

Using Large Language Models for Automated Corpus Annotation and Linguistic Analysis: A Critical Methodological Framework

Maria Ibrar ¹

¹ Riphah Institute of Language and Literature, Riphah International University, Islamabad, Pakistan

Received: 13 August 2026; **Accepted:** 26 August 2026; **Published:** 04 September 2026

Abstract: Large language models (LLMs) are increasingly used to classify, label, summarize, and interpret large text collections, creating new possibilities for corpus linguistics. Their capacity for zero-shot and few-shot instruction following could reduce the cost of linguistic annotation and extend analysis beyond the categories handled by conventional part-of-speech taggers, parsers, and dictionary-based tools. At the same time, LLM outputs are probabilistic, prompt-sensitive, model-dependent, and potentially biased, raising fundamental questions about measurement validity, annotation reliability, and reproducibility. This article critically synthesizes foundational corpus-annotation principles with recent evidence on LLM-based text annotation and develops a validated human–LLM workflow for corpus research. The framework distinguishes token-, span-, sentence-, document-, and discourse-level annotation; requires a human-coded gold sample before large-scale deployment; treats prompt design as part of the annotation manual; and evaluates accuracy, precision, recall, F1, inter-annotator agreement, stability across repeated runs, subgroup performance, and error types. Recent studies show that LLMs can approach or exceed crowd-worker performance on some well-specified classification tasks, but that performance varies substantially across datasets, languages, models, prompts, text lengths, and annotation complexity. Span-level annotation and context-dependent semantic or pragmatic coding remain particularly challenging. The article therefore argues against unvalidated full automation and proposes selective automation, disagreement-based human adjudication, model/version documentation, and preservation of raw outputs. For corpus linguistics, the strongest near-term use of LLMs is as flexible annotators within a transparent, theory-driven, and auditable pipeline rather than as replacements for linguistic expertise. The resulting framework supports scalable corpus annotation while preserving the empirical principles on which corpus-based linguistic inference depends.

Keywords: Corpus linguistics; large language models; automated annotation; linguistic analysis; human-in-the-loop; reliability; reproducibility; natural language processing.

Introduction: Corpus linguistics studies recurrent language patterns in principled collections of naturally occurring texts. Its empirical value depends not only on corpus size but also on transparent sampling, metadata, replicable procedures, and reliable linguistic annotation (Biber, Conrad, & Reppen, 1998; McEnery & Hardie, 2012). Annotation connects textual forms to analytic categories ranging from lemmas and part-of-speech tags to stance, discourse functions, speech acts, metaphor, evaluation, and ideological framing. The

validity of subsequent frequency, variation, register, or discourse analyses therefore depends on the quality of those labels.

Traditional corpus annotation relies on human coders, rule-based tools, or supervised machine learning. Human annotation can capture theoretically rich distinctions but is costly; rules are transparent but brittle; supervised models can be accurate but require labeled training data. Standardized resources such as Universal Dependencies show the importance of

explicit guidelines for morphological and syntactic annotation (Nivre et al., 2020). Yet many semantic and pragmatic categories remain expensive to annotate at scale.

Large language models (LLMs) create a new option. Transformer-based pretraining and instruction-following allow models to perform many tasks from natural-language definitions and a small number of examples (Vaswani et al., 2017; Brown et al., 2020). A corpus researcher can therefore ask an LLM to label stance, rhetorical moves, semantic roles, evaluative polarity, or discourse functions without building a new classifier for every variable. LLMs can also return structured outputs and explanatory rationales.

The scientific problem is that plausibility is not validity. LLM outputs can vary across prompts, repeated runs, model versions, languages, domains, and text lengths. Gilardi, Alizadeh, and Kubli (2023) reported strong ChatGPT performance on selected annotation tasks, and Heseltine and von Hohenberg (2024) found high GPT-4 accuracy for several political-text variables. However, Bhat and Varma (2023) observed poor performance on two of three low-resource language tasks, Reiss (2023) documented prompt and run sensitivity, and Kristensen-McLachlan et al. (2025) found supervised DistilBERT more consistently effective than several LLM configurations. Span-level annotation also remains difficult (Kasner et al., 2026).

These limitations matter especially in corpus linguistics because systematic annotation errors can become apparent linguistic differences. If one register, period, dialect, or language is labeled less accurately than another, downstream comparisons may partly measure model bias. This article therefore asks: (1) which corpus annotation tasks are most suitable for LLM assistance, (2) how should model labels be validated before linguistic inference, and (3) how can LLMs be integrated without sacrificing transparency and reproducibility?

The central argument is that an LLM should be treated as a probabilistic annotator within a measurement process, not as an unquestioned linguistic authority. The article develops a validated human–LLM workflow combining corpus design, expert gold annotation, prompt specification, repeated model coding, quantitative validation, error analysis, selective human

adjudication, and provenance tracking. The objective is not maximum automation, but scalable annotation that remains theoretically defensible and empirically auditable.

