

Out-of-Distribution Checkpoint Selection Has a Resolution Problem: Auditing Sparse Exact Match with Token Likelihood

Kunhee Ryu Keeheon Lee*

Yonsei University

Seoul, Republic of Korea

rgh00826@yonsei.ac.kr

keeheon@yonsei.ac.kr

Abstract

Exact match (EM) can make checkpoint selection ambiguous when a small labeled out of distribution (OOD) validation set leaves several checkpoints tied at zero or at the maximum. Without an explicit tie rule, the reported metric does not identify which checkpoint produced the final test score. We study this validation resolution problem in semantic parsing and morphological reinflection. Our protocol reports the reference budget, token likelihood choice under teacher forcing, and latest tied candidate, while using EM for a unique nonzero maximum. Across 600 draws from 30 runs, 8 references leave every early checkpoint at zero in 48% of draws and the maximum tied in 57%. Within tied draws, mean full split regret falls from 0.063 under random selection to 0.050 with token likelihood. The latest tied candidate performs better on average for mostly monotone trajectories, but token likelihood independently orders candidates that hard EM cannot distinguish. These results support reporting both orderings when small labeled OOD validation sets guide checkpoint selection.

1 Introduction

Suppose two checkpoints both fail all eight examples in a labeled out of distribution (OOD) validation set. Exact match (EM) assigns both checkpoints a score of zero, even if they achieve different accuracies on the full held out generalization split. A report that says *selected by OOD EM* therefore does not identify whether the earlier checkpoint, the later checkpoint, or another tied candidate produced the reported final OOD score. We call this inability of the validation metric to rank candidate checkpoints a *validation resolution problem*.

Checkpoint selection and final evaluation place different demands on EM. For final evaluation,

EM directly reflects whether a task output is entirely correct. For checkpoint selection on a small validation set, however, its binary scores for each example form a coarse signal. A model can assign increasing probability to correct output tokens before greedy decoding produces an entirely correct sequence. We therefore compare decoded OOD EM, which we call *hard EM*, with the mean log likelihood of the gold OOD tokens under teacher forcing, which we call *token likelihood*. Under teacher forcing, each gold token is scored conditional on the preceding gold tokens, allowing token likelihood to order checkpoints that hard EM ties.

We use these two signals to audit how much resolution a labeled OOD validation set provides for checkpoint selection. For a given labeled OOD reference subset, whose size we call the reference budget, the protocol first asks whether hard EM identifies a unique checkpoint. When hard EM has a unique nonzero maximum, the protocol selects that checkpoint directly. When the maximum is tied, including when every checkpoint scores zero, it reports both the token likelihood tie break and the latest checkpoint among the tied candidates. Each report also states this budget and its sampling scheme, the frequency and size of ties, and the final OOD score separately from the validation statistics. Because the protocol requires gold OOD outputs, its scope is labeled OOD validation with a small budget.

Explicit tie handling matters most when OOD performance improves late in training. We call this pattern a *delayed OOD transition*, in which in distribution accuracy rises early while OOD EM remains near zero until later checkpoints. Across 30 independently seeded training runs, we sample 20 labeled OOD reference subsets, or *draws*, per run. With eight references, every checkpoint in the first 30% of the evaluated trajectory has zero hard EM in 48% of run/draw pairs, while the max-

*Corresponding author.

imum across all candidates is tied in 57%. Generalization exact match (GEN EM) is EM on the full held out generalization split. Selection regret is the gap between the best candidate’s and the selected checkpoint’s GEN EM. Within tied draws, tie breaking by token likelihood reduces mean regret relative to random tie breaking. The latest tied checkpoint nevertheless achieves lower regret on average when GEN EM rises over training. At larger reference budgets from 16 to 64, hard EM more often identifies a unique maximum, reducing the need for an auxiliary tie breaker.

