STED: MEASURING CONSISTENCY IN LLM STRUCTURED OUTPUTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Structured responses are common in industrial applications of large language models (LLMs), with tool-calling agents a prominent use case. Yet even at temperature $T = 0$, repeated prompts can produce inconsistent tool choices and argument values despite parseable responses. We propose **STED** (Semantic Tree Edit Distance), a type-aware score for comparing hierarchical structured outputs. It matches terminal values at corresponding tree paths and aggregates their graded agreement using soft precision and recall. On 208 structural human comparisons used during development, STED makes 88.0% correct decisions versus 79.8% for a field-path similarity baseline, counting metric ties as unsuccessful decisions. Frozen numerical-equivalence tests isolate a benefit of interpreting numerical literals, with within-value-group AUCs of .999–1.000 versus .612–.619 for a control with matched text processing. Textual relatedness and external model judgments show no consistent improvement. We apply the fixed score to 2.1M repeated generations from 18 LLMs, measuring consistency separately from validity and prompt coverage. The observed rankings and model-dependent schema-enforcement effects show why repeatability adds information to model selection alongside task quality.

1 INTRODUCTION

Structured responses are a common interface for industrial LLM adoption. Applications include content moderation, which returns categories and policy decisions, and information extraction from bank forms, which maps account details and monetary amounts to predefined fields. Tool-calling agents extend this interface to action: models emit function names and structured arguments to automate business workflows. We study hierarchical structured outputs, focusing on JSON and JSON-encoded tool calls. Reliability requires consistent labels, extracted values, and tool-call arguments across repeated invocations, alongside parseable responses.

Repeatability cannot be assumed even at temperature $T = 0$. Atil et al. (2024) document output and accuracy variation under settings expected to be deterministic. Finite-precision arithmetic and batch-dependent computation can alter greedy token choices (Yuan et al., 2025; He & Thinking Machines Lab, 2025); remedies requiring backend access are unavailable to closed-model API callers. Single-output evaluations also miss variation across generations (Song et al., 2024). These findings motivate repeated-run evaluation under a fixed deployment.

Acceptable variation depends on the application: reordering a collection can be harmless, but rotating a numerical vector can change a function's return; two spellings can be accepted aliases, while a one-character date change can be consequential. String equality, structural alignment, and semantic similarity capture different aspects of these relations. Evaluating a consistency metric requires testing its rankings, invariances, and acceptance decisions against explicit targets.

Our validation labels concern reference-relative closeness or argument equivalence. We then aggregate the same comparison score across repeated generations; that application corpus has no independent pair-equivalence labels. The distinction separates validation of comparison rules from measurement of repeated-output consistency.

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

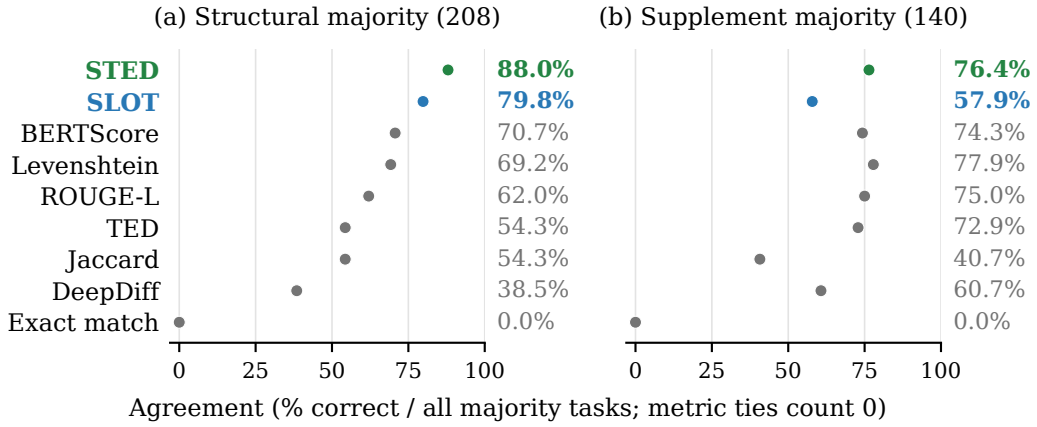


Figure 1: **Human preference agreement on both development sets.** Correct decisions use all binary majorities: 208 structural and 140 supplement tasks. Metric ties count as unsuccessful decisions, matching the ablations. Three annotators' labels informed configuration selection. Conditional rates and exact counts appear in App. Table 19.

Our contributions are a consistency measure and its validation and application:

- **Type-aware comparison** (Sections 3 and 5): STED retains SLOT's exact field-path correspondence and soft-F1 normalization and introduces graded, type-dependent terminal comparisons. One fixed configuration is evaluated throughout.
- **Validation and application** (Section 4): human comparisons, accepted alternatives and external tests characterize the rules' benefits and limits. We assemble 2.1M repeated generations and apply the fixed score to study consistency as a model-selection criterion.

2 Related Work

Nondeterminism in LLM inference. Output variation at $T = 0$ (Atil et al., 2024) and numerical analyses (Yuan et al., 2025) show that greedy decoding alone is insufficient for repeatability. He & Thinking Machines Lab (2025) obtain identical completions with batch-invariant vLLM kernels, but this requires inference-stack access unavailable to closed-model API callers. Repeatability claims must therefore specify serving conditions.

Evaluating variation across generations. Repeated sampling reveals benchmark-score variation hidden by a single run (Song et al., 2024). Atil et al. (2024) distinguish raw-response agreement from parsed-answer agreement, but exact rates do not grade partial changes. Semantic uncertainty clusters responses by meaning (Kuhn et al., 2023); SelfCheckGPT uses sampled agreement to assess factual reliability (Manakul et al., 2023). Neither specifies field identity or array order. Wu et al. (2025) find that consistency surrogates' agreement with users depends on the metric and task, motivating validation for the intended outputs.

Structured-output evaluation. Constrained generation restricts admissible outputs (Beurer-Kellner et al., 2023; Willard & Louf, 2023; Geng et al., 2025), but schemas permit different choices and values. Requesting JSON through prompting can also change diversity (Parikh, 2026), separately from decoder-level enforcement. Extraction and tool-calling benchmarks assess task performance against references or execution criteria (Singh et al., 2026; Feng et al., 2025; Zhou et al., 2025; Patil et al., 2025); these targets do not validate graded agreement between repeated outputs. Structure–Content Decomposition (Zhang et al., 2026) separates placement from value errors, highlighting the need to consider both.

Metrics for comparing structured outputs. Tree edit distance compares structures through edit operations (Tai, 1979; Zhang & Shasha, 1989). JSON-specific distances make different ordering assumptions: JEDI preserves array order (Hütter et al., 2022), whereas HMIL-compatible comparison

treats arrays as bags (Šopík & Strenáčik, 2022). BERTScore (Zhang et al., 2020) aligns contextual tokens; applied to serialized JSON, it lacks explicit field-identity constraints. SLOT's content metric anchors semantic comparisons to field paths (Wang et al., 2025a); implementations must also fix array-index handling. STED shares its precision/recall normalization, making value kernels and correspondence rules the relevant comparison. Stickler offers default and custom field comparators (AWS Labs, 2025); ExtractBench specifies metrics in schemas and uses LLM-based array alignment (Ferguson et al., 2026); Object Aligner additionally handles references across renamed identifiers (Drchal, 2026). These alternatives highlight the need to state and validate comparison policies: structural alignment and semantic proximity alone do not establish consumer equivalence.

3 METHOD

JSON encodes recursive containment: objects contain keyed members, arrays contain indexed elements, and primitive values terminate the nesting. Each nested occurrence has a unique syntactic parent, and containment is acyclic; a parsed output therefore defines a rooted tree (Hütter et al., 2022). Field labels and indices locate values within this hierarchy.

For outputs A and B, let T_A and T_B denote their rooted, labeled trees. STED retains SLOT's exact terminal-path correspondence and soft-F1 normalization (Wang et al., 2025a), changing the value-comparison rules. It assigns costs at fixed paths without searching for an optimal tree-edit alignment.

3.1 TREE REPRESENTATION AND SEMANTIC COMPARISON

The root represents the complete output. Each terminal is a primitive value or an empty container, located by a tuple of typed object-key and array-index steps. Thus a key containing a dot cannot alias a nested path, and object key `"0"` differs from array index 0. Distinct markers retain empty objects and arrays; missing fields have no terminal. Paths distinguish fields with similar values but different roles.

Our experiments target JSON and tool calls. XML and HTML also admit tree representations, but are outside the experiments reported here. Tables and graphs can be serialized as JSON, but STED compares their tree encoding without modeling relational semantics or graph equivalence.

For matched values a, b, let $k(a, b) \in [0, 1]$ denote local similarity and $c(a, b) = 1 - k(a, b)$ substitution cost. Boolean, null and empty-container matches require the same kind. Numeric kernels use normalized equality and capped near-miss credit; literal kernels use identity or normalized edit similarity. For prose, we use

$$k_{\text{text}}(a, b) = \max\left(0, \frac{e(a)^\top e(b)}{\|e(a)\| \, \|e(b)\|}\right), \tag{1}$$

where e is a configurable text encoder. If either string falls below a configurable length threshold, normalized character similarity preserves spelling distinctions. Symmetric chunk matching reduces truncation in long strings and accounts for content in both inputs. App. A specifies the routing rules and constants.

3.2 NORMALIZED TERMINAL-EDIT COST

STED assigns a normalized edit cost to terminals at fixed tree paths. We report its complementary agreement score $s_{\text{STED}} = 1 - d_{\text{STED}}$ as **STED**; higher scores indicate greater agreement.

Let P_A and P_B be the terminal-path sets of A and B. Terminals at the same path correspond; absent paths receive zero credit. Object serialization order has no effect; arrays are treated as sequences. Set-valued fields and independent parallel calls may require a different correspondence policy, whose validity must be assessed separately. Callee names use the same type-aware leaf kernel as other values, allowing graded agreement while retaining their field identity.

Exact values at shared paths receive full credit. We charge one for each unmatched terminal and $2(1 - k_p)$ for each matched pair. Let $M(A, B) = \sum_{p \in P_A \cap P_B} k_p(a_p, b_p)$. The normalized distance

and agreement score are

$$d_{\mathrm{STED}}(A,B) = \frac{|P_A \triangle P_B| + 2\sum_{p \in P_A \cap P_B}(1 - k_p(a_p, b_p))}{|P_A| + |P_B|}, \tag{2}$$

$$s_{\mathrm{STED}}(A,B) = 1 - d_{\mathrm{STED}}(A,B) = \frac{2M(A,B)}{|P_A| + |P_B|}. \tag{3}$$

Here $\triangle$ counts missing and extra paths. Equation 3 is SLOT's soft precision/recall aggregation with the specified terminal kernels: the harmonic mean of $M/|P_B|$ and $M/|P_A|$. Every parsed JSON value has at least one terminal, including an empty root container.

Since $0 \le M \le \min(|P_A|, |P_B|)$, Eq. 3 lies in $[0, 1]$; symmetric, reflexive leaf kernels give symmetry and self-agreement one. Its complementary distance is not a metric in the strict mathematical sense: soft-F1 normalization can violate the triangle inequality even with exact leaf equality, a limitation shared with SLOT (App. A). A high aggregate score need not imply agreement in every field.

Extracting and joining paths is linear in the number of terminals, apart from path construction and comparison-kernel costs.