2. Corpus Annotation as Linguistic Measurement

A corpus is a designed sample of language, and its annotations are measurements layered onto that sample. Sinclair (2005) emphasizes representativeness, sampling, and balance, while Leech (2005) describes annotation as the addition of interpretive linguistic information. Some labels are close to observable form, such as sentence boundaries or lemmas; others, such as sarcasm, stance, rhetorical purpose, or politeness, require substantial interpretation. LLM suitability therefore depends on both annotation unit and conceptual difficulty.

Five levels are particularly useful. Token-level annotation assigns labels to words; span-level annotation identifies textual stretches such as entities, metaphors, or multiword expressions; clause- or sentence-level annotation captures functions such as stance or speech acts; document-level annotation assigns genre, topic, or overall orientation; and discourse-level annotation represents relations across units. The higher levels may provide more context, but they also introduce more abstraction and ambiguity.

Annotation must also be separated from analysis. Annotation produces structured labels; corpus analysis uses those labels to calculate distributions, concordances, collocations, register differences, or statistical relationships. Deterministic operations such as token counts, normalization, and keyness should remain reproducible through scripts or corpus tools. LLMs are most valuable where contextual interpretation is needed, provided the resulting labels have been independently validated.

Established annotation projects offer an important lesson: a tagset is part of a theory of language. Universal Dependencies publishes explicit, cross-linguistically motivated guidelines rather than relying on intuitive labels (Nivre et al., 2020). The same principle should govern LLM prompts. “Identify stance” is not a scientific annotation scheme. A usable prompt must specify the unit, construct definition, permissible labels, inclusion and exclusion rules, examples, ambiguity procedures, and output format.

Prompt engineering is therefore better understood as annotation-manual engineering. Changing definitions or examples can change the operationalized variable, so prompt versions should be archived and reported. Human “gold” labels also require scrutiny: for interpretive categories, at least two trained annotators and an adjudication procedure are preferable. Where disagreement reflects genuine ambiguity, preserving that information can be more informative than pretending that every instance has an obvious true label.

3. What LLMs Change in the Annotation Process

LLMs shift annotation work from task-specific programming toward instruction design and validation. A researcher can describe a taxonomy in natural language and supply few-shot demonstrations instead of training a separate model from a large labeled dataset. This flexibility is attractive for changing or specialized coding schemes and for semantic or pragmatic phenomena that depend on context.

Three capabilities are especially relevant. First, LLMs can integrate lexical, syntactic, and discourse cues when assigning a category. Second, they can return several related labels from the same passage, such as stance type, target, and polarity. Third, they can provide a short rationale that helps reviewers inspect difficult cases. Rationales should be treated as auditable explanations rather than faithful windows into internal model reasoning.

The same flexibility produces risks. Natural-language instructions can be interpreted differently from the researcher’s theoretical definition; few-shot examples can create unintended shortcuts; output formats can drift; and identical inputs can produce different labels. Closed services may also update models without allowing researchers to reconstruct the original system. Current evidence is therefore mixed but coherent. Gilardi et al. (2023) showed that ChatGPT can outperform crowd-workers on selected tasks. Ziems et al. (2024), across 25 computational social-science benchmarks, found that LLMs did not consistently beat strong fine-tuned classifiers on taxonomic labeling but could contribute usefully to human annotation teams. Bhat and Varma (2023) reported substantial variation

across low-resource languages and tasks. Volkanovska (2025) found few-shot LLM named-entity recognition insufficiently reliable as a general solution for specialized environmental text.

Text length and task structure also matter. Heseltine and von Hohenberg (2024) found particularly strong performance on short political texts and lower performance on longer news articles. Reiss (2023) showed that small prompt changes and repeated runs can alter labels. Pangakis, Wolken, and Fasching (2023) consequently recommend task-by-task validation against human annotation. Kristensen-McLachlan et al. (2025) further showed that a supervised DistilBERT classifier generally outperformed the tested LLM configurations on their binary tasks.

Span annotation deserves special caution because a model must identify both category and boundaries. Kasner et al. (2026) report only moderate LLM–human agreement across several span tasks, despite potentially attractive cost. Corpus studies of metaphor, error spans, multiword units, or discourse markers should therefore report strict exact-span metrics as well as token-overlap measures.

The conclusion is task-sensitive: LLMs are neither universally superior nor inherently unsuitable. They are most compelling when annotation categories are contextual and flexible, training data are limited, and validation shows a meaningful advantage over simpler alternatives. If an established tagger or compact supervised model is more stable, cheaper, and equally accurate, using a generative LLM offers little methodological benefit.

4. A Validated Human–LLM Framework for Corpus Annotation

Figure 1 presents the proposed workflow. The process begins with corpus design and a documented annotation scheme. A stratified subset is then coded by human experts to test the guidelines and create an independent benchmark. The gold sample should represent the genres, periods, proficiency levels, dialects, languages, or other strata on which later comparisons will be made. Rare but important categories may require targeted oversampling.

