1143/2363	0.48	GB160	1	1153/2377	0.49	GB326	1	941/2182	0.43
GB074	1	1245/2443	0.51	GB165	0	2125/2214	0.96	GB327	1	1429/2188	0.65
GB075	0	1098/2390	0.46	GB166	0	2090/2213	0.94	GB328	0	1106/2187	0.51
GB079	1	1468/2313	0.63	GB167	0	1488/2195	0.68	GB329	0	1234/2184	0.57
GB080	1	1894/2365	0.80	GB170	0	1667/2264	0.74	GB330	0	1235/2172	0.57
GB081	0	1828/2359	0.77	GB171	0	1622/2243	0.72	GB331	0	1263/2172	0.58
GB082	0	1482/2392	0.62	GB172	0	1935/2228	0.87	GB333	1	1234/2189	0.56
GB083	1	1189/2393	0.50	GB177	0	1742/2196	0.79	GB334	0	1394/2188	0.64
GB084	0	1244/2402	0.52	GB184	0	1199/2202	0.54	GB335	0	1400/2132	0.66
GB086	1	1424/2339	0.61	GB185	0	1103/2211	0.50	GB336	0	1650/2130	0.77
GB089	0	1379/2391	0.58	GB186	0	1686/2195	0.77	GB400	0	1400/1893	0.74
GB090	0	1352/2396	0.56	GB187	0	813/2194	0.37	GB401	?	1050/1891	0.56
GB091	0	1464/2404	0.61	GB188	0	1087/2183	0.50	GB402	0	1065/1889	0.56
GB092	0	1430/2405	0.59	GB192	0	1953/2228	0.88	GB403	0	965/1889	0.51
GB093	0	1520/2410	0.63	GB193	2	1237/2262	0.55	GB408	0	978/1909	0.51
GB094	0	1790/2407	0.74	GB196	0	2051/2215	0.93	GB409	0	1379/1919	0.72
GB095	0	1798/2362	0.76	GB197	0	2143/2216	0.97	GB410	1	1218/1890	0.64
GB096	0	1799/2357	0.76	GB198	0	1664/2208	0.75	GB415	0	1497/1955	0.77
GB098	0	1679/2359	0.71	GB203	2	795/2194	0.36	GB421	?	778/1889	0.41
GB099	0	1796/2351	0.76	GB204	?	1056/2183	0.48	GB422	0	984/1889	0.52
GB103	0	1364/2413	0.57	GB250	0	855/2174	0.39	GB430	0	1714/1961	0.87
GB104	0	1676/2398	0.70	GB252	0	1105/2171	0.51	GB431	0	1475/1961	0.75
GB105	1	1075/2394	0.45	GB253	0	1179/2170	0.54	GB432	0	1335/1960	0.68
GB107	0	1225/2416	0.51	GB254	0	865/2171	0.40	GB433	0	1195/1963	0.61
GB108	0	1306/2391	0.55	GB256	0	1114/2170	0.51	GB519	0	956/1898	0.50
GB109	0	1648/2390	0.69	GB257	1	835/2254	0.37	GB520	0	1027/1904	0.54
GB110	0	1692/2404	0.70	GB260	0	1706/2251	0.76	GB521	0	1282/1925	0.67
GB111	0	1318/2357	0.56	GB262	0	1513/2234	0.68	GB522	1	1135/1893	0.60

Table 12: Majority voting labels for all Grambank features used in this study. Count shows the number of languages coded with the majority label out of total valid codings. Pr. indicates the proportion.

System Hyperparameters		
Category	**Parameter**	**Value**
LLM	Temperature	0.1
	Max tokens	32768
	Force max tokens	True
	Dtype	bfloat16
Agent	Max iterations per feature	10
	Confidence threshold	0.75
	Min queries per feature	5
Retrieval	Embedding model	paraphrase-multilingual-MiniLM-L12-v2
	Embedding batch size	64
	BM25 k_1	1.5
	BM25 b	0.75
	Hybrid α (dense weight)	0.5
	MMR λ	0.6

Table 13: System hyperparameters. The agent uses a ReAct loop capped at 10 iterations per feature, with a confidence threshold of 0.75 for accepting conclusions. Retrieval combines BM25 and dense embeddings with equal weight ($\alpha = 0.5$), followed by MMR reranking ($\lambda = 0.6$) for diversity.