3.3 Consistency across repeated outputs

For n runs of prompt i with $m \ge 2$ parseable outputs, define validity and pairwise consistency as

$$r_v = m/n, \qquad c_i = \frac{2}{m(m-1)} \sum_{j<k} s_{\mathrm{STED}}(A_{ij}, A_{ik}).$$

Across N prompts, let N_s have at least two parseable outputs. We report the scoreable-prompt mean $\bar{c}_s = N_s^{-1} \sum_{i:m_i \ge 2} c_i$, coverage N_s/N, and validity separately. The zero-filled mean retained for comparison with stored releases is $\bar{c}_0 = (N_s/N)\bar{c}_s$. Graded agreement measures output consistency; it does not establish task correctness or execution equivalence.

4 Experiments

We first test three aspects of the comparison policy: agreement with human preferences, recognition of equivalent numerical arguments, and acceptance of benchmark-approved argument alternatives. Controlled ablations isolate the rules behind these results. We then apply the same score to repeated generations to examine model differences and schema enforcement.

4.1 Experimental setup

Baselines and measurement policy. Baselines are TED, DeepDiff (Dehpour, 2026), SLOT, BERTScore, exact match, Jaccard (Jaccard, 1901), Levenshtein (Levenshtein, 1966), and ROUGE-L (Lin, 2004), on shared inputs where available. Primary STED uses `all-MiniLM-L6-v2` (Sentence Transformers, 2026) as e; its encoder, kernels, and correspondence are fixed across validation and application.

Metric validation criteria. Primary preference agreement is correct decisions divided by all binary-majority tasks; metric ties count zero. Numerical tests average within-reference-group AUC; pooled AUC is secondary. Accepted-alternative tests report positive and false acceptance after threshold calibration.

Statistical comparisons. Holm correction uses eight baseline, seven matched-control and six rule-ablation contrasts per human set, and eight contrasts across model judges. Each table names its family; repeated contrasts can have different adjusted p-values. Correction does not account for configuration selection.

Enforcement comparison. For each model and temperature, I contains prompts with at least two parseable outputs in both modes. We compute $\Delta c = |I|^{-1} \sum_{i \in I}(c_i^{\mathrm{enforced}} - c_i^{\mathrm{free}})$ with a 95% interval from 10,000 paired-prompt bootstrap resamples (seed 2026). Validity uses all prompts.

Table 1: **Validation targets and observed results.** Scores answer different questions and are not comparable across rows. Human results are development evidence; external model judgments are automated labels. Full results and protocols appear in the cited sections.

Target	STED and comparator results
Human preference (Figure 1)	Correct/all majorities: structural 88.0% vs. SLOT 79.8%; supplement 76.4% vs. SLOT 57.9% and Levenshtein 77.9%.
Numerical equivalence (§4.3.2)	Group-macro AUC .999–1.000 vs. matched-text SLOT .612–.619; literal routing accounts for the gain.
BFCL alternatives (§4.3.3; App. F)	At the 1% budget, acceptance is 36.1% vs. index-free SLOT 33.7%. With hard negatives at 0.1%, TP/FP is 85/3 vs. 210/4. Corpus-negative pooled AUC is .655 vs. .851.
External text (App. C.1; App. C.2)	PAWS AUC .615 vs. SLOT .631. SICK and STS-B scores equal matched-text SLOT pairwise.
CPython return equality (App. C.2)	Family-macro AUC .694 vs. matched-text SLOT .722.
External model judgments (App. D)	Source-macro agreement 80.46/78.71% vs. matched-text SLOT 81.25/79.52% for the two judges; no established difference.
Local sandbox returns (App. H)	Pooled AUC .854 vs. index-free SLOT .816; within-operator macro AUC .265 vs. .485.

4.2 Experimental data

Table 1 summarizes the validation targets and result directions. Detailed protocols and full comparator tables appear in the cited appendices.

Human comparisons. Three annotators judge 256 structural and 200 supplement tasks, selecting the closer of two candidates or a tie. Binary majorities require two matching votes. The two sets are analyzed separately.

Accepted alternatives. BFCL's answer key (Patil et al., 2025) supplies 1,885 accepted-alternative pairs, such as "New York, NY" and "NYC". Other questions supply 1,343 negatives outside the constructed alias component; the hard extension adds 1,384 character-nearest negatives. Unresolved aliases can remain.

Repeated-generation corpora. The large corpus contains 2,140,380 outputs from 18 models: 1,006 Toucan prompts (beyoru, 2025) and 75 ShareGPT-formatted prompts (Arun63, 2025b;a), with ten runs at eleven temperatures. Enforcement uses 22,500 outputs on the same ShareGPT subset at $T = 0, .5, 1$. Five malformed references are excluded from both analyses (App. J).

4.3 Results

4.3.1 Agreement with human preferences

On the structural set's 208 majority tasks, STED makes **88.0%** correct decisions (183/208), versus SLOT's 79.8% (166/208), counting metric ties as unsuccessful decisions (Figure 1). The difference is 8.2 percentage points (source-cluster 95% interval $[2.8, 13.7]$; Holm-adjusted $p = .005$, eight-baseline family).

On the supplement, selected for opposing preferences of an earlier configuration and SLOT, STED makes 107/140 correct decisions (76.4%), SLOT 81/140 (57.9%), and Levenshtein 109/140 (77.9%). STED's lead over SLOT is 18.6 points ($[8.6, 28.6]$; adjusted $p = .006$, eight-baseline family). All three annotators' judgments informed configuration selection, so these comparisons describe development data. App. Table 19 also reports conditional rates excluding metric ties.

4.3.2 Numerical argument equivalence

Equivalent amounts can have different string representations, whereas a small textual edit can change their value. We test literal routing on official decimal.js and big.js fixtures, from two JavaScript libraries for arbitrary-precision arithmetic (Mclaughlin, 2026b;a). Metrics, encoder and sampling

rules are frozen before retrieval. Library equality operations label numerical-string arguments; values previously evaluated in CPython are excluded (Table 2).

Table 2: **Numerical argument equivalence.** Official decimal-arithmetic tests, excluding previously evaluated CPython values: 1,884 pairs/618 reference groups for decimal.js; 3,000/974 for big.js. Macro AUC weights groups equally; pooled AUC combines pairs. Slashes group configurations equal in every displayed column. STED rule removals: $-N$ near-miss credit, $-L$ literal routing, $-C$ cross-type numeric comparison. Bold marks maxima.

	decimal.js		big.js	
Metric	Macro	Pooled	Macro	Pooled
STED / $-N$ / $-C$	**1.000**	**1.000**	**.999**	**.999**
SLOT / SLOT: truncated cosine	.598	.618	.623	.624
Matched-text SLOT / $-L$	.619	.591	.612	.593
Exact leaf values + soft-F1	.500	.500	.500	.500

STED reaches within-group AUCs of **1.000/.999**, versus .619/.612 for matched-text SLOT: gains of .381 (group-bootstrap 95% interval $[.351, .409]$) and .387 ($[.363, .410]$). Removing literal routing recovers the control's scores; removing near-miss or cross-type numeric credit leaves AUC unchanged. The gain concerns numerical equivalence, which canonicalization can solve directly. The libraries share an author. App. C.3 gives the full protocol.

4.3.3 Recognition of accepted argument alternatives

Thresholds calibrated on other questions target a 1% empirical false-accept budget. STED accepts 36.1% of BFCL alternatives versus 33.7% for index-free SLOT, at observed false-accept rates of 1.04%/1.19% (Table 3). Hard negatives lower acceptance to 19.9%/16.7%, at false-accept rates of 1.03%/1.06%. These gains concern the 1% operating point; Table 1 also reports the stricter-budget and ranking results. App. F gives uncertainty.

Table 3: **Recognition of BFCL-accepted argument alternatives.** All scored external baselines are shown. Thresholds are calibrated on other questions to 1% empirical negative acceptance. TPR is the accepted fraction of 1,885 positive pairs; FPR is the accepted fraction of 1,343 negatives, or 2,727 with hard negatives. Both are measured on held-out questions and must be read jointly. Bold marks the best displayed value in each column. Additional calibration budgets, a normalization control, and intervals are in App. Table 21.

	Corpus negatives		With hard negatives	
Metric	TPR (%) ↑	FPR (%) ↓	TPR (%) ↑	FPR (%) ↓
STED	**36.1**	1.04	**19.9**	1.03
SLOT (strict paths)	28.5	1.19	11.3	1.06
SLOT (index-free paths)	33.7	1.19	16.7	1.06
Exact match	0.0	**0.00**	0.0	**0.00**
Levenshtein	6.5	1.19	3.7	1.10
Jaccard	5.8	0.67	6.1	0.84
ROUGE-L	1.5	1.27	1.5	1.14

Scope across targets. Table 1 separates gains on numerical equivalence and selected acceptance budgets from mixed text, program-return and external model-judge results. These instruments test different notions of similarity; their outcomes do not support a target-independent ordering of STED and SLOT.

4.3.4 Consistency of LLM structured outputs

At $T = .5$, mean $\bar{c}_s$ across models is .930 for Toucan and .669 for ShareGPT. These corpora differ in prompts, structure, and serving paths; 48/75 ShareGPT prompts request quizzes. Their difference does not isolate task format. Consistency is below one even at $T = 0$ for every Toucan model (Table 4); ShareGPT results appear in App. Table 24.

Table 4: **STED consistency on Toucan.** r_v is generation validity and coverage is the percentage of prompts with at least two valid outputs, both at $T = .5$. $\bar{c}_s$ averages conditional pairwise agreement over scoreable prompts. Rows describe the recorded model–serving configurations. Coverage groups organize the display; 90% is not a deployment threshold. † denotes proxy extraction. Means have no uncertainty intervals; their order is descriptive. Bold marks column maxima at the reported precision.

Model	r_v%	Coverage%	$\bar{c}_s^{T=0}$	$\bar{c}_s^{T=.5}$	$\bar{c}_s^{T=1}$
Coverage $\geq 90\%$					
Claude-Opus-4	97.1	97.3	.999	**.965**	.945
Nova-2-Lite†	96.6	97.1	.970	.964	**.952**
Claude-3.5-Haiku	97.8	98.2	.999	.959	.949
Claude-3.7-Sonnet	96.7	97.4	.968	.956	.944
Claude-3.5-Sonnet	97.3	97.9	.967	.955	.946
Claude-Sonnet-4	97.0	97.4	.977	.954	.938
Claude-Opus-4.5	97.6	97.8	.982	.947	.933
Claude-Sonnet-4.5	96.8	97.3	.995	.943	.926
Claude-Haiku-4.5	95.3	96.6	.992	.919	.894
Qwen3-32B†	96.8	97.4	.946	.916	.876
Qwen3-235B-A22B†	**98.1**	98.1	.928	.915	.889
Mimo-V2-Flash†	97.9	**99.1**	.941	.912	.879
Grok-4.1-Fast†	94.9	96.9	.876	.869	.861
Minimax-M2†	90.5	94.7	.881	.855	.809
Coverage $< 90\%$					
Llama-3.3-70B†	32.4	42.6	.979	**.965**	.946
GPT-4.1-Mini†	76.1	77.4	.927	.947	.926
Gemini-2.5-Flash-Lite†	54.0	63.1	**1.000**	.910	.877
GPT-OSS-120B†	82.3	89.5	.903	.885	.870

Temperature trends. Higher temperature generally increases output randomness (Renze & Guven, 2024). Across eleven temperatures, its Spearman association with $\bar{c}_s$ is negative for 17/18 Toucan and all 18 ShareGPT models; GPT-4.1-Mini is the Toucan exception ($\rho = .455$). This checks expected direction, not equivalence accuracy.