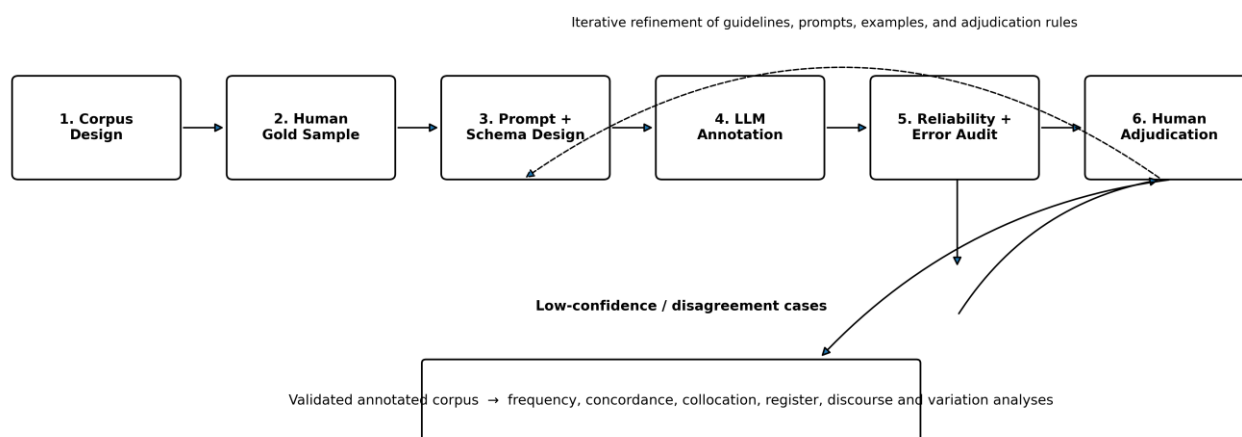


Figure 1. Proposed validated human-LLM workflow for automated corpus annotation and downstream linguistic analysis.

For interpretive coding, at least two human annotators are desirable. Pre-adjudication agreement should be reported using an appropriate coefficient such as Cohen's kappa for two nominal coders (Cohen, 1960) or Krippendorff's alpha for more flexible designs. Agreement values should be accompanied by qualitative discussion because prevalence and conceptual ambiguity can affect the statistics.

The coding manual is then translated into a structured LLM prompt. The prompt should state the annotation role, unit of analysis, theoretical definition, allowed labels, inclusion and exclusion criteria, positive and negative examples, uncertainty rules, and a machine-readable output schema. Few-shot examples should come from a development set rather than the held-out gold evaluation sample. Contrastive examples are especially useful for borderline categories.

Before full-corpus annotation, the model is evaluated on the gold set. Classification tasks should report accuracy, macro-F1, and per-class precision, recall, and F1. Span tasks should report exact-boundary and overlap metrics. Ordinal or continuous ratings require agreement measures appropriate to their scales. Results should be broken down by corpus strata, not only averaged globally.

Repeated annotation of the same validation items is essential. Self-agreement across identical runs, prompt variants, or multiple models provides an uncertainty signal. Items with unstable labels can be routed to humans. This logic parallels active annotation approaches such as LLMaAA, which uses LLM labels selectively rather than assuming every pseudo-label is

equally trustworthy (Zhang et al., 2023).

A practical decision rule has three zones. In the green zone, pre-specified metrics are met across important strata and no serious systematic error is found; routine cases can be automated with periodic audits. In the amber zone, performance is adequate overall but weak for particular categories or varieties; only high-confidence cases should be accepted automatically. In the red zone, labels are unstable, subgroup gaps are large, errors threaten the construct, or a simpler baseline performs better; the LLM should not be the primary annotator.

Calderon, Reichart, and Dror (2025) make the same principle explicit by proposing a statistical test for whether an LLM can replace human annotators. Whether or not that exact procedure is adopted, replacement should be demonstrated empirically. Summary metrics should also be supplemented by error analysis. False positives and negatives can be grouped into lexical shortcuts, missing context, negation, boundary errors, figurative language, terminology, multilingual interference, or unclear guidelines. Some failures can be corrected by prompt refinement; others indicate that the annotation construct itself needs revision.

Human adjudication then focuses on disagreements and uncertainty rather than every item. Heseltine and von Hohenberg (2024) show that reconciling disagreement across repeated LLM codings can improve accuracy. A corpus project can extend this strategy: accept stable cases, review uncertain cases, and manually audit a random sample of accepted

labels. Each output should preserve a text identifier, automatic label, model/version, prompt version, and adjudication status.

After validation, downstream linguistic analysis should use standard reproducible procedures. Normalized frequencies, dispersion, concordance inspection, collocation, multivariate analysis, and statistical tests should be computed from archived annotation files and code. LLM-generated interpretations may assist discussion, but core quantitative findings should not depend on an unrecoverable conversational interaction.