Domain	**Feature IDs**
Word Order (10)	GB074, GB075, GB130, GB131, GB132, GB133, GB134, GB135, GB136, GB260
Simple Clauses (66)	GB068, GB070, GB071, GB072, GB073, GB089, GB090, GB091, GB092, GB093, GB094, GB095, GB096, GB098, GB099, GB103, GB104, GB105, GB107, GB108, GB109, GB113, GB114, GB115, GB116, GB117, GB126, GB137, GB138, GB139, GB140, GB147, GB148, GB149, GB155, GB156, GB250, GB252, GB253, GB254, GB256, GB257, GB262, GB263, GB264, GB265, GB266, GB270, GB273, GB275, GB276, GB285, GB286, GB291, GB297, GB302, GB303, GB304, GB322, GB323, GB324, GB401, GB408, GB409, GB410, GB522
Nominal Categories (44)	GB020, GB021, GB028, GB030, GB031, GB035, GB036, GB037, GB038, GB039, GB041, GB042, GB043, GB044, GB046, GB051, GB052, GB053, GB054, GB057, GB058, GB165, GB166, GB167, GB184, GB185, GB186, GB187, GB188, GB192, GB196, GB197, GB198, GB313, GB314, GB315, GB316, GB317, GB318, GB319, GB320, GB321, GB325, GB415
Nominal Syntax (23)	GB022, GB023, GB024, GB025, GB026, GB027, GB059, GB065, GB069, GB170, GB171, GB172, GB193, GB203, GB204, GB301, GB305, GB306, GB326, GB430, GB431, GB432, GB433
Verbal Categories (19)	GB082, GB083, GB084, GB086, GB110, GB111, GB119, GB120, GB121, GB146, GB177, GB298, GB299, GB309, GB312, GB400, GB519, GB520, GB521
Morphology (11)	GB047, GB048, GB049, GB079, GB080, GB081, GB122, GB124, GB158, GB159, GB160
Complex Sentences (11)	GB118, GB150, GB151, GB152, GB327, GB328, GB329, GB330, GB331, GB421, GB422
Lexicon (11)	GB123, GB127, GB129, GB296, GB300, GB333, GB334, GB335, GB336, GB402, GB403

Table 14: Grambank features used in this study, grouped by domain (195 total).

Query `GB049 · NMZObject · Coding{0, 1, ?}`

Is there a productive morphological pattern for deriving an object noun from a verb?

Phenomena: [object nominalization, noun–verb compounds, derivational morphology]

Iter 1 `search_text("object nominal patient derivation", top_k=5)`

Thought: The current evidence is contradictory regarding the productivity of nominalization prefixes and their semantic scope (action vs. object). Claim C003 suggests they form action nominals, while the query asks for object nouns. Claim C001 suggests Object–Verb compounds are productive, but it is unclear if these are considered "derived from a verb" in the strict morphological sense or just compounding.

Finding: This will identify if there is a specific derivational process for object nouns distinct from action nominals or compounds.

Iter 2 `extract_author_claims("Nominal morphology > Nominal derivations")`

Thought: The current evidence is contradictory and low confidence. Supporting claims suggest Object–Verb compounds are productive, while contradicting claims suggest nominalizing prefixes are not fully productive. Crucially, the query asks for deriving an *object* noun from a verb.

Finding: This will reveal if the grammar describes a specific derivational process for creating object/patient nouns from verbs, distinct from compounding or action nominalization.

⋮

Repeated calls

Iter 10 `extract_author_claims("Nominal morphology > Nominal derivations")`

Thought: The evidence so far indicates that Object–Verb compounds are productive (C001), but the query asks about deriving an object noun *from a verb*. Compounding involves a noun and a verb, not deriving a noun solely from a verb. The prefixation of nominalizing prefixes creates action nominals or abstract nouns, not necessarily object nouns (patient/result).

Finding: This will reveal if the grammar describes a specific derivational process for creating object nouns from verbs, distinct from compounding or action nominalization.

Conclude `predict_label`

Label reasoning: The evidence indicates that object nominalization is primarily achieved through compounding (Object–Verb compounds) rather than a specific morphological element attaching to verbs to form an object nominalization; the available affixes form action/abstract nominals and are not fully productive for object derivation.

Answer: In Japhug, there is no single, fully productive morphological affix dedicated exclusively to deriving object nouns from verbs. Instead, the language primarily utilizes a highly productive syntactic pattern of Object–Verb nominal compounds to express this relationship.

Output 27 LLM calls · 46,228 tokens (40,366 in / 5,862 out)

Label: `0` **Confidence:** `0.10`

Audit: `overturned` **Human review:** `true`

Review reason: The answer is factually incorrect as the evidence explicitly identifies productive Object–Verb nominal compounds.