Consistency informs model selection. On Toucan at $T = .5$, observed consistency is .959 for Claude-3.5-Haiku and .947 for Claude-Opus-4.5, with similar validity (.978/.976). Without paired uncertainty estimates, this ordering is descriptive. Table 4 groups configurations by coverage and flags proxy extraction. A native-endpoint Gemini spot-check shifts validity by 6.3 points (App. J.1).

Coverage changes interpretation. Llama-3.3-70B's Toucan zero-filled mean $\bar{c}_0$ rises from .396 to .447 between $T = 0$ and $T = 1$, while conditional consistency falls from .979 to .946. Validity rounds to 32.5% at both endpoints, but prompt coverage rises from 40.5% to 47.2%. The identity $\bar{c}_0 = (N_s/N)\bar{c}_s$ explains this reversal and motivates reporting all three quantities under a fixed aggregation policy (App. I).

Cost of repeated-output scoring. With cached encodings and reuse of identical-pair scores (retaining pair multiplicity), scoring 8,436,780 valid pairs takes 650 seconds with 64 CPU workers on a `c7i.48xlarge` instance.

4.3.5 Effect of schema enforcement

Enforcement benefits depend on the model. Qwen3, GPT-OSS, and Minimax gain $+.139$ to $+.244$, with all nine paired-bootstrap 95% intervals above zero (Table 5). For Haiku and Sonnet, all six intervals include zero. These Claude checkpoints already achieve 1.000 free-form validity and .603–.966 conditional consistency, whereas Minimax has .689–.909 validity and .255–.464 consistency. Thus the Claude runs provide stronger free-form baselines, with no established additional consistency gain from enforcement in this setting.

Table 5: **Applying STED to schema enforcement.** All five models use the fixed primary score. n is the paired intersection; validity uses all 75 common prompts. 95% confidence intervals use 10,000 bootstrap resamples of paired prompts. Haiku and Sonnet denote Claude models. App. J records checkpoint and serving-protocol differences. Bold marks higher within-row consistency and validity at the reported precision.

Model	T	n	Free	Enforced	Δc [95% CI]	r_v: free/enf.
Qwen3-235B	0	74	.491	**.661**	.169 [.112,.232]	.956/**.979**
Qwen3-235B	0.5	74	.445	**.618**	.173 [.113,.237]	.972/**.989**
Qwen3-235B	1	74	.401	**.563**	.162 [.097,.232]	.965/**.983**
GPT-OSS-120B	0	53	.574	**.714**	.139 [.071,.206]	**.969**/.493
GPT-OSS-120B	0.5	60	.352	**.581**	.229 [.153,.306]	**.967**/.495
GPT-OSS-120B	1	57	.323	**.556**	.233 [.157,.309]	**.967**/.491
Minimax-M2	0	63	.464	**.642**	.179 [.106,.253]	.689/**.711**
Minimax-M2	0.5	75	.305	**.549**	.244 [.169,.320]	.909/**.952**
Minimax-M2	1	75	.255	**.473**	.218 [.140,.297]	.884/**.945**
Haiku-4.5	0	74	**.966**	**.966**	.000 [-.015,.014]	**1.000**/.987
Haiku-4.5	0.5	74	.657	**.668**	.011 [-.007,.029]	**1.000**/.985
Haiku-4.5	1	74	.603	**.617**	.014 [-.003,.036]	**1.000**/.987
Sonnet-4.5	0	74	**.823**	.820	-.003 [-.030,.024]	**1.000**/.987
Sonnet-4.5	0.5	74	**.692**	.688	-.004 [-.020,.011]	**1.000**/.987
Sonnet-4.5	1	74	.631	**.635**	.004 [-.017,.024]	**1.000**/.987

Valid format does not ensure stable outputs. Qwen3's free-form validity is already .956–.972, yet consistency is only .401–.491 and increases by .162–.173 under enforcement. Its principal measured weakness is repeatability, despite high parseability.

Consistency gains can carry validity costs. For GPT-OSS at $T = 0$, conditional consistency rises from .574 to .714 while validity falls from .969 to .493. Both Claude models also lose validity, to .985–.987. Enforcement therefore requires joint assessment of consistency and validity. Checkpoint and serving-protocol differences appear in App. J.

5 Ablation Study

Design and measurement. Human ablations use all binary-majority development judgments, counting metric ties as incorrect. The frozen external numerical test separately checks literal routing.

Matched text processing on development comparisons. We match SLOT's cosine mapping and text processing to STED, retaining its non-string comparisons (Table 6). STED makes 183 versus 170 correct structural decisions and 107 versus 83 supplement decisions: gains of 6.3 percentage points (source-cluster 95% interval $[2.6, 10.2]$; Holm-adjusted $p = .011$) and 17.1 points ($[8.6, 25.7]$; $p = .002$).

Table 6: **STED versus controlled SLOT variants.** Correct decisions (%) use all 208 structural and 140 supplement majorities; ties count as no correct decision. All variants share the encoder, exact paths, positional arrays and soft-F1 aggregation. Matched text processing includes STED's preprocessing, short-text fallback and chunking, with SLOT's non-string comparator. Bold marks the highest percentage in each column.

Metric	Structural (%) ↑	Supplement (%) ↑
STED	**88.0**	**76.4**
SLOT	79.8	57.9
SLOT: truncated cosine	80.8	62.1
SLOT: matched text processing	81.7	59.3

Individual-rule controls. Table 7 varies one setting at a time; App. E.2 reports the separate six-control paired analysis.

Table 7: **Controlled ablations of STED.** Each row changes one setting. Correct decisions use all binary-majority tasks as the denominator (208 structural; 140 supplement), including metric ties as no correct decision. This differs from the conditional agreement in Figure 1. Ties are reported separately; bold marks the highest correct-decision percentage in each set.

	Structural		Supplement	
Configuration	Correct (%) ↑	Ties	Correct (%) ↑	Ties
STED	**88.0**	5	76.4	1
Without near-miss credit	81.3	15	54.3	4
Without literal routing	87.5	3	**77.9**	1
Without cross-type numeric comparison	87.0	5	76.4	1
Require exact callee identity	85.6	7	49.3	4
Minimum over the same leaf scores	25.5	144	40.7	73
Align tool calls by name	82.7	5	75.0	2

Graded leaf comparisons distinguish near misses. Removing near-miss credit reduces correct decisions from 88.0% to 81.3% on the structural set and from 76.4% to 54.3% on the supplement, increasing ties. Retaining it adds 6.7 points (source-cluster 95% interval $[2.4, 12.2]$; Holm-adjusted $p = .025$) and 22.1 points ($[14.3, 30.0]$; $p < .001$), respectively. Other leaf-rule removals have smaller effects without established human-preference gains.

Exact callee identity lowers supplement agreement by 27.1 points ($p_{\mathrm{adj}} < .001$), without an established structural difference ($p_{\mathrm{adj}} = .694$). Graded name credit measures closeness without implying tool interchangeability.

Literal routing supports accepted-alternative recognition. Without literal routing, BFCL positives accepted fall from 36.1% to 23.2% at the 1% calibration budget (observed false acceptance 1.04% and 1.12%). With hard negatives, acceptance falls from 19.9% to 11.9% (false acceptance 1.03% and .95%). Stricter budgets do not uniformly favor STED (App. C.1).

The frozen numerical test localizes its .381/.387 AUC gains to this rule (Table 2); other numerical ablations leave AUC unchanged. Date, datetime, IP and text tests establish no general advantage (App. C.2).

Graded aggregation preserves decision coverage. Minimum aggregation can tie candidates through a zero-credit leaf. It decides only 64/208 structural and 67/140 supplement majorities, versus 203/208 and 139/140 for STED. Soft-F1, shared with SLOT, preserves coverage without guaranteeing agreement in every field.

Call correspondence is task-dependent. Name alignment changes structural correct decisions from 88.0% to 82.7% and supplement decisions from 76.4% to 75.0%; neither difference passes paired multiplicity correction. These development sets establish no benefit over positional correspondence.

6 Conclusion

STED measures structured-output agreement through tree-path correspondence, type-aware comparison, and graded aggregation. Human development comparisons support graded leaf credit, while frozen numerical tests isolate a benefit of literal routing and BFCL alternatives demonstrate acceptance gains at comparable measured error rates. Applying the fixed score to 2.1M generations from 18 LLMs reveals imperfect repeatability even at $T = 0$ and model-dependent enforcement effects. Consistency adds information for model selection alongside task quality; it should be reported together with validity and prompt coverage under a fixed comparison policy.

Limitations. Co-author human labels and local sandbox calls informed development; threshold splits do not remove this exposure. Frozen model judges do not replace independent human annotation. BFCL negatives can contain unresolved aliases. The generation corpus has narrow prompt coverage, heterogeneous serving paths, and no independent pair-equivalence labels. No experiment establishes a universal equivalence detector or population false-accept guarantee.

AI USE STATEMENT

In this work, we used generative AI tools to **implement methods** and to **design and give feedback on experiments and analysis code**: the metric implementation, the experiment harnesses, and the statistical analysis scripts were written with LLM coding assistance, and LLM critique was used to stress-test experimental designs and identify confounds in our validation instruments. We also used generative AI tools for tasks with recommended disclosure: drafting and editing parts of the manuscript for readability, formatting references, searching for and summarizing related literature, and generating figures from computed result tables. We also use Claude Opus 5.5 and GPT-6 Astra as blinded model judges in App. D; their choices are automated validation labels, not human annotations. Model generations form the repeated-output corpus. Every reported number is computed by a recorded script from frozen evaluation inputs and stored outputs, and the scripts are in the submission repository (App. K). Human judgments in §4.3.1 are from human annotators, not from LLMs; where we use LLM generations as experimental inputs, they are identified as such rather than treated as human labels. We have reviewed all AI-assisted work: LLM-written analysis code was verified against independently recomputed reference values, and the literature discussion draws on original papers and official technical documentation. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work involves human annotation and a public dataset release. Annotation: §4.3.1 reports judgments collected from co-author annotators on machine-generated JSON outputs; no personal or sensitive data is elicited, and annotators are not the subject of study. The rubric, interface, co-author participation, blinding status, and majority protocol are documented in App. E. Dataset release: the experimental data comprise metric-evaluation tasks and model generations derived from publicly released corpora (Toucan, ShareGPT, BFCL, and τ-bench). External text checks use PAWS, SICK and STS-B; program checks use CPython, decimal.js and big.js fixtures and pinned official τ-bench tools. The submission repository includes the co-author annotation audit and source records; the public archive contains derived generations and validation artifacts. The datasheet in App. K documents intended and out-of-scope uses. The main foreseeable misuse is treating a consistency score as a correctness score — a model that reliably emits the same wrong output scores highly. The datasheet and §6 state this limit.

REPRODUCIBILITY STATEMENT

All reported numbers are produced by recorded scripts from frozen evaluation inputs and stored model outputs; randomized analyses record their seeds. The submission repository contains the implementation and analysis scripts; the full raw-output deposit is publicly accessible at `https://huggingface.co/datasets/socb-anon-2441/socb`. App. K identifies the recorded model and serving metadata, frozen scoring environment, and sampling and analysis protocols. App. A specifies the primary score and its constants. App. E gives the complete annotation protocol. App. D records the frozen external selection, model-judge protocol, controls, and official-tool replay. Where a result depends on a choice that could be made otherwise (kernel definition, tie handling, aggregation), the alternative is reported alongside it.

REFERENCES

Arun63. sharegpt-quizz-generation-json-output. Hugging Face dataset, 2025a. URL `https://huggingface.co/datasets/Arun63/sharegpt-quizz-generation-json-output`.