Validation sample size should be determined by the intended claim rather than by a fixed percentage of the corpus. A small pilot may be enough to reject a poor prompt, but a model intended to replace substantial human coding requires narrower confidence intervals around its error estimates. Researchers should report the number of gold instances per class and per analytically important stratum. Rare labels deserve explicit attention because a high micro-average can be achieved while missing most instances of a low-frequency phenomenon. Where feasible, bootstrap confidence intervals for F1 or accuracy can show how uncertain the estimated performance is.

Prompt development should also be separated from final testing. A development set can be used to compare definitions, label descriptions, few-shot examples, and output formats. Once the prompt is frozen, performance should be reported on a held-out test set that was not used to tune the instructions. This mirrors standard machine-learning practice and reduces the temptation to keep modifying the prompt until it happens to fit the evaluation sample. The final article should identify which prompt decisions were made before versus after seeing errors.

For large corpora, quality control can continue during production. One strategy is to divide the corpus into batches and manually audit a small random sample from each batch. Control items with known labels can be inserted periodically to detect drift. If accuracy declines for a later batch, annotation can be paused and recalibrated instead of discovering the problem after all texts have been processed. This is especially important when external providers update models or when the corpus changes substantially across time

periods or registers.

Finally, researchers should conduct a downstream sensitivity analysis. The practical importance of an annotation error depends on the analysis that follows. A five-percent random error may have little effect on a very strong frequency difference, whereas a smaller but systematic error concentrated in one group can reverse a comparison. Recomputing the principal linguistic results under alternative plausible labels for uncertain items can show whether the substantive conclusion is robust to annotation uncertainty.

5. Integrating LLM Annotation with Corpus-Linguistic Analysis

The strongest integration is modular: deterministic operations remain deterministic, while validated LLMs handle categories that genuinely require context. Tokenization, frequency counts, normalization, keyword statistics, collocational measures, and regressions usually do not benefit from generative uncertainty. Semantic, pragmatic, and rhetorical coding often does.

In academic-writing corpora, conventional tools can identify lemmas, grammatical features, lexical bundles, and syntactic patterns. An LLM can then classify the communicative function of selected forms—for example, whether a modal expresses epistemic probability, obligation, or ability, or whether a sentence functions as a claim, limitation, recommendation, or methodological justification. This enriches corpus annotation without replacing established quantitative methods.

Media corpora offer a similar division. Statistical keyword analysis can locate salient vocabulary and concordance lines; an LLM can annotate the stance target, frame, or evaluative function of those lines after validation. The model should not decide statistical keyness by intuition. Learner corpora may use LLMs to classify error type or revision function, but validation must guard against prescriptive bias toward standardized varieties.

Register studies require particular care. Biber-style multidimensional analysis derives patterns from co-occurring feature frequencies. If an LLM has different recall across registers, a factor may partly reflect annotation bias. Gold validation should therefore estimate category-specific performance separately for

each register before feature counts are entered into multivariate analysis.

Historical and multilingual corpora also require specialized validation. Contemporary pretraining may impose present-day meanings on historical language, while multilingual training distributions are uneven. An annotation scheme validated in English should not automatically be transferred to Urdu, Arabic, Chinese, code-switching, or dialectal data. Translation into English may improve classification but can erase linguistic evidence and should only be used when the construct is explicitly language-independent.

LLMs are also useful during exploration. They can suggest candidate categories from sampled concordances or cluster similar passages, after which the researcher formalizes the scheme and tests it systematically. The same model should not be allowed to invent a category, label the corpus, and independently “confirm” the resulting pattern without external validation; separating exploration, annotation, and inference reduces circularity.

Table 3 summarizes recommended metrics. The key principle is that evaluation should match the annotation unit. Document labels need class-sensitive metrics, token labels need token-level performance, spans need boundary-aware measures, and continuous judgments need error and agreement statistics rather than correlation alone.

A useful implementation pattern is to preserve the original corpus and store LLM labels in a separate

annotation layer. The annotation table can contain one row per textual unit with fields for document ID, start and end offsets, label, model, prompt version, run number, and adjudication status. This structure prevents automatic labels from being confused with original text and allows researchers to compare successive annotation versions without altering the corpus source.

The distinction between corpus-based and corpus-driven research also remains relevant. In a corpus-based study, an existing linguistic framework can be encoded in the prompt and evaluated against human examples. In a more corpus-driven study, an LLM may help summarize recurrent patterns in concordance samples and propose candidate distinctions. Those candidate distinctions should then be converted into explicit categories and applied systematically. The model’s exploratory suggestions are not themselves evidence of frequency or representativeness.

LLMs may additionally support qualitative follow-up after quantitative patterns have been established. For example, after a significant difference in stance categories is identified between two registers, the researcher can sample concordance lines from each category and ask a model to group recurrent contextual realizations. Those groupings should be checked by the analyst and linked back to source examples. Used this way, the LLM accelerates close reading without replacing the traceability that is central to corpus interpretation.