Audit objections: Evidence cites productive OV compounds, contradicting "no pattern"; answer ignores supporting claims C001–C003.

Figure 6: Agent trace for Grambank feature **GB049** (*NMZObject*) in Japhug, produced by the grammar-grounded pipeline with Qwen3.6-27B. The trace shows that the system predicts label `0` with low confidence, flags the case for human review, and the audit overturns the conclusion.

Query U30 · Implication

If the verb has categories of person-number or of gender, it always has tense-mode categories.

Domain: TENSE_ASPECT_MODALITY; logic: antecedent ⇒ consequent; languages: 10

Plan parse_universal(U30)

Antecedent: verb has categories of person-number or of gender.

Consequent: verb also has tense-mode categories.

Parsing rationale: from CSV

Language assessments aggregate_language_verdicts

Assessment	Languages
VALID_SUPPORT	Aguaruna, Chácobo, Moloko, Komnzo, Pite Saami
VALID_VIOLATION	Ik
ANTECEDENT_NA	Papuan Malay, Kalamang
INSUFFICIENT	Chakali, Eastern Geshiza

Counts: support = 5; violation = 1; antecedent N/A = 2; insufficient = 2.

Counterexample inspect_language("Ik")

Antecedent evidence: Ik verbs are richly inflected for person and number, utilizing a complex system of prefixes and suffixes for subject and object agreement.

Consequent evidence: Ik lacks a dedicated, obligatory tense paradigm; temporal distinctions are conveyed through aspectual markers, auxiliaries, or context rather than fused tense-mode.

Assessment: VALID_VIOLATION; confidence = 0.85.

Conclude conclude

Conclusion: The universal is false because there is at least one valid violation where a language exhibits person-number agreement on the verb but lacks obligatory tense-mode marking. While many languages support the correlation, the existence of counterexamples like Ik demonstrates that person-number agreement does not strictly necessitate tense-mode categories.

Cross-linguistic notes: Ik serves as a critical counterexample. In contrast, Aguaruna, Chácobo, Moloko, Komnzo, and Pite Saami support the universal by exhibiting both features. Papuan Malay and Kalamang are irrelevant to the test because they lack the requisite person-number agreement on verbs.

Output 10 languages evaluated

Verdict:	FALSE	**Confidence:**	0.95
Support:	5	**Violation:**	1
Antecedent N/A:	2	**Insufficient:**	2

Figure 7: Agent trace for typological hypothesis testing on Greenberg Universal **U30**. The system parses the universal as an implication, evaluates the antecedent and consequent across 10 languages, identifies Ik as a valid counterexample, and concludes that the universal is false with confidence 0.95.

(a) Domain Extraction Prompt

```
You are a linguistic typologist. Below is the
table of contents of a reference grammar for
{language} with section summaries, and
quantitative IGT corpus statistics.

TABLE OF CONTENTS WITH SECTION SUMMARIES:
{toc_with_summaries}

QUANTITATIVE IGT SUMMARY:
{igt_summary}

GLOSS ABBREVIATION LEGEND:
{abbrev_legend}

Task: Identify major typological domains and
generate 3-5 specific, answerable feature
questions per domain grounded in (a) section
summaries and (b) IGT tag frequencies.

Heuristics:
- PST/PFV frequent -> generate TMA questions
- EVID at 0% -> EVIDENTIALITY likely absent
- No agreement tags -> AGREEMENT may be "No"

Output ONLY valid JSON:
{
"domains": [{
"domain_id": "D001",
"domain_name": "TENSE_ASPECT_MODALITY",
"relevant_sections": ["Verbs > Tense"],
"igt_signals": ["PST: 140 (8.4%)"],
"candidate_features": [{
"feature_id": "F001",
"question": "Does {language} grammaticalize
 tense?",
"igt_tags_to_check": ["PST","FUT","PRF"],
"prior": "likely_yes"
}]
}]
}
```