Arun63. sharegpt-structured-output-json. Hugging Face dataset, 2025b. URL `https://huggingface.co/datasets/Arun63/sharegpt-structured-output-json`.

Berk Atil, Sarp Aykent, Alexa Chittams, Lisheng Fu, Rebecca J. Passonneau, Evan Radcliffe, Guru Rajan Rajagopal, Adam Sloan, Tomasz Tudrej, Ferhan Ture, Zhe Wu, Lixinyu Xu, and Breck Baldwin. Non-determinism of "deterministic" LLM settings. *arXiv preprint arXiv:2408.04667*, 2024. URL `https://arxiv.org/abs/2408.04667`. Version v5.

AWS Labs. Stickler: Structured object evaluation for GenAI. https://github.com/awslabs/stickler, 2025. Software documentation; accessed 24 September 2026.

Luca Beurer-Kellner, Marc Fischer, and Martin Vechev. Prompting is programming: A query language for large language models. In *PLDI*, 2023.

beyoru. Toucan-1.5M-structured-Qwen, 2025. URL https://huggingface.co/datasets/beyoru/Toucan-1.5M-structured-Qwen.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 1–14. Association for Computational Linguistics, 2017. doi: 10.18653/v1/S17-2001. URL https://aclanthology.org/S17-2001/.

Seperman Dehpour. DeepDiff: Deep difference and search of any Python object/data, 2026. URL https://github.com/seperman/deepdiff. Software documentation; accessed 26 September 2026.

Jan Drchal. Object aligner: A configurable json schema similarity score for graphs, applied to llm prompt optimization. https://arxiv.org/abs/2607.01972, 2026.

Di Feng, Kaixin Ma, Feng Nan, Haofeng Chen, Bohan Zhai, David Griffiths, Mingfei Gao, Zhe Gan, Eshan Verma, Yinfei Yang, Zhifeng Chen, and Afshin Dehghan. SO-Bench: A Structural Output Evaluation of Multimodal LLMs, 2025. URL https://arxiv.org/abs/2511.21750.

Nick Ferguson, Josh Pennington, Narek Beghian, Aravind Mohan, Douwe Kiela, Sheshansh Agrawal, and Thien Hang Nguyen. Extractbench: A benchmark and evaluation methodology for complex structured extraction. https://arxiv.org/abs/2602.12247, 2026.

Saibo Geng, Hudson Cooper, Michał Moskal, Samuel Jenkins, Julian Berman, Nathan Ranchin, Robert West, Eric Horvitz, and Harsha Nori. JSONSchemaBench: A rigorous benchmark of structured outputs for language models. *arXiv preprint arXiv:2501.10868*, 2025. URL https://arxiv.org/abs/2501.10868v3. Version v3.

Horace He and Thinking Machines Lab. Defeating nondeterminism in LLM inference. Thinking Machines Lab Blog, September 2025. https://thinkingmachines.ai/blog/defeating-nondeterminism-in-llm-inference/.

Thomas Hütter, Nikolaus Augsten, Christoph M. Kirsch, Michael J. Carey, and Chen Li. JEDI: These aren't the JSON documents you're looking for. In *Proceedings of the 2022 International Conference on Management of Data (SIGMOD '22)*. Association for Computing Machinery, 2022. Extended version: arXiv:2201.08099.

Paul Jaccard. Étude comparative de la distribution florale dans une portion des alpes et du jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37:547–579, 1901. doi: 10.5169/seals-266450. URL https://doi.org/10.5169/seals-266450.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2302.09664.

V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966. URL https://nymity.ch/sybilhunting/pdf/Levenshtein1966a.pdf.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81. Association for Computational Linguistics, 2004. URL https://aclanthology.org/W04-1013/.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *EMNLP*, 2023.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 216–223. European Language Resources Association (ELRA), 2014. URL https://aclanthology.org/L14-1314/.

Michael Mclaughlin. big.js: A small, fast JavaScript library for arbitrary-precision decimal arithmetic, 2026a. URL https://github.com/MikeMcl/big.js/tree/v7.0.1. Software and test fixtures, version 7.0.1; accessed 26 September 2026.

Michael Mclaughlin. decimal.js: An arbitrary-precision decimal type for JavaScript, 2026b. URL https://github.com/MikeMcl/decimal.js/tree/v10.6.0. Software and test fixtures, version 10.6.0; accessed 26 September 2026.

Tapan Parikh. Structured Output Collapses Answer Diversity Across 44 Language Models, 2026. URL https://arxiv.org/abs/2607.18476.

Shishir G. Patil, Huanzhi Mao, Fanjia Yan, Charlie Cheng-Jie Ji, Vishnu Suresh, Ion Stoica, and Joseph E. Gonzalez. The berkeley function calling leaderboard (BFCL): From tool use to agentic evaluation of large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 48371–48392. PMLR, 2025. URL https://proceedings.mlr.press/v267/patil25a.html.

Python Software Foundation. CPython 3.13.7 test suite. Versioned software test fixtures, 2025. URL https://github.com/python/cpython/tree/v3.13.7/Lib/test.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of EMNLP-IJCNLP*, pp. 3982–3992, 2019. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

Matthew Renze and Erhan Guven. The effect of sampling temperature on problem solving in large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 7346–7356, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.432. URL https://aclanthology.org/2024.findings-emnlp.432/.

Sentence Transformers. all-MiniLM-L6-v2 model card. Hugging Face, 2026. URL https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2. Model documentation; accessed 26 September 2026.

Abhinav Kumar Singh, Harsha Vardhan Khurdula, Yoeven D Khemlani, and Vineet Agarwal. The structured output benchmark: A multi-source benchmark for evaluating structured output quality in large language models. https://arxiv.org/abs/2604.25359, 2026.

Yifan Song, Guoyin Wang, Sujian Li, and Bill Yuchen Lin. The good, the bad, and the greedy: Evaluation of LLMs should not ignore non-determinism. *arXiv preprint arXiv:2407.10457*, 2024.

Břetislav Šopík and Tomáš Strenáčik. Tree edit distance for hierarchical data compatible with HMIL paradigm. https://arxiv.org/abs/2208.00782, 2022.

Kuo-Chung Tai. The tree-to-tree correction problem. *Journal of the ACM*, 26(3):422–433, 1979.

Darren Yow-Bang Wang, Zhengyuan Shen, Soumya Smruti Mishra, Zhichao Xu, Yifei Teng, and Haibo Ding. Slot: Structuring the output of large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 472–491, 2025a. doi: 10.18653/v1/2025.emnlp-industry.32. URL https://aclanthology.org/2025.emnlp-industry.32/.

Guanghui Wang, Jinze Yu, Xing Zhang, Dayuan Jiang, Yin Song, Tomal Deb, Xuefeng Liu, and Peiyang He. STED and consistency scoring: A framework for evaluating LLM structured output reliability. https://arxiv.org/abs/2512.23712, 2025b.

Brandon T. Willard and Rémi Louf. Efficient guided generation for large language models. *arXiv preprint arXiv:2307.09702*, 2023.

Xiaoyuan Wu, Weiran Lin, Omer Akgul, and Lujo Bauer. Estimating llm consistency: A user baseline vs surrogate metrics. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025. doi: 10.18653/v1/2025.emnlp-main.1554. URL https://aclanthology.org/2025.emnlp-main.1554/.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*, 2024. URL https://arxiv.org/abs/2406.12045.

Jiayi Yuan, Hao Li, Xinheng Ding, Wenya Xie, Yu-Jhe Li, Wentian Zhao, Kun Wan, Jing Shi, Xia Hu, and Zirui Liu. Understanding and mitigating numerical sources of nondeterminism in LLM inference. *arXiv preprint arXiv:2506.09501*, 2025. URL https://arxiv.org/abs/2506.09501.

Kaizhong Zhang and Dennis Shasha. Simple fast algorithms for the editing distance between trees and related problems. *SIAM Journal on Computing*, 18(6):1245–1262, 1989.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *ICLR*, 2020.

Yiwei Zhang, Chengke Wu, Li Wang, and Jianqiang Li. Where vs what: Decomposing structural and content failures in LLM-generated structured outputs. https://arxiv.org/abs/2608.25358, 2026. arXiv v1, submitted 26 August 2026.

Yuan Zhang, Jason Baldridge, and Luheng He. PAWS: Paraphrase Adversaries from Word Scrambling. In *Proceedings of NAACL-HLT*, pp. 1298–1308, 2019. doi: 10.18653/v1/N19-1131. URL https://aclanthology.org/N19-1131/.

Zhicheng Zhou, Jing Li, Suming Qiu, Junjie Huang, Linyuan Qiu, and Zhijie Sun. DeepJSONEval: Benchmarking Complex Nested JSON Data Mining for Large Language Models, 2025. URL https://arxiv.org/abs/2509.25922.

A Terminal-Path Score and Leaf Rules

All primary results compare terminals at identical paths, using positional array indices and no recursive alignment. Paths are tuples of typed object-key or array-index steps. Empty objects and arrays have distinct terminal markers; null is a value and missing fields have no terminal. For terminal-path sets P_A, P_B and shared-path credit $M = \sum_{p \in P_A \cap P_B} k_p(a_p, b_p)$,

$$d_{\mathrm{STED}} = \frac{|P_A \triangle P_B| + 2\sum_{p \in P_A \cap P_B}(1 - k_p)}{|P_A| + |P_B|}, \qquad s_{\mathrm{STED}} = 1 - d_{\mathrm{STED}} = \frac{2M}{|P_A| + |P_B|}.$$

The distance charges unmatched terminals one and matched pairs $2(1 - k_p)$. The score inherits SLOT's soft precision/recall aggregation. Every parsed JSON value has at least one terminal.

Leaf policy. Boolean/null values require exact type and value; empty-container markers match only their own kind. Number pairs and number–numeric-string pairs receive one within relative tolerance 10^{-9}, with absolute tolerance 10^{-12} only when at least one value is exactly zero. Otherwise,

$$k_{\mathrm{num}}(a, b) = .9 \max\left(0, 1 - 10\frac{|a - b|}{\max(|a|, |b|)}\right).$$

Zero–zero uses equality. Numeric-list literals take the minimum elementwise comparison; unequal lengths score zero.

Identical strings receive one; case-only differences after stripping surrounding whitespace receive .95. Detected literals use type-specific equality, then the numeric rule above or, for identifier-, code-, and word-like literals of at most 128 characters on both sides, $.9(1 - d_{\mathrm{lev}}/\max(|a|, |b|))$. Other unequal

Table 8: **Reading the STED score.** Each row changes only the stated part of the reference call. M is matched leaf credit. These exact examples illustrate the definition; they carry no independent equivalence labels.

Change from the reference	$\lvert P_B \rvert$	M	STED
Reorder object keys	3	3	1
`city: "Seattle"` → `"seattle"`	3	2.95	.9833
Remove `days`	2	2	.8000
`days: 1` → `2`	3	2	.6667

detected literals score zero. Callee names follow these ordinary leaf rules, without an identity gate on shared arguments.

Remaining strings use normalized Levenshtein similarity when $\min(|a|, |b|) < 4$ and truncated cosine otherwise. The encoder is `all-MiniLM-L6-v2`, with separator/case preprocessing. Strings reaching 300 characters use chunks of size 250 with overlap 50 and a symmetric mean of best chunk similarities.

Scope. Human judgments informed kernel design and configuration selection (App. E). The four-character boundary is a fixed heuristic whose effect is not isolated here; short abbreviations can be related despite little character overlap. Chunking limits encoding truncation but does not certify equivalence. Shared-argument credit despite callee changes, numeric normalization and near-miss credit likewise preclude a universal execution-equivalence interpretation. Corpus parsing and aggregation follow App. J.