Table 1. Recommended validation metrics for common corpus annotation tasks

Task type	Core metrics	Additional checks	Why it matters
Binary / multiclass classification	Accuracy; macro-F1; per-class precision/recall/F1	Confusion matrix; prevalence; subgroup metrics	Overall accuracy can hide minority-class failure.
Token sequence labeling	Token accuracy; macro-F1; per-label F1	Out-of-vocabulary and non-standard subsets	Corpus feature counts depend on individual-token decisions.
Span annotation	Exact-span P/R/F1; token-overlap P/R/F1	Boundary error distribution; false-span audit	Correct category with wrong boundaries can distort frequency and context.

Task type	Core metrics	Additional checks	Why it matters
Ordinal ratings	Weighted kappa; MAE	Distribution shift; calibration	Order and distance between labels are analytically meaningful.
Continuous ratings	MAE/RMSE; ICC; correlation	Bland–Altman-style level bias	High correlation can coexist with systematic score inflation.
Repeated LLM runs	Self-agreement; label entropy / disagreement rate	Prompt-order and model-version sensitivity	Measures annotation stability rather than only agreement with humans.

6. Reproducibility, Ethics, and Open-Science Requirements

Reproducibility is a central challenge for LLM-assisted corpus annotation. Closed services can change models while retaining similar product names, and stochastic decoding can alter outputs. Kristensen-McLachlan et al. (2025) therefore emphasize transparency, cost, privacy, and Open Science concerns alongside accuracy.

A methods section should report more than “we used ChatGPT.” Researchers should record provider or local checkpoint, exact model identifier when available, access date, system and user prompts, few-shot examples, decoding parameters, output schema, preprocessing, batching, retries, and post-processing. Prompt templates should be released in supplementary material, and raw outputs should be retained when licensing permits. If the model changes during a project, the researcher should test a fixed calibration set to determine whether batch effects were introduced.

Open-weight models offer more local control but still require documentation of checkpoint, inference library, chat template, quantization, context length, and decoding settings. They can be preferable for confidential corpora because data need not leave the research environment.

Privacy must be considered before sending interviews, student writing, emails, medical narratives, or identifiable social-media text to external services. Ethics approval, consent, corpus licensing, and data-

processing rules may restrict third-party transfer. De-identification or local deployment may be necessary.

Bias is equally important. LLMs can encode social stereotypes and standard-language preferences. Annotation of politeness, toxicity, professionalism, grammaticality, or sentiment may therefore perform differently across groups and varieties. Wang, Morgenstern, and Dickerson (2025) demonstrate broader risks of identity flattening when models substitute for human populations. Corpus projects should perform subgroup error analysis wherever social or linguistic varieties are analytically relevant.

Figure 2 summarizes a three-layer audit. Measurement validity asks whether the coding scheme operationalizes the intended construct. Annotation reliability asks whether LLM labels agree sufficiently with adjudicated human labels and remain stable across runs and subgroups. Reproducibility asks whether another researcher can reconstruct the process from saved materials. A project should not scale annotation until all three layers are satisfactory.

Table 4 converts these principles into a reporting checklist. The goal is not administrative detail for its own sake: provenance is what allows future researchers to distinguish linguistic evidence from model-specific decisions.

Cost and environmental reporting can also improve methodological comparison. API expenditure, local inference time, hardware requirements, and human adjudication hours should be reported at least approximately. A system that is marginally more

accurate but requires much greater cost may not be preferable for routine annotation. Conversely, a slightly more expensive model may be justified if it sharply reduces expert review for a rare, difficult phenomenon. Reporting these trade-offs allows future researchers to choose methods on evidence rather than model prestige.

Corpus licensing deserves separate attention from

privacy. A text may be publicly visible but still be distributed under terms that restrict automated processing, redistribution, or transfer to third parties. Researchers should verify the corpus license and the model provider's data-handling terms. When source text cannot be shared, reproducibility can still be improved by releasing document identifiers, prompts, code, aggregate results, and derived labels where legally permitted.