(b) ReAct Action-Selection Prompt

```
You are deep-searching a reference grammar
for a typological feature.

LANGUAGE: {language}
FEATURE QUESTION: {question}
CURRENT EVIDENCE GRAPH: {evidence_summary}
EVIDENCE GAPS: {gap_analysis}
Progress: iteration {iteration}/{max_iter}

AVAILABLE TOOLS:
 1. read_full_section(query)
 2. follow_cross_references(query)
 3. extract_author_claims(query)
 4. search_text(query, top_k)
 5. analyse_tag(tag)
 6. analyse_construction(tags)
 7. analyse_absence(category)
 8. compare_tags(tag_a, tag_b)
 9. get_section_igt(query)
10. search_translations(query)
11. get_triline_examples(query)
12. conclude[only when ALL constraints met]

Gap-to-tool guidance:
- NO IGT EVIDENCE-> analyse_tag / absence
- NO GRAMMAR PROSE -> read_full_section
- CONTRADICTIONS -> follow_cross_references

Output ONLY valid JSON:
{
"thought": "what gap this action fills",
"action": "tool_name",
"args": {},
"evidence_type": "grammar_prose|igt_quant|...",
"claim_to_add": "one-sentence claim",
"supports_hypothesis": true | false | null
}
```

Figure 8: (a) Domain extraction prompt: the model reads grammar section summaries and IGT tag frequencies to discover typological domains and generate candidate feature questions. (b) ReAct action-selection prompt: at each iteration the agent receives its evidence graph and gap analysis, then selects one of 11 tools; `conclude` is permitted only when all required evidence types are present.

(c) Contradiction Resolution Prompt

```
You are resolving a contradiction in the
evidence for a typological feature.

LANGUAGE: {language}
FEATURE QUESTION: {question}

CLAIM A: {claim_a_text}
Source: {claim_a_source}

CLAIM B: {claim_b_text}
Source: {claim_b_source}

Consider:
- Is one source more authoritative?
- Could both be true (variation, different
construction types)?
- Does one apply to a restricted context
the other does not mention?

Output ONLY valid JSON:
{
"resolution": "how contradiction is resolved",
"resolved": true | false,
"preferred_claim": "A"|"B"|"both_partial"
 |"neither",
"revised_confidence_a": 0.0-1.0,
"revised_confidence_b": 0.0-1.0,
"synthesis": "one sentence: nuanced truth"
}
```

(d) Post-hoc Auditor Prompt

```
You are a critical auditor reviewing a
typological conclusion.

LANGUAGE: {language}
FEATURE: {question}
CONCLUSION: {value} -- "{value_detail}"
CONFIDENCE: {confidence}
EVIDENCE GRAPH SUMMARY:
{evidence_graph_summary}

Try to DISPROVE this conclusion using only
the evidence listed:
- Are IGT patterns diagnostic or correlated?
- Could grammar prose describe marginal usage?
- Are contradictions ignored in conclusion?
- Should "Yes" be "Partial" given evidence?

Output ONLY valid JSON:
{
"verdict": "upheld"|"weakened"|"overturned",
"objections": [
"objection 1 (max 60 chars)",
"objection 2 (max 60 chars)"
],
"revised_value": "same or corrected",
"revised_confidence": 0.0-1.0,
"audit_notes": "one sentence summary"
}
```

Figure 9: (c) Contradiction resolution prompt: invoked when the evidence graph detects a conflict between two claims; the ReAct loop pauses until resolution is complete. (d) Post-hoc auditor prompt: an adversarial pass that attempts to overturn or weaken the conclusion; the verdict and revised confidence are stored in the final feature record.

(e) Feature Conclusion Prompt

```
You are writing a final typological feature
entry.

LANGUAGE: {language}
FEATURE QUESTION: {question}
COMPLETE EVIDENCE GRAPH: {evidence_graph}
CHUNK SUMMARIES: {chunk_summaries}

Evidence weighting:
- GRAMMAR_STATEMENT > IGT_PATTERN alone
- ABSENCE_EVIDENCE with zero counts =
strong negative evidence
- Unresolved contradictions -> lower
confidence, flag for review
- "Yes" requires BOTH grammar prose AND
IGT quantitative support

Output ONLY valid JSON:
{
"linguistic_definition": "lang-agnostic
definition (1-2 sentences)",
"structural_description": "realization
in {language}: forms, positions",
"value": "Yes|No|Partial|Unclear|?",
"value_detail": "key markers + refs",
"confidence": 0.0-1.0,
"key_evidence": [
"Finding: ... Source: Sec [chunk_id].
 Justification: ...",
"Finding: PST 140/1668 (8.4%), preverbal.
 Source: analyse_tag(PST)."
],
"typological_notes": "cross-ling. notes",
"needs_human_review": true | false,
"review_reason": "one sentence or empty"
}
```