A.1 Worked Example of the Primary Score

Consider the reference call

`{name: "weather", arguments: {city: "Seattle", days: 1}}`.

Its three terminal paths are the callee name and two arguments. Table 8 shows how near-miss credit and missing paths affect the score.

Exact match would tie the value-change and missing-field rows at zero. Terminal credit exposes their graded ordering; acceptability still depends on consumer semantics. Array positions preserve sequence sensitivity (alignment control: Section 5).

Distance axioms. For $A = \{\mathtt{x} : 1\}$, $B = \{\mathtt{y} : 1\}$ and $C = \{\mathtt{x} : 1, \mathtt{y} : 1\}$, both STED and strict-path SLOT give $d(A, B) = 1$ but $d(A, C) = d(C, B) = 1/3$, violating the triangle inequality even with exact leaves. STED also gives distance zero to $\{\mathtt{x} : 1\}$ and $\{\mathtt{x} : \mathtt{"1"}\}$, violating identity of indiscernibles on raw JSON. "Evaluation metric" denotes a scoring criterion, not a metric space.

B Generation Data and Sampling Protocol

The generation corpus contains 1,006 Toucan tool-calling prompts and 75 ShareGPT prompts. Each model–prompt condition has ten runs at eleven temperatures from 0 to 1. These are convenience samples with substantial repetition, not a representative sample of industrial workloads.

ShareGPT. We use `Arun63/sharegpt-quizz-generation-json-output` and `Arun63/sharegpt-structured-output-json` (Arun63, 2025a;b); their cards describe ShareGPT-formatted conversations rather than establish origin in the original ShareGPT collection. Our 80-prompt pool contains 48 quiz-generation and 32 other structured-output prompts. Five malformed references are excluded, leaving 48 quiz and 27 other prompts. The same 75-prompt subset is used for enforcement. Table 9 describes its structure and input lengths.

Toucan. The subset contains 515 single-turn-original and 491 single-turn-diversify examples, sampled from 119,287 single-turn examples of the upstream Toucan corpus (beyoru, 2025). Table 10 describes the 993 reference tools and argument types. A non-ASCII query is not necessarily multilingual; no language-stratified human validation is claimed for the English-oriented MiniLM encoder.

Table 9: ShareGPT Dataset Statistics

Metric	Value	Details
Sample Composition		
Total samples	75	(5 excluded[†])
Quiz Generation	48	64.0%
Structured Output	27	36.0%
Structure Statistics		
Max depth	4.0±1.0	range 2–7
Total fields	42.3±37.8	range 4–228
Arrays per sample	4.7±2.7	range 0–11
Nested objects	12.1±13.6	range 0–98
Field Type Distribution		
string	7,276	68.2%
integer	1,968	18.4%
array	907	8.5%
object	524	4.9%
Input Characteristics		
Prompt length	4,330±8,963	mean±std (tokens)
Range	172–73,522	tokens
Median	1,795	tokens

[†]5 samples excluded due to malformed ground truth JSON (truncated data or invalid syntax).

Table 10: Toucan Dataset Statistics

Metric	Value	Details
Sample Composition		
Total samples	1,006	–
Unique tools	993	across all samples
Tools per sample	95.7	mean available tools
Non-ASCII queries	56.4%	non-ASCII queries
Tool Call Distribution		
Total calls	1,583	across all samples
Calls per sample	1.57	mean (range 1–14)
1 call	61.9%	623 samples
2 calls	23.5%	236 samples
3 calls	11.9%	120 samples
4+ calls	2.7%	27 samples
Parameter Statistics		
Total parameters	2,366	across all calls
Params per call	1.48	mean (range 0–19)
string	48.7%	1,152 params
float	24.8%	587 params
int	18.2%	431 params
list	4.3%	102 params
bool	3.1%	73 params
dict	0.9%	21 params
Tree Depth (per single tool call)		
Mean depth	2.9±0.5	range 2–8
Query Characteristics		
Query length	128±75	mean±std (tokens)
Range	13–563	tokens
Median	105	tokens

C SLOT Baseline and Comparison Policy

SLOT (Wang et al., 2025a) provides a formula but no reference implementation used by this study. Our reimplementation compares values at corresponding ancestral key paths, gives missing paths zero credit, and takes the harmonic mean of soft precision and recall. It uses the same `all-MiniLM-L6-v2` encoder as STED.

Value comparison. The published formula specifies Sentence-BERT (Reimers & Gurevych, 2019) comparisons but leaves numerical implementation choices open. Our reimplementation maps cosine to $[0, 1]$ by $(1+\cos)/2$; identical strings receive one. Non-string leaves use equality with a string-form fallback on type mismatch. These are disclosed implementation choices, not uniquely specified by SLOT's formula. STED's prose kernel instead truncates cosine at zero and its typed leaf rules grade

selected literal changes. Thus comparisons between the complete metrics do not isolate any one kernel change.

Array paths. The SLOT formula does not uniquely specify list-index handling. The primary baseline uses the same typed terminal paths and empty-container markers as STED, with positional list indices (strict paths); the accepted-alternatives and execution tables also report an index-free reading. We do not treat lower scores under one reading as evidence of lower accuracy. Array-order equivalence depends on the intended task; primary STED fixes positional correspondence across all applications.

C.1 Matched Text Processing and Target Dependence

Controls. All eight variants share the encoder, typed terminal paths, positional arrays and inherited soft-F1 aggregation (App. A). SLOT controls replace affine with truncated cosine, then add STED's preprocessing, short-string fallback and symmetric chunk matching while retaining SLOT's non-string comparisons. Matched text adds neither case-only credit nor literal routing. STED controls remove near-miss credit, literal routing or cross-type numeric comparison, or use exact typed leaves.

Human development comparisons. The same 456 tasks yield 208 structural and 140 supplement majorities. Metric ties earn no correct decision: STED scores 183/208 (88.0%) and 107/140 (76.4%), with five and one ties; matched-text SLOT scores 170/208 and 83/140, with four and one ties. Table 11 uses 5,000 paired source-cluster bootstrap resamples and 10,000 source-level sign swaps (seed 20260925). Holm correction covers seven controls within each set, separately from the six-ablation and eight-baseline families; structural STED versus SLOT has adjusted $p = .018$. These judgments informed configuration selection, so the comparisons remain development evidence. Conditional agreement excluding metric ties is retained in Table 19.

Table 11: **Paired tests for the frozen SLOT comparison.** Differences are STED minus the named control, in percentage points of correct decisions over all majority tasks. Intervals use source-cluster resampling; p_{adj} uses Holm correction over these seven controls within each set. This family differs from the six ablations and eight external baselines reported elsewhere.

Control	Structural Δ [95% CI]	p_{adj}	Supplement Δ [95% CI]	p_{adj}
SLOT	$+8.2$ $[+2.8, +13.7]$	.018	$+18.6$ $[+8.6, +28.6]$	.004
SLOT: truncated cosine	$+7.2$ $[+3.0, +11.9]$	.015	$+14.3$ $[+5.7, +22.9]$	.009
SLOT: matched text processing	$+6.3$ $[+2.6, +10.2]$	.011	$+17.1$ $[+8.6, +25.7]$	.002
Without near-miss credit	$+6.7$ $[+2.4, +12.2]$	.018	$+22.1$ $[+14.3, +30.0]$	$< .001$
Without literal routing	$+0.5$ $[-1.0, +2.2]$	1.000	-1.4 $[-4.3, +1.4]$	1.000
Without cross-type numeric comparison	$+1.0$ $[+0.0, +2.5]$	.994	$+0.0$ $[+0.0, +0.0]$	1.000
Exact leaf values + soft-F1	$+22.6$ $[+14.7, +30.3]$	$< .001$	$+38.6$ $[+26.4, +50.0]$	$< .001$

External comparison. Frozen τ-bench model judgments on 145 triples from 91 sources support graded comparison over typed exact leaves, but none of the three primary STED–SLOT contrasts is significant (App. D). They establish no independent advantage for STED-specific rules and supply no new human validation.

Accepted alternatives. The controls score 1,885 BFCL positives and two negative sets under leave-one-question-out calibration (App. F). Literal routing improves acceptance at the 1% budget; removing near-miss credit improves hard-set acceptance at 0.1% from 85 to 157 positives with the same three false accepts (Table 12). These effects depend on the target. SLOT uses shared positional paths here; the main comparison also includes index-free SLOT.

Text scope. PAWS-Wiki (Zhang et al., 2019) has 8,000 development and 8,000 test paraphrase pairs, each sentence wrapped in one JSON field. Configurations are frozen; development labels set thresholds. Test AUC is .615 for STED versus .631 for SLOT: difference $-.0159$, paired 95% interval $[-.0208, -.0110]$, using 1,000 bootstraps over 1,616 shared-sentence components. Article IDs are unavailable. Splits share no exact sentences; encoder-training exposure is unknown.

All test strings have 32–216 characters; neither short-string fallback, chunking nor paired literal routing applies. Matched-text SLOT falls to .6151, localizing the difference to preprocessing. The controlled perturbation suite is near saturation: STED/SLOT correctly order 736/737 and 737/737 source- and ratio-matched semantic/expression comparisons. These text checks do not establish a general semantic advantage or validate complete structured outputs.

Table 12: **Target dependence of the frozen controls.** BFCL reports TP/FP: 1,885 positives, 1,343 corpus negatives, 2,727 negatives with the hard extension. Budgets refer to calibration, not achieved test error. PAWS AUC uses 8,000 externally labeled test pairs as JSON text fields. Slashes in metric labels group configurations equal in every displayed column. STED rule removals: $-N$ near-miss credit, $-L$ literal routing, $-C$ cross-type numeric comparison. Bold marks maximal TP or AUC; read TP with FP.

Metric	BFCL corpus, 1%	BFCL hard, 1%	BFCL hard, 0.1%	PAWS AUC
STED / $-C$	**680**/14	376/28	85/3	.615
SLOT / SLOT: truncated cosine	538/16	213/29	109/3	**.631**
Matched-text SLOT	442/15	251/26	114/3	.615
$-N$	674/14	**383**/28	**157**/3	.615
$-L$	437/15	225/26	38/3	.615
Exact leaf values + soft-F1	17/10	23/24	0/0	.501

C.2 External Labels and Standard-Library Returns

Design and scope. STED and seven controls are fixed before input retrieval, with no retuning or new human judgments or LLM generations. These tests assess textual relatedness and parser-return equivalence; they do not independently validate preferences over complete structured outputs. The earlier human judgments remain development evidence.

Human relatedness labels. SICK (Marelli et al., 2014) supplies 4,906 official Zenodo test pairs; STS-B (Cer et al., 2017) supplies 1,379 from `sentence-transformers/stsb`. Sentences occupy one JSON text field. Table 13 reports Pearson (linear) and Spearman (rank) correlations with human scores.

Table 13: **External human relatedness labels.** Pearson r and Spearman ρ use all 4,906 SICK and 1,379 STS-B test pairs; each sentence is one JSON text field. STED and matched-text SLOT scores coincide; exact-leaf scores are constant (undefined correlations). Slashes group configurations equal in every displayed column. STED rule removals: $-N$ near-miss credit, $-L$ literal routing, $-C$ cross-type numeric comparison. Bold marks maxima.

	SICK		STS-B	
Metric	ρ	r	ρ	r
STED / Matched-text SLOT / $-N$ / $-L$ / $-C$	.7712	.8355	.8179	.8276
SLOT	**.7715**	**.8362**	**.8203**	.8274
SLOT: truncated cosine	**.7715**	.8357	**.8203**	**.8288**
Exact leaf values + soft-F1	—	—	—	—

STED, matched-text SLOT and all three STED leaf-rule removals score all 6,285 pairs identically, establishing no additional-rule advantage. Exact leaf equality is constant, so its correlation is undefined.