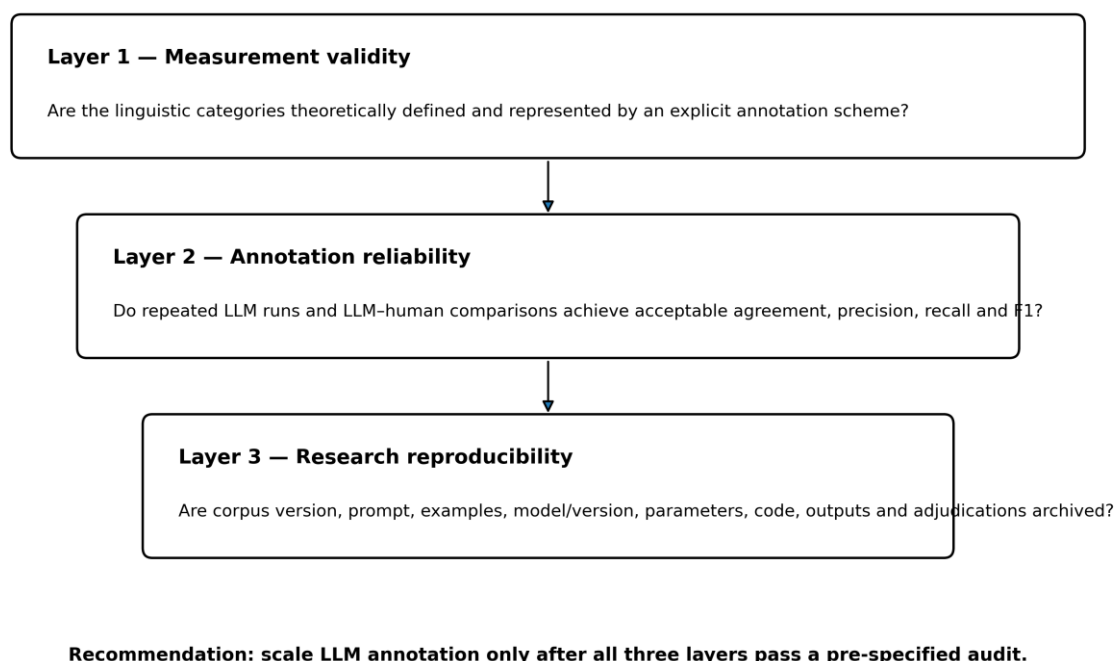


Figure 2. Three-layer audit for deciding whether LLM-generated annotations are suitable for scientific corpus analysis.

Ethical review should also consider whose linguistic norms are embedded in the coding scheme. An LLM may appear consistent because it reproduces dominant conventions, while the research question concerns

legitimate variation. For categories such as error, civility, professionalism, or appropriateness, the paper should state the normative reference explicitly and avoid presenting culturally or socially contingent judgments as universal linguistic facts.

Table 2. Minimum reporting checklist for reproducible LLM-assisted corpus annotation

Component	Minimum information to report
Corpus	Corpus name/version, sampling frame, inclusion criteria, languages/registers, preprocessing, metadata variables.
Human benchmark	Gold-sample selection, annotator expertise, coding manual, pre-adjudication agreement, adjudication procedure.

Component	Minimum information to report
LLM	Provider or checkpoint, exact model/version when available, access date, local/API/interface mode.
Prompt	Full system/user prompt, definitions, labels, few-shot examples, example order, uncertainty instruction, output schema.
Inference	Temperature/decoding settings, context window, batch size, retries, response parsing and validation rules.
Evaluation	Accuracy/F1 or task-appropriate metrics, repeated-run stability, subgroup performance, baseline comparisons, error analysis.
Provenance	Raw model output, prompt version, automatic vs adjudicated label, reviewer changes, timestamps/version identifiers.
Release	Analysis code, prompt templates, derived annotation files where licensing permits, limitations and known failure modes.

7. Synthesis of Current Evidence

Five conclusions emerge from the literature. First, LLMs can be highly effective for bounded, well-defined classification tasks, sometimes reaching or exceeding crowd-worker performance. Their speed and flexible instruction following make them valuable for large projects and specialized schemes.

Second, performance is conditional rather than universal. Differences across Gilardi et al. (2023), Bhat and Varma (2023), Heseltine and von Hohenberg (2024), Kristensen-McLachlan et al. (2025), Volkanovska (2025), and Kasner et al. (2026) are best understood as evidence that model success depends on corpus, task, language, prompt, and annotation unit. “Does LLM annotation work?” is therefore the wrong question; the relevant question is whether it works for a defined construct and target corpus at an acceptable error rate.

Third, supervised baselines remain essential. If a small gold dataset exists, an encoder classifier can be more stable and economical than repeated generative inference. LLMs provide greatest value when categories are context-dependent, taxonomies change,

training data are limited, or structured explanations help review.

Fourth, human labor should be reallocated rather than eliminated. Experts can focus on scheme design, gold annotation, rare cases, disagreements, and error analysis, while the model handles repetitive high-confidence items. This is a more defensible form of automation than indiscriminate full-corpus labeling.

Fifth, corpus-linguistic methodology itself strengthens LLM evaluation. Stratification by register, time, speaker group, or language can reveal systematic errors hidden by an overall F1 score. Concordance-based inspection helps researchers understand why a label fails in context. Table 2 illustrates the range of outcomes in selected studies; it is a methodological synthesis rather than a meta-analysis.

The likely future is therefore hybrid. Deterministic preprocessing, specialized taggers, compact supervised classifiers, LLMs for context-sensitive cases, and expert adjudication can coexist in one pipeline. Automation becomes a calibrated allocation of cases according to evidence of reliability rather than a simple human-versus-machine choice.