(f) Grambank Labeler Prompt

```
You are a typology expert applying a
Grambank coding scheme to a grammar.

FEATURE: {grambank_id} -- {query}
CODING POLICY (apply EXACTLY):
{coding_policy}

EVIDENCE (language: {language})
Answer summary: {answer}
Key evidence items: {key_evidence}
Structural description: {structural_desc}
IGT examples cited: {igt_examples}

Task:
1. Note what grammatical behaviour each
 piece of evidence demonstrates.
2. Match evidence to coding criteria.
3. Select the label most directly
 supported by a concrete IGT example
 or explicit grammar statement.

Rules:
- Label must be exactly one code from
the Coding Policy (e.g. 0, 1, 2, ?).
- Use "?" only when evidence cannot
distinguish between substantive codes.

Output ONLY valid JSON:
{
"label": "<code from Coding Policy>",
"reasoning": "2-4 sentences citing
specific evidence items",
"confidence": 0.0-1.0
}
```

Figure 10: (e) Feature conclusion prompt: synthesizes the full evidence graph into a structured feature record; evidence items must cite grammar sections by chunk ID and IGT tools by name. (f) Grambank labeler prompt: maps a `QueryResult` to a Grambank categorical code by matching retrieved evidence against the official coding policy.

(g) IGT-Only Domain Extraction Prompt

```
You are a linguistic typologist. There is
NO reference grammar. Infer typological
domains purely from IGT corpus statistics.

LANGUAGE: {language}
IGT CORPUS STATISTICS: {igt_summary}
CONSTRUCTION PATTERNS: {construction_patterns}
GLOSS ABBREVIATION LEGEND: {abbrev_legend}

Reading the data:
- Tags >5% coverage-> likely grammaticalized
- Tags <0.5% -> marginal or absent
- Consistent position -> morphosyntactic slot
- Complementary tags (PST/FUT, SG/PL)
-> paradigm systems

Procedure:
1. Group tags by domain (TMA, Agreement,
 Case, Negation, ...)
2. For each domain: PRESENT (>5%),
 PARTIAL (<5%), ABSENT (near-zero)
3. Generate 3-5 feature questions per domain
4. Set prior: "likely_yes" / "likely_no"
/ "uncertain"

Output ONLY valid JSON:
{
"domains": [{
"domain_id": "D001",
"domain_name": "TENSE_ASPECT_MODALITY",
"igt_signals": ["PST: 140 (8.4%)",
"FUT: 0 (0.0%)"],
"candidate_features": [{
"feature_id": "F001",
"question": "Does {language}
 grammaticalize tense?",
"igt_tags_to_check": ["PST","FUT"],
"prior": "likely_yes"
}]
}]
}
```

(h) IGT-Only ReAct Decision Prompt

```
You are deep-searching an IGT corpus for a
typological feature. There is NO grammar.

LANGUAGE: {language}
FEATURE QUESTION: {question}
CURRENT EVIDENCE GRAPH: {evidence_summary}
EVIDENCE GAPS: {gap_analysis}
Progress: iteration {iteration}/{max_iter}

AVAILABLE TOOLS (IGT-only):
 1. get_tag_inventory
 2. get_construction_inventory
 3. find_tag_cluster(seed_tag)
 4. analyse_tag(tag)
 5. analyse_construction(tags)
 6. analyse_absence(category)
 7. compare_tags(tag_a, tag_b)
 8. get_section_igt(query)
 9. analyse_semantic_context(tag)
10. search_translations(query)
11. get_triline_examples(query)
12. analyse_morpheme_position(tag)
13. parse_example_structure(query)
14. get_morpheme_forms(tag)
15. conclude [only when ALL constraints met]

Gap-to-tool guidance:
- NO IGT EVIDENCE -> analyse_tag
- NO ABSENCE CHECK-> analyse_absence
- NO COUNTER-EV.-> compare_tags
- CONTRADICTIONS-> find_tag_cluster

Output ONLY valid JSON:
{
"thought": "what gap this action fills",
"action": "tool_name",
"args": {},
"evidence_type": "igt_quantitative|...",
"claim_to_add": "one-sentence claim",
"supports_hypothesis": true | false | null
}
```

Figure 11: (g) IGT-only domain extraction prompt: when no grammar is available, domain discovery is driven entirely by corpus statistics; high-frequency tags signal grammaticalized categories, complementary tags signal paradigms, and near-zero counts signal absence. (h) IGT-only ReAct decision prompt: mirrors the grammar-grounded variant but expands the tool set to 15 IGT-specific tools including morpheme position analysis and LLM-powered clause parsing.