Paired intervals use 1,000 bootstraps (seed 20260926) over shared-normalized-sentence components: 865 for SICK (largest 357 pairs), 1,178 for STS-B (largest 25). STED minus truncated-cosine SLOT has Spearman difference $-.000329$ (pointwise 95% interval $[-.000951, -.000008]$) and $-.002385$ ($[-.005071, .000391]$), respectively. Clusters capture sentence reuse, not original image/document grouping. Test sentences have no exact normalized overlap with prior human or PAWS inputs, but share 3,959/291 sentences with their own training/development splits. Those labels did not select configurations; encoder exposure is unknown.

Program protocol. String literals come from CPython 3.13.7 test fixtures (Python Software Foundation, 2025), with date/datetime candidates restricted to `fromisoformat` tests. Each becomes a named call with one string argument. Native return equality under Python 3.13.7's `date.fromisoformat`, `datetime.fromisoformat`, `ip_address` and `Decimal` supplies labels; fixture code is read statically.

We exclude invalid inputs, strings over 256 characters, and decimal strings over 64 characters, nonfinite values or absolute adjusted exponents above 100. Positives pair distinct strings with equal

returns; negatives use fixed random sampling and character proximity without metric scores. Return-value groups are hash-split into calibration/test, with negatives confined to their split and positives capped at 1,000 per family and split by hash. Independent replay verifies all 2,543 labels without discrepancies; no return-value group crosses splits.

Table 14: **External standard-library return equivalence.** AUC uses 35 date, 117 datetime, 42 IP and 2,019 Decimal test pairs; Macro averages these four families. Decimal† excludes zero-reference groups post hoc (36 positives, 59 negatives, 23 return-value groups), without adding a Macro term. Slashes group frozen configurations equal in every displayed column. STED rule removals: $-N$ near-miss credit, $-L$ literal routing, $-C$ cross-type numeric comparison. Bold marks maxima.

Metric	Date	Datetime	IP	Decimal	Decimal†	Macro
STED / $-C$	.594	.641	.609	**.934**	.909	.694
SLOT / SLOT: truncated cosine	.605	.654	**.649**	.810	.671	.680
Matched-text SLOT	**.769**	**.757**	.642	.720	.614	**.722**
$-N$	.500	.565	.510	**.934**	**.914**	.627
$-L$	**.769**	**.757**	.642	.720	.613	**.722**
Exact leaf values + soft-F1	.500	.500	.500	.500	.500	.500

Results. The 330 calibration and 2,213 test pairs include 9 date, 19 datetime, 13 IP and 24 Decimal positive test groups. Equal-family macro AUC is .694 for STED versus .722 for matched-text SLOT. Predeclared 1%/5% negative-calibration budgets are not test-error guarantees: STED's Decimal false acceptance is 32/1,019 (3.14%) at 1%. Date/IP calibration has only 6/14 negatives, limiting threshold resolution.

Post-hoc Decimal diagnostic. Of 1,000 positives and 1,019 negatives, 964 positives represent zero. Excluding zero-reference groups leaves 95 pairs from 23 groups (36 positives, 59 negatives): AUC is .909 for STED, .671 for SLOT, .614 for matched-text SLOT and .613 without literal routing. Full-set equal-reference-group AUC is .907 versus .556 for STED/matched-text SLOT. Both diagnostics were chosen after observing concentration; neither changes the original test, scores or labels. The small nonzero sample supports a numerical gain beyond zero, but removing near-miss credit improves AUC to .914. Date, datetime, IP and text comparisons establish no general advantage.

C.3 Balanced External Numerical Confirmation

Protocol. Following the Decimal diagnostic, we fix all eight configurations, terminal representation, encoder, sampling and analysis before retrieving official decimal.js 10.6.0 and big.js 7.0.1 fixtures (Mclaughlin, 2026b;a). The libraries share an author and some values, so they are two collections rather than independent replications.

Static extraction retains quoted decimal strings, excluding comments, escapes, invalid/nonfinite values, strings over 64 characters and absolute adjusted exponents above 100. No strings are synthesized. Library constructors and equality methods supply labels; fixture code is not executed. Exact Python `Decimal` independently verifies accepted returns and all 8,222 labels without discrepancies.

Balanced construction. Each exact-value group contributes at most two hash-selected positives between distinct strings. Each positive anchors at most one hash-random and one character-nearest different-value negative; duplicates are removed. Candidate pools contain at most two strings per value. Group hashes assign calibration/test 30:70 consistently across libraries, with no value crossing splits and a cap of 1,000 groups per library and split. Calls have a fixed library name and one string argument.

Primary analysis excludes pairs containing a recorded CPython value, leaving 649 positives/1,235 negatives for decimal.js and 1,016/1,984 for big.js. Every retained reference group has both labels. Table 2 reports all eight configurations.

Results and uncertainty. The primary endpoint weights reference-group AUCs equally. STED exceeds matched-text SLOT by .381 on decimal.js (pointwise 95% interval $[.351, .409]$) and .387 on big.js ($[.363, .410]$). Removing literal routing gives matched-text SLOT's AUCs; removing near-miss credit or cross-type numeric comparison leaves STED unchanged.

Table 15: **Frozen external LLM judges.** Primary source-macro agreement (%): 125 decisive judgments each, spanning 84/83 sources for Opus/Astra. Metric ties earn no correct decision. All six metrics are frozen; truncate is secondary. Bold marks column maxima at reported precision.

Metric	Opus 5.5	GPT-6 Astra
STED	80.46	78.71
SLOT affine	80.36	78.61
SLOT truncate (secondary)	**81.55**	**79.82**
SLOT matched text	81.25	79.52
SLOT matched text + numeric	80.65	78.92
Typed exact leaf F1	59.13	60.44

Intervals use 5,000 reference-group bootstraps (seed 20260926); negatives can recur across groups, and four prespecified contrasts lack multiplicity adjustment. Archived secondary analyses report pooled AUC, all eligible values, magnitude strata and 1%/5% calibration budgets with achieved false-accept rates.

Scope. The evidence supports literal routing for numerical argument equivalence; library canonicalization solves this target by construction. It supplies neither independent human-preference validation nor an execution guarantee. CPython-value exclusion does not establish absence from earlier research or encoder training; human development exposure is unchanged.

D Frozen External Judges and Official Tool Replay

Source and selection. Historical τ-bench (Yao et al., 2024) airline/retail data contain 1,980 GPT-4o and Claude Sonnet 3.5 trajectories from 165 tasks. For each generator and identical task specification, we take the lowest three complete, nonempty, parseable trials. Hash selection (seed 20260926; cap 20 per view/domain/generator) uses no rewards, scores or judgments. Each triple has three distinct outputs: a shared tool-name/occurrence selected by hash, or complete ordered call sequences, retaining all arguments. The 145 evaluation triples contain 65 full-call and 80 sequence comparisons from 91 domain–task sources and 378 trajectories. Sources span generators/views; reactive user utterances can differ. There is no calibration split.

Frozen metrics and rubric. Selection, inputs, six metrics, `all-MiniLM-L6-v2` encoder and rubric are frozen before judgments, without retuning. Metrics share terminal correspondence and aggregation (App. C.1): STED, affine/truncated-cosine SLOT, matched-text SLOT, its finite-`Decimal`-equality extension (excluding booleans), and typed exact leaf F1. The rubric asks which candidate preserves reference meaning and structure across values, types, nesting, collections and arguments, without metric formulas or fixed weights. The reference need not be correct; choices are A, B, tie or unclear.

Blinded model judgments. We configure Opus 5.5 and GPT-6 Astra through Bedrock, using default sampling and high reasoning effort for Astra; temperature is not controlled at zero. Opaque IDs and seeded candidate randomization hide metric names/scores, generators, rewards, success labels and paper claims. Tools and external retrieval are prohibited and audited. All 80 batches succeed: 316 labels, comprising 145 primary judgments and 13 position swaps per model. These are model labels; the earlier human judgments informed metric development.

Primary agreement and uncertainty. Each judge gives 125 decisive choices, 20 ties and no unclear/failed labels. Correct decisions are averaged within source, then across 84 Opus/83 Astra decisive sources. Judge ties are excluded; metric ties earn zero within the denominator, as in the primary human analysis. Paired inference uses 5,000 source-cluster bootstraps for pointwise percentile 95% intervals and 10,000 two-sided source-mean sign swaps (seed 20260926), with Holm adjustment over eight contrasts (two judges, four controls). Truncated-cosine SLOT is predeclared secondary: STED differences are -1.09/-1.10 points, outside that family.

Interpretation and judge reliability. Graded comparison beats exact leaves, partly through fewer metric ties; no STED–SLOT superiority or equivalence is established. Judges agree on 111/119 jointly decisive items (93.3%), or 125/145 including ties (86.2%). Each preserves 12/13 swapped

Table 16: Primary source-macro contrasts: STED minus control, in percentage points. Intervals are pointwise source-cluster 95% CIs; p_{H} is Holm-adjusted across eight contrasts.

	Opus 5.5		GPT-6 Astra	
Control	Difference [95% CI]	p_{H}	Difference [95% CI]	p_{H}
SLOT affine	$+0.10$ $[-4.37, +5.16]$	1.000	$+0.10$ $[-4.42, +4.92]$	1.000
SLOT matched text	-0.79 $[-2.98, +0.79]$	1.000	-0.80 $[-3.01, +0.80]$	1.000
SLOT text + numeric	-0.20 $[-1.79, +1.19]$	1.000	-0.20 $[-1.81, +1.20]$	1.000
Typed exact leaf F1	$+21.33$ $[+14.09, +29.17]$	0.000800	$+18.27$ $[+11.65, +25.90]$	0.000800

Table 17: Secondary micro agreement: correct/125 (%), with metric-tie counts on the same decisive items. Bold marks agreement maxima at reported precision.

	Opus 5.5		GPT-6 Astra	
Metric	Correct (%)	Ties	Correct (%)	Ties
STED	101/125 (80.8)	4	98/125 (78.4)	5
SLOT affine	102/125 (81.6)	3	98/125 (78.4)	4
SLOT truncate (secondary)	**103/125 (82.4)**	3	**99/125 (79.2)**	4
SLOT matched text	102/125 (81.6)	3	**99/125 (79.2)**	4
SLOT matched text + numeric	101/125 (80.8)	3	98/125 (78.4)	4
Typed exact leaf F1	71/125 (56.8)	42	71/125 (56.8)	39

judgments; changes involve ties, with no decisive reversal (8/8 and 9/9 repeats). This small check cannot establish position invariance, and shared sources preclude treating judges as independent source replications.

Official replay and outcome labels. We restore each official initial database and execute actual calls in order: all 378/378 trajectories qualify and 2,804/2,804 returns exactly match historical logs, with no incompatibility exclusions. The official data and tool code share one pinned revision; a finite-arithmetic guard precedes the calculator without rewriting accepted expressions. Labels compare final full database states within task, independently of reward or success. All three sequence pairs per triple, deduplicated across views, yield 378 pairs, 181 positives and 91 sources; 45 sources have both labels for macro AUC. Requiring at least one changed final state leaves 296 pairs, 99 positives and 78 sources (33 macro-eligible). The other 82 pairs are unchanged-state positives, with undefined AUC.

Table 18: Descriptive final-state equality AUCs for all frozen metrics. Macro averages sources with both labels (45 overall; 33 with a changed state). Bold marks column maxima at reported precision.