Table 3. Selected evidence on LLM-assisted text annotation relevant to corpus research

Study	Task / setting	Main finding for corpus methodology
Gilardi et al. (2023)	Multiple text-annotation tasks; ChatGPT vs crowd-workers	Demonstrated that LLM annotation can match or exceed crowd-worker quality on selected, well-defined tasks.
Bhat & Varma (2023)	NLI, sentiment and news classification in low-resource languages	Performance varied strongly across tasks/languages; off-the-shelf multilingual LLM use was unsuitable for two of three tasks.
Zhang et al. (2023)	NER and relation extraction with active learning	LLM annotations can be useful within selective/active annotation loops rather than as indiscriminate full-corpus labels.
Pangakis et al. (2023)	27 annotation tasks across 11 social-science datasets	Performance was promising but dataset- and task-dependent; human validation is required.
Heseltine & von Hohenberg (2024)	Political text in four countries	High accuracy on several variables; shorter texts easier; repeated LLM coding plus expert reconciliation improved quality.
Ziems et al. (2024)	13 LMs, 25 computational social-science benchmarks	LLMs did not consistently beat best fine-tuned classifiers for taxonomic labels but were useful as human-team annotators and for explanations.
Kristensen-McLachlan et al. (2025)	Open/closed LLMs vs supervised classifiers	Substantial task/prompt variability; DistilBERT was generally more consistent; highlights Open Science and reproducibility concerns.
Volkanovska (2025)	NER in specialized environmental scientific text	Few-shot LLM NER was not a general solution; prompt, class and data format affected results.
Calderon et al. (2025)	Ten datasets, multiple LLMs/prompts	Introduced a statistical test for deciding whether human annotators can be replaced, emphasizing formal justification.
Kasner et al. (2026)	Three span-annotation tasks	LLM-human agreement was only moderate overall; span annotation needs boundary-sensitive validation.

8. Implications for Future Corpus-Linguistic Research

Future research should build benchmarks tailored to

corpus linguistics rather than relying mainly on generic NLP or social-science classification datasets. Gold resources are needed for discourse moves, stance

functions, semantic preference, pragmatic markers, metaphor, phraseological functions, learner errors, register features, and historical meanings across multiple genres and languages.

Uncertainty calibration is another priority. Verbal confidence from a generative model should not be trusted automatically. More useful signals include disagreement across repeated runs, prompt variants, or independent models. Research should test how well these signals predict actual error and whether they can support reliable human-routing policies.

Cross-corpus transfer also deserves systematic study. A prompt validated on research articles may fail on classroom talk or journalism even with the same label set. Leave-one-register-out and temporal-holdout evaluations can show whether a model captures the intended construct or relies on superficial cues.

Multilingual work must test conceptual as well as computational equivalence. Categories such as politeness, stance, evidentiality, or offensiveness can differ across linguistic communities. Language experts should participate in gold annotation, and performance should be reported separately by language and variety.

Human–LLM disagreement can itself become a research object. Recurrent disagreement clusters may reveal unclear boundaries in a coding scheme, overlooked subcategories, or language-specific realizations. LLMs can therefore function as probes that help refine linguistic theory, not merely as cheaper coders.

Finally, corpus releases should include an AI annotation card documenting corpus scope, label definitions, gold annotation, model and prompt, evaluation metrics, subgroup performance, provenance, and known failure modes. Such documentation would make AI-generated labels inspectable instead of allowing them to become invisible assumptions inside a corpus.

Another important direction is model comparison at equal annotation budgets. Many studies compare a costly LLM with a supervised classifier that has already benefited from hundreds or thousands of human labels. Future corpus research should ask what quality can be achieved for a fixed amount of money or expert time. The comparison might include direct expert annotation, crowd annotation with adjudication, few-shot LLM coding, active learning, and a small

supervised classifier trained on the same gold sample. Such budget-normalized evaluation would make claims of efficiency substantially more informative.

Research is also needed on annotation ensembles. Independent models may make different errors, and agreement among models can provide a stronger acceptance signal than confidence from a single model. However, ensembles can reproduce shared biases if the models have similar training data or architectures. Studies should compare same-family and cross-family ensembles and determine whether disagreement genuinely predicts linguistic ambiguity or merely technical instability.

A final opportunity concerns longitudinal reproducibility. Because LLM technology changes quickly, benchmark corpora with frozen gold labels can be re-annotated annually with newer models. This would show whether progress is occurring uniformly across linguistic phenomena or whether certain categories remain resistant to automation. Corpus linguistics, with its emphasis on stable textual datasets and replicable measurement, is particularly well suited to provide such long-term evaluation resources.

9. Conclusion

Large language models provide corpus linguistics with a flexible interface for annotation beyond conventional lexical and morphosyntactic tagging. They can apply natural-language definitions, adapt to specialized categories with few examples, incorporate context, and generate structured labels at scale. These capabilities are particularly promising for semantic, pragmatic, rhetorical, and discourse-level research.

Evidence through 2026 nevertheless shows substantial variation across tasks, prompts, languages, text lengths, and annotation units. Strong results coexist with instability, weaker performance in specialized or low-resource settings, span-boundary errors, subgroup risks, and reproducibility problems. Conventional supervised classifiers also remain competitive.