To test the pattern across tasks and model families, we study semantic parsing on COGS (Kim and Linzen, 2020), using standard and higher capacity model settings, and on the SLOG variable free split (Li et al., 2023). English and Turkish reinflection tasks from SIGMORPHON provide contrasting morphology regimes (Cotterell et al., 2017). We also evaluate pretrained T5-small (Raffel et al., 2020) and BART-base (Lewis et al., 2020) models on COGS. Together, these settings reveal the prevalence of hard EM ties, the cases in which token likelihood distinguishes tied checkpoints, and the reference budgets at which hard EM regains sufficient resolution.

Contributions. We make three contributions. First, we quantify how often small labeled OOD validation sets leave ties and how much an unreported tie breaker changes the selected result. Second, we compare hard EM, token likelihood, the latest tied checkpoint, and two decoded partial credit metrics (token F1 and normalized edit similarity) on the same OOD references, accounting for repeated draws within runs and removing selected references in a separate evaluation. Third, we provide a validation resolution protocol that uses hard EM for a unique nonzero maximum.

2 Related Work and Positioning

Compositional generalization is commonly studied with controlled benchmarks for sequences and semantic parsing. The Simplified version of the CommAI Navigation tasks (SCAN) tests recombination of navigation commands (Lake and Baroni, 2018; Loula et al., 2018), and Compositional Freebase Questions (CFQ) tests mappings from questions to logical forms (Keysers et al., 2020). COGS targets semantic interpretation (Kim and Linzen, 2020). ReCOGS exposes sensitivity to incidental logical form details (Wu et al., 2023).

SLOG adds structurally complex semantic parsing cases (Li et al., 2023). Results across this literature depend on split design, target representation, architecture, and dataset artifacts (Oren et al., 2020; Shaw et al., 2021; Csordás et al., 2021; Ontañón et al., 2022; Hupkes et al., 2023). We use COGS and SLOG as temporal testbeds for checkpoint selection. SIGMORPHON reinflection supplies a morphology contrast (Cotterell et al., 2017), and its learning curves are more gradual in our experiments.

Work on training dynamics shows what final accuracy can obscure. Example forgetting and dataset cartography characterize learning from example histories (Toneva et al., 2019; Swayamdipta et al., 2020). Grokking and double descent show that generalization can change after training accuracy saturates or vary in a nonmonotonic way (Power et al., 2021; Liu et al., 2022; Belkin et al., 2019; Nakkiran et al., 2021). Mechanistic studies use internal progress measures to trace how abrupt behaviors emerge (Olsson et al., 2022; Nanda et al., 2023). These findings motivate our study of when token likelihood on labeled OOD references can rank checkpoints that hard OOD EM or in distribution development EM do not distinguish.

Checkpoint selection also occurs under distribution shift. Domain generalization evaluations show that model selection choices can affect conclusions (Gulrajani and Lopez-Paz, 2021). WILDS provides standardized evaluations under real distribution shifts and uses OOD validation data for model selection (Koh et al., 2021). Work on invariant risk and shortcut learning explains why in distribution performance need not predict OOD behavior (Arjovsky et al., 2019; Geirhos et al., 2020). In our setting, checkpoints are scored by hard EM and token likelihood on the same labeled OOD examples. When hard EM ties, the stated selector does not identify a checkpoint. We therefore report the tie break alongside the final OOD score.

Linear probes were introduced as diagnostics of intermediate representations (Alain and Bengio, 2017) and have since been used to test which linguistic properties are recoverable from language representations (Conneau et al., 2018; Hewitt and Manning, 2019; Tenney et al., 2019). Representation similarity methods compare internal states across models or training stages (Raghu et al., 2017; Morcos et al., 2018; Kornblith et al., 2019). Probe results require controls and careful interpre-

tation (Hewitt and Liang, 2019; Belinkov, 2022). Appendix B reports structural probes and nearest neighbor alignment as auxiliary diagnostics. The central checkpoint selection analysis uses OOD token likelihood and jointly reports hard EM resolution, the latest tied checkpoint, and tie breaking by token likelihood.