	All 378 pairs		296 changed-state pairs	
Metric	Pooled	Macro	Pooled	Macro
STED	0.6284	**0.6047**	0.6070	0.5867
SLOT affine	**0.6364**	0.5917	**0.6133**	**0.5910**
SLOT truncate (secondary)	0.6296	0.5975	0.6072	0.5830
SLOT matched text	0.6285	0.5975	0.6103	0.5830
SLOT matched text + numeric	0.6271	0.6019	0.6067	0.5830
Typed exact leaf F1	0.5825	0.5749	0.5455	0.5691

Requiring identical ordered observations as well leaves four positives among 378 pairs: pooled AUCs in table row order are 0.9853, 0.9846, 0.9846, 0.9853, 0.9853 and 0.9933; all macro AUCs are 1.0000 on only four eligible sources. This sparse result supplies no strong validation claim.

Scope. Final-state equality measures effects, not complete intent, dialogue meaning or policy compliance. STED leads overall macro AUC; SLOT affine leads the other three columns, establishing no general advantage. Coverage is limited to two tool domains and historical generators; public inputs may have appeared in judge training.

E PAIRWISE HUMAN ANNOTATION PROTOCOL

Annotators and interface. Three co-author annotators completed the 256-task structural set and 200-task supplement as a research contribution. Each screen shows a reference output and two candidate outputs, with differing fields listed. Metric names and candidate model names are hidden, and left/right positions are randomized. Annotators know the research hypothesis and are instructed not to confer.

Decision rubric. The instruction is to "Pick the candidate closer in meaning to the reference." Annotators judge closeness to the reference, rather than which output they prefer in general. Both candidates can contain errors. *Equally close* is permitted when they cannot distinguish the candidates, and *Can't tell* when they cannot interpret the reference; neither response is penalized. This is a preference task, not an execution-equivalence label.

Analysis unit. Whole-document choices define the reported judgments. Some tasks additionally request field-level choices or offer long-text reading aids; those auxiliary responses are excluded from the 1,368 document-level judgments. A binary majority requires at least two identical candidate choices. App. E.1 reports the set-specific denominators, tie handling, design exposure, and annotator agreement.

Primary estimand and inference. Correct decisions use all binary-majority tasks, with metric ties earning zero: STED scores 183/208 (88.0%) on the structural set and 107/140 (76.4%) on the supplement. Score differences of at most 10^{-6} are ties. Conditional agreement excluding metric ties is secondary: 183/203 (90.1%) and 107/139 (77.0%). Paired inference uses 5,000 source-cluster bootstraps for pointwise percentile 95% intervals and 10,000 two-sided source-level sign swaps (seed 20260925). Holm correction applies separately within each set to three families: eight external baselines, seven matched controls (App. C.1), and six primary-rule controls (App. E.2). All three annotators' labels informed final configuration selection; these corrections do not make the comparisons independent validation.

E.1 THREE-ANNOTATOR PAIRWISE AUDIT

Structural-set composition. The 256 structural tasks come from a 704-task pool. Of 208 binary-majority tasks, 120 have unanimous binary decisions; 48 tasks lack a binary majority. Equal/unclear votes are not assigned a winner.

Against A/B/C individually, STED agrees on 173/212, 165/193, and 180/212 metric-decided structural tasks; SLOT agrees on 164/213, 153/199, and 163/213. On the unanimous subset, agreement is 115/119 for STED and 107/119 for SLOT. These denominators exclude each metric's ties.

Agreement between annotators. On jointly binary-decided structural tasks, A–B agree on 142/180 (78.9%; Cohen $\kappa = .578$), A–C on 150/194 (77.3%; .546), and B–C on 156/184 (84.8%; .695). These kappas exclude either annotator's equal/unclear choices; they are not three-category or three-annotator kappa. C gives 217 binary choices, 33 equal, and 6 unclear. The main pool contains 212 source items, of which 132 occur in the structural subset; several comparison tasks can share a source item.

Disagreement supplement. All three annotators completed 200 tasks selected for opposite orderings of a callee-identity STED configuration and SLOT, with margins at least .05. Selection did not require opposite orderings under the final primary score. Against A/B/C, STED agrees on 105/150, 109/141, and 120/182 metric-decided tasks; SLOT agrees on 80/151, 84/142, and 88/182. There are 140 binary-majority tasks, including 69 unanimous decisions, leaving 60 without a binary majority. Jointly binary-decided annotator agreement is 87/113 for A–B, 88/141 for A–C, and 103/134 for B–C.

Secondary conditional comparisons. Baselines compare the displayed candidates' raw scores; membership in a larger candidate top set is not a pairwise preference. Table 19 reports both primary correct/all-majority fractions and secondary conditional agreement. STED:SLOT discordance is 22:6 on structural tasks (nominal exact $p = .004$) and 41:14 on the supplement ($p < .001$). These task-level tests condition on jointly decided disagreements; primary inference uses the source-cluster protocol above.

Table 19: Agreement with the three-annotator majority. STED denotes the primary path-F1 score; the other rows are external baselines. Primary entries are correct/N (%), with common denominators $N = 208$ (structural) and $N = 140$ (supplement); metric ties count zero. Secondary conditional entries are correct/decided (%), excluding metric ties. All tie counts are shown; 0/0 (–) means conditional agreement is undefined because every metric decision ties. These are descriptive results on development data used for configuration selection.

	Structural majority ($N = 208$)			Supplement majority ($N = 140$)		
Metric	Primary	Conditional	Ties	Primary	Conditional	Ties
STED	183/208 (88.0)	183/203 (90.1)	5	107/140 (76.4)	107/139 (77.0)	1
SLOT	166/208 (79.8)	166/205 (81.0)	3	81/140 (57.9)	81/140 (57.9)	0
BERTScore	147/208 (70.7)	147/208 (70.7)	0	104/140 (74.3)	104/140 (74.3)	0
Levenshtein	144/208 (69.2)	144/205 (70.2)	3	109/140 (77.9)	109/139 (78.4)	1
ROUGE-L	129/208 (62.0)	129/172 (75.0)	36	105/140 (75.0)	105/137 (76.6)	3
TED	113/208 (54.3)	113/150 (75.3)	58	102/140 (72.9)	102/130 (78.5)	10
Jaccard	113/208 (54.3)	113/168 (67.3)	40	57/140 (40.7)	57/116 (49.1)	24
DeepDiff	80/208 (38.5)	80/175 (45.7)	33	85/140 (60.7)	85/139 (61.2)	1
Exact match	0/208 (0.0)	0/0 (–)	208	0/140 (0.0)	0/0 (–)	140

Table 20: **Paired configuration diagnostics.** Δ is primary STED minus the named control, in percentage points on the common majority-task denominator. Positive values favor primary STED. Intervals are source-cluster bootstrap 95% intervals; p_{adj} uses Holm correction over six controls per set. These are development-exposed diagnostics. Effect signs and uncertainty are reported without ranking the rules.

Control	Structural Δ [95% CI]	p_{adj}	Supplement Δ [95% CI]	p_{adj}
near-miss credit	$+6.7$ $[+2.4, +12.2]$	.025	$+22.1$ $[+14.3, +30.0]$	$< .001$
literal routing	$+0.5$ $[-1.0, +2.2]$	1.000	-1.4 $[-4.3, +1.4]$	1.000
cross-type numeric comparison	$+1.0$ $[+0.0, +2.5]$	.994	$+0.0$ $[+0.0, +0.0]$	1.000
Require exact callee identity	$+2.4$ $[-0.8, +5.7]$	.694	$+27.1$ $[+15.7, +37.9]$	$< .001$
Minimum over the same leaf scores	$+62.5$ $[+53.6, +71.6]$	$< .001$	$+35.7$ $[+27.1, +44.3]$	$< .001$
Align tool calls by name	$+5.3$ $[+1.4, +10.5]$	.128	$+1.4$ $[-2.9, +5.7]$	1.000

E.2 Controlled primary-score diagnostics

The same 256 structural and 200 supplement tasks are scored with STED and six alternatives. Three remove near-miss credit, string-literal routing, or cross-type numeric comparison; one requires exact callee identity. Two replace aggregation with a minimum or align calls by name. Alignment changes correspondence while retaining graded callee scoring. All variants use the same inputs and labels.

Common denominator and paired uncertainty. Table 20 reports primary-minus-control differences using all binary-majority tasks, with metric ties earning zero. Inference follows the setup above, with the six-rule Holm family applied within each set.

Interpretation. Near-miss credit improves correct-decision fractions on both sets and reduces ties. Exact callee identity lowers agreement on the selected supplement; literal routing and numeric comparison have smaller effects. These findings characterize the tested development data and do not establish benefits on unseen sources.

F BFCL-Accepted Argument Alternatives

Construction. From BFCL v4's eight AST categories, the canonical call takes the first accepted value of each parameter and omits optional parameters. For parameters with multiple accepted values, every pair of up to four distinct values yields a positive pair differing at that parameter only. This produces 1,885 positives: 120 are equal after case, spacing, punctuation and JSON-escaping normalization, and 1,765 are substantive alternatives. The labels describe BFCL's accepted values, rather than equality of observed execution returns.

Negatives pair the canonical value with a value accepted in another question for the same function and parameter, falling back to the same parameter name in another function. Alias components join normalized values connected through accepted alternatives; 137 of 1,480 candidate negatives are removed by this rule, leaving 1,343. The harder set adds 1,384 character-nearest values from other questions outside the alias component and absent from existing negatives. Unclosed aliases can remain, and these negatives are not a random deployment sample.

Held-out threshold calibration. For source question q, sort the M_q negative scores from all other questions as $z_{(1)} \geq \cdots \geq z_{(M_q)}$. At empirical calibration budget α, use

$$\theta_{-q}(\alpha) = z_{(\lfloor \alpha M_q \rfloor + 1)}, \qquad a_i = \mathbf{1}[s_i > \theta_{-q}(\alpha)]. \tag{4}$$

Strict acceptance handles ties conservatively. We sum accepted positives and negatives over held-out questions. The calibration budget need not equal the measured test false-accept rate. Only threshold fitting is held out: alias construction and hard-negative selection use the assembled corpus. This procedure does not supply a population risk guarantee.

Uncertainty. Table 21 includes both 1% and 0.1% calibration budgets across 4,612 retained pairs. Its 95% intervals use 1,000 source-question bootstrap resamples of fixed held-out decisions, without refitting thresholds. They describe variation among the observed questions and exclude uncertainty from configuration selection and alias construction.

Table 21: Leave-one-question-out thresholds on BFCL's accepted-alternative key (1,885 positives). FA is observed false accepts; acc is positives accepted, with a source-question bootstrap 95% interval. Bold marks lowest FA and highest acc in each block. Test false-accept rates can exceed calibration budgets.