The appropriate methodological response is validated selective automation. A stratified human-coded gold sample should establish the benchmark; the prompt should function as a formal coding manual; performance should be measured with task-appropriate metrics, repeated runs, subgroup analysis, baseline comparisons, and error inspection; uncertain

cases should be routed to experts; and model versions, prompts, raw outputs, and adjudications should be archived.

The objective is not to replace linguistic expertise. It is to create scalable measurements that remain valid, reliable, transparent, and reproducible. Under those conditions, LLMs can become durable collaborators in corpus construction and analysis rather than opaque shortcuts.

The proposed framework also has a deliberate limitation: it does not prescribe one universal model, prompt, or numerical acceptance threshold. Corpus questions differ too widely for a single cutoff to be scientifically defensible. The appropriate standard must reflect the annotation unit, category prevalence, consequences of error, available human expertise, and the magnitude of downstream linguistic claims. What should be universal is the requirement to state that standard in advance and demonstrate that the chosen annotation procedure meets it.

References

1. Bhat, S., & Varma, V. (2023). Large language models as annotators: A preliminary evaluation for annotating low-resource language content. In *Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems* (pp. 100–107). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.eval4nlp-1.8>
2. Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press.
3. Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.
4. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.
5. Calderon, N., Reichart, R., & Dror, R. (2025). The Alternative Annotator Test for LLM-as-a-Judge: How to statistically justify replacing human annotators with LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 16051–16081). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.acl-long.782>
6. Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46. <https://doi.org/10.1177/001316446002000104>
7. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019* (pp. 4171–4186). Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-1423>
8. Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd-workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30), e2305016120. <https://doi.org/10.1073/pnas.2305016120>
9. Gries, S. T. (2009). *Quantitative corpus linguistics with R: A practical introduction*. Routledge.
10. Heseltine, M., & von Hohenberg, B. C. (2024). Large language models as a substitute for human experts in annotating political text. *Research & Politics*, 11(1), 1–10. <https://doi.org/10.1177/20531680241236239>
11. Kasner, Z., Zouhar, V., Schmidtová, P., Kartáč, I., Onderková, K., Platek, O., Gkatzia, D., Mahamood, S., Dusek, O., & Balloccu, S. (2026). LLMs as span annotators: A comparative study of LLMs and humans. In *Proceedings of the First Workshop on Multilingual Multicultural Evaluation* (pp. 1–22). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2026.mme-main.1>
12. Kristensen-McLachlan, R. D., Canavan, M., Kárdos, M., Jacobsen, M., & Aarøe, L. (2025). Are chatbots reliable text annotators? Sometimes. *PNAS Nexus*, 4(4), pgaf069. <https://doi.org/10.1093/pnasnexus/pgaf069>
13. Krippendorff, K. (2018). *Content analysis: An introduction to its methodology* (4th ed.). SAGE.
14. Leech, G. (2005). Adding linguistic annotation. In M. Wynne (Ed.), *Developing linguistic corpora: A guide to good practice* (pp. 17–29). Oxbow Books.
15. McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511981395>

16. Nasuto, A., Iacus, S., Rowe, F., et al. (2026). Large language models identify immigration attitudes in online discourse regardless of language. *Scientific Reports*, 16, 20118. <https://doi.org/10.18653/v1/2023.findings-emnlp.872>
17. Nivre, J., de Marneffe, M.-C., Ginter, F., Hajič, J., Manning, C. D., Pyysalo, S., Schuster, S., Tyers, F., & Zeman, D. (2020). Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 4034–4043). European Language Resources Association.
18. Pangakis, N., Wolken, S., & Fasching, N. (2023). Automated annotation with generative AI requires validation. *arXiv*. <https://doi.org/10.48550/arXiv.2306.00176>
19. Reiss, M. V. (2023). Testing the reliability of ChatGPT for text annotation and classification: A cautionary remark. *arXiv*. <https://doi.org/10.48550/arXiv.2304.11085>
20. Sinclair, J. (2005). *Corpus and text: Basic principles*. In M. Wynne (Ed.), *Developing linguistic corpora: A guide to good practice* (pp. 1–16). Oxbow Books.
21. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
22. Volkanovska, E. (2025). Large language models as annotators of named entities in climate change and biodiversity: A preliminary study. In *Proceedings of the 1st Workshop on Ecology, Environment, and Natural Language Processing* (pp. 24–33). University of Tartu Library.
23. Wang, A., Morgenstern, J., & Dickerson, J. P. (2025). Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, 7, 400–411.
24. Wynne, M. (Ed.). (2005). *Developing linguistic corpora: A guide to good practice*. Oxbow Books.
25. Zhang, R., Li, Y., Ma, Y., Zhou, M., & Zou, L. (2023). LLMaAA: Making large language models as active annotators. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 13088–13103). Association for Computational Linguistics.
26. Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2024). Can large language models transform computational social science? *Computational Linguistics*, 50(1), 237–291. https://doi.org/10.1162/coli_a_00502