	1% calibration budget			0.1% calibration budget		
Metric	FA	acc	95% CI	FA	acc	95% CI
Corpus negatives (1,343)						
Normalized equality	**0**	120	[94, 149]	**0**	120	[93, 150]
Exact match	**0**	0	[0, 0]	**0**	0	[0, 0]
Levenshtein	16	123	[95, 155]	2	3	[0, 7]
Jaccard	9	109	[68, 154]	**0**	95	[56, 140]
ROUGE-L	17	29	[9, 56]	**0**	0	[0, 0]
STED	14	**680**	[603, 770]	2	**367**	[316, 422]
SLOT (strict paths)	16	538	[465, 612]	3	174	[137, 214]
SLOT (index-free paths)	16	636	[550, 718]	3	276	[216, 339]
Corpus + hard negatives (2,727)						
Normalized equality	**0**	120	[92, 148]	**0**	120	[93, 148]
Exact match	**0**	0	[0, 0]	**0**	0	[0, 0]
Levenshtein	30	70	[50, 90]	3	2	[0, 5]
Jaccard	23	115	[73, 162]	**0**	95	[54, 139]
ROUGE-L	31	29	[8, 57]	**0**	0	[0, 0]
STED	28	**376**	[325, 435]	3	85	[63, 111]
SLOT (strict paths)	29	213	[173, 256]	3	109	[83, 139]
SLOT (index-free paths)	29	315	[254, 382]	4	**210**	[159, 267]

Operating-point dependence. At the 1% budget, STED accepts 680 corpus positives and 376 with hard negatives, compared with 636 and 315 for index-free SLOT. At the 0.1% budget, the counts are 367 versus 276 on corpus negatives, but 85 versus 210 with hard negatives. Actual false accepts must be read alongside these counts. The complete ranking also favors SLOT on the original accepted-alternative corpus (pooled AUC .803 for strict paths and .851 for index-free paths, versus .655 for STED). Thus STED's observed advantage at the reported 1% operating points does not establish dominance across thresholds or ranking targets.

G Diff-Taxonomy Audit: Empirical Distribution of Pair Variation in the Generation Corpus

We run a rule-based tagger from the submission repository on 129,808 output pairs from the generation corpus (4 models: Claude-3.5-Haiku, Claude-Sonnet-4, GPT-4.1-Mini, Llama-3.3-70B; 2 datasets; 3 temperatures $\{0.0, 0.5, 1.0\}$; 10 random pairs per sample) to characterize how disagreements are tagged across seven variation categories. Categories are non-exclusive and rule-based, not human-validated labels.

Table 22: Rule-based per-category pair frequency on the audited corpus subset. Categories are non-exclusive; "other" is the residual.

Category	Toucan (n=120,720)	ShareGPT (n=9,088)
R1 synonym	5.6%	**64.7%**
R1 format-whitespace/case	0.3%	3.5%
R1 format-numeric	0.02%	0.03%
R1 format-date	0.01%	0.1%
R2 type coercion	0.1%	0.03%
R3 reorder	0.1%	2.9%
R4 case convention	0.2%	0.6%
R4 key rename	0.01%	0.0%
R5 aggregation	1.6%	**69.0%**
R6 nesting	**10.6%**	**61.2%**
R7 compound	6.3%	**69.1%**
other	88.9%	24.1%

Interpretation and label limits. The audit identifies multiple changed leaves and nested differences in real generations. On Toucan, 88.9% of pairs fall into the residual category; a follow-up probe of 960 pairs and 1,763 leaf differences tags 91.8% of those differences as one-sided fields. This motivates comparing presence and value together. The "synonym" category uses MiniLM cosine at least .85, rather than human equivalence judgments; "aggregation" means more than three leaf differences and "nesting" means a changed path deeper than two levels. These rule-based frequencies are descriptive and cannot validate the semantic accuracy of the same embedding kernel.

H Local Sandbox Execution Diagnostic

BFCL v4 supplies schemas and reference bindings for 29 locally implemented pure Python functions; 207 of 208 references execute. Variants cover numeric jitter, formatting/coercion, argument swaps/deletion, callee substitution, sign changes, list edits/permutations and defaults. We exclude 621 convention-dependent rebracketing, key-renaming and wrapper pairs.

Labels and scope. Positives share one observed return: numbers use relative/absolute tolerances 10^{-9}, dictionaries compare by keys and sequences by position; exceptions against retained references count as different. This local target depends on coercion/error handling and the constructed edit mixture, not official BFCL execution or external services. Return equality cannot establish behavioral or intent equivalence: 74 intended-changing edits share a return.

AUC comparisons. Pooled AUC compares 647 positives against 1,105 negatives; 98.6% of comparisons cross operators. Macro AUC averages six operators with both labels. STED leads the compared baselines on pooled AUC (.854) but trails them on macro AUC (.265). Pair-weighted within-operator AUC is .274, so the reversal is not solely equal-operator weighting (Table 23).

I Additional Primary Consistency Results

Table 24 applies the same STED score as the human evaluation to all 18 ShareGPT model–serving configurations. It reports conditional consistency, validity and scoreable-prompt coverage separately.

Coverage example. On the 1,006 Toucan prompts, Llama-3.3-70B has 407 scoreable prompts at $T = 0$ and 475 at $T = 1$. The zero-filled mean rises from .396 to .447, while the scoreable-prompt mean falls from .979 to .946. Validity is .324851 and .325050, respectively. The increase in coverage explains the reversal through $\bar{c}_0 = (N_s/N)\bar{c}_s$; the scoreable prompt sets can also differ in

Table 23: **Agreement with execution outcomes.** Pooled AUC uses 1,752 pairs; macro AUC averages the six edit operators containing both labels (higher is better). STED uses the same primary configuration as the human and corpus evaluations. The labels describe equality of a single observed return in the local sandbox. Bold marks column maxima at the reported precision.

Metric	Pooled AUC	Macro AUC
STED	**.854**	.265
SLOT (strict paths)	.789	.311
SLOT (index-free paths)	.816	.485
TED	.599	.446
DeepDiff	.466	.464
Exact match	.613	**.500**
Jaccard	.654	.430
Levenshtein	.616	.320
ROUGE-L	.531	.324

Table 24: **STED consistency on ShareGPT.** r_v is generation validity and coverage is the percentage of prompts with at least two valid outputs, both at $T = .5$. $\bar{c}_s$ averages conditional pairwise agreement over scoreable prompts. Rows describe the recorded model–serving configurations. Bold marks column maxima at the reported precision.

Model	r_v%	Coverage%	$\bar{c}_s^{T=0}$	$\bar{c}_s^{T=.5}$	$\bar{c}_s^{T=1}$
Claude-Sonnet-4	72.0	93.3	.829	**.734**	.662
Claude-Opus-4	97.2	98.7	.973	**.734**	.658
Claude-3.5-Sonnet	**100.0**	**100.0**	.798	.732	.678
Claude-Opus-4.5	96.4	98.7	.878	.729	**.679**
Claude-3.5-Haiku	97.9	**100.0**	**.997**	.710	.610
Claude-Sonnet-4.5	87.7	98.7	.947	.706	.642
Llama-3.3-70B	97.1	**100.0**	.814	.697	.646
Claude-3.7-Sonnet	**100.0**	**100.0**	.747	.682	.639
Claude-Haiku-4.5	**100.0**	**100.0**	.981	.669	.614
Qwen3-235B-A22B	99.9	**100.0**	.743	.666	.614
Gemini-2.5-Flash-Lite	**100.0**	**100.0**	**.997**	.653	.605
GPT-4.1-Mini	**100.0**	**100.0**	.807	.652	.599
Nova-2-Lite	**100.0**	**100.0**	.738	.648	.598
Grok-4.1-Fast	98.4	**100.0**	.675	.642	.647
Qwen3-32B	97.3	97.3	.795	.620	.545
Mimo-V2-Flash	74.8	76.0	.660	.608	.564
GPT-OSS-120B	**100.0**	**100.0**	.614	.590	.567
Minimax-M2	99.9	**100.0**	.636	.567	.473

composition. This descriptive example does not isolate a causal temperature effect on a fixed set of valid outputs.

Decision scope. In ShareGPT at $T = .5$, Sonnet-4 has conditional consistency .734 with validity 72.0%, whereas 3.5-Sonnet has .732 with validity 100.0%. A consistency ranking alone therefore does not identify the preferred deployment. Task correctness, costs and latency are not measured in this comparison and require separate evaluation.

J SCHEMA-ENFORCEMENT PROTOCOL

The collection contains 30 frozen checkpoints: five models, three temperatures ($T = 0, .5, 1$), and two modes (free-form/enforced), each originally containing 76 prompts and ten runs (22,800 outputs). Eighteen checkpoints cover Qwen3-235B-A22B, GPT-OSS-120B and Minimax-M2; twelve cover Claude-Haiku-4.5 and Claude-Sonnet-4.5. The same STED score serves human validation and both generation analyses.

Common cohort. Both generation analyses use 75 prompts from the full 80-source ShareGPT pool, excluding malformed references at zero-based positions 38, 45, 60, 67 and 68. All enforcement validity, means and intervals use these same 75 identifiers, yielding 22,500 analyzed outputs.

Sources and serving protocols. The Claude checkpoints were collected on 15 April 2026 and store parsed responses, with `None` for invalid/missing outputs; raw HTTP text is unavailable. Each Claude model–temperature comparison has 74 paired scoreable prompts. Claude uses Bedrock InvokeModel schema enforcement; the other models use the Bedrock Mantle OpenAI-compatible interface, with model-specific strictness. Checkpoints retain dated source and serving metadata and are separate from the Jan–Feb generation archive. Between-model differences therefore combine model identity, endpoints, enforcement implementations and collection runs.

Paired estimand. Each free/enforced comparison uses the same prompt intersection for both means and their difference. Validity averages all outputs; conditional agreement requires at least two parseable outputs in each mode. Paired-prompt bootstraps use 10,000 resamples (seed 2026), without adjustment across fifteen comparisons. Intervals describe prompt sampling within each recorded run, excluding between-run generation uncertainty. Conditional consistency gains do not imply correctness or validity gains.

J.1 Provider extraction paths

Several non-Claude Toucan configurations use a unified proxy whose extraction can miss native tool-call fields. A Gemini-2.5-Flash-Lite spot-check at $T = 0$ reruns the 460 prompts for which the proxy recorded no valid output, using a native endpoint with a matched 2,048-token limit. It recovers 636 valid runs across 93 prompts, implying a validity change from 54.2% to 60.5% for that cell. The stored corpus is retained: replacing one cell would mix collection protocols, and the effect has not been characterized across other providers and temperatures. The spot-check illustrates serving-path sensitivity and does not identify a general model capability or the source of all $T = 0$ variation.

K Reproducibility and Release Scope

Artifacts and use. The repository contains evaluator/analysis code, frozen environments, the 456-task annotation records and scores, and labeled pairs. The anonymous archive at `https://huggingface.co/datasets/socb-anon-2441/socb` supplies prompts, stored outputs and validation artifacts. The corpus analysis covers 396 cells, 2,140,380 outputs and 8,436,780 valid pairs. To reproduce a result, use its frozen inputs and the score in App. A; the repository's `implementation_notes/README_generation_records.md` links the recorded environments and analysis instructions.

Parsing and aggregation. Empty/missing Toucan or ShareGPT corpus responses are invalid; enforcement treats any non-`None` parsed value as valid. Use raw-text parsing where available and stored parsed responses for Claude. Preserve duplicate-pair multiplicity in prompt means and the common 75-prompt ShareGPT cohort (App. J). Existing outputs support new scoring analyses; a new model requires new generations under a documented serving protocol.

Analysis protocols. Human denominators and Holm families are specified in App. E; BFCL calibration in App. F; paired enforcement in App. J. Execution AUCs are descriptive. The 316 external model judgments, including swaps, are distinct from human labels; all 378 selected trajectories pass exact official-tool return replay (App. D).

Release limits. Generations support repeated-output and deployment-specific comparisons but have no train/test split or independent pair-equivalence labels; upstream prompts may occur in pretraining. Repeated wrong calls can score highly. Dated metadata identifies recorded configurations; provider access and future checkpoint availability remain outside the release's control.

Prior dissemination. A non-archival workshop version and public preprint (Wang et al., 2025b) introduced STED and the basic consistency framework. This submission adds the fixed-score comparisons, three-annotator evaluation and expanded generation/enforcement analyses.