3 Soft and Hard OOD Checkpoint Signals

OOD token likelihood. Let θ_t denote the parameters of checkpoint t and $\mathcal{D}_{ood}$ the labeled OOD reference pairs (x, y) . Teacher forcing scores each gold token y_i given the input x and the gold prefix $y_{<i}$. Writing $|y|$ for the output length, we first average gold token log probabilities within each output and then across the reference set as follows.

$$\begin{aligned}\ell_t(x, y) &= \frac{1}{|y|} \sum_{i=1}^{|y|} \log p_{\theta_t}(y_i \mid x, y_{<i}), \\ R_t^{\text{tok}} &= \frac{1}{|\mathcal{D}_{ood}|} \sum_{(x,y) \in \mathcal{D}_{ood}} \ell_t(x, y).\end{aligned}\tag{1}$$

This score requires gold OOD outputs and is therefore available only for labeled OOD validation. Final performance remains measured by EM. Token likelihood aggregates changes at individual positions, whereas EM changes only when an entire decoded output becomes correct. A checkpoint can therefore improve many gold token probabilities before greedy decoding produces an exactly correct output.

Selection definitions. Generalization exact match (GEN EM) is greedy decoding EM on the held out generalization split. Development EM is the corresponding in distribution development metric. A *run* is an independently seeded training trajectory. Within a run, a *draw* is a labeled OOD reference subset held fixed across candidate checkpoints. Its *reference budget* is the number of input and output pairs. Draws with tiny budgets come from a fixed OOD reference pool with 2,100 examples. The *analyses with all references* instead score the entire held out generalization split. Selection regret, also called oracle regret, is the gap between the full split GEN EM of the best candidate checkpoint and that of the checkpoint chosen by a selection rule. The *early window* is the first 30% of evaluated checkpoints. An *early zero condition* occurs when every checkpoint in this window has zero hard OOD EM. A *tied hard*

EM maximum occurs when multiple candidate checkpoints attain the same highest hard OOD EM on a draw. We use *hard EM plateau* for the candidate set that hard EM leaves unresolved under either condition.

To compare token likelihood movement across runs, write $L_t = R_t^{\text{tok}}$ and normalize it within each run as $(L_t - L_{\text{first}})/(\max_s L_s - L_{\text{first}})$. If this denominator is zero, we use minimum and maximum scaling. A constant trajectory is set to zero. A *checkpoint distinguished only by token likelihood* has zero hard OOD EM while normalized token likelihood has reached at least 0.10. We call a trajectory *mostly monotone* when GEN EM on the full split generally improves toward later checkpoints despite local reversals. All primary EM values use deterministic greedy decoding. Section 5.1 reports decoded partial credit and beam search controls.

Reporting protocol. Our empirical results motivate the following protocol for labeled OOD validation. Given a fixed set of candidate checkpoints, evaluate every checkpoint on the same labeled OOD reference subset. If hard OOD EM has a unique nonzero maximum, select that checkpoint directly. If the maximum is shared, including when all candidates score zero, report both the latest checkpoint among the tied candidates and the token likelihood tie break defined by Equation 1. This ordering makes the target metric decisive when it resolves the candidates and makes the tie breaking choice explicit when hard EM does not. The resulting audit table accompanies each final OOD score. For each regime, it reports the reference budget and sampling scheme, the early zero and tied maximum rates for hard EM, the unique hard EM choice when one exists, and both tie break choices otherwise. Together, these fields make the selection decision traceable.

4 Experimental Regimes

We choose regimes to cover distinct learning patterns, tasks, and initialization conditions. The baseline COGS regime uses the semantic parsing benchmark introduced by Kim and Linzen (2020). It is our primary setting with a delayed transition, and in our runs its in distribution EM becomes high while OOD EM remains low across a long checkpoint interval. A COGS variant with higher capacity tests whether this pattern depends on model capacity. The SLOG benchmark sup-

plies a second semantic parsing setting (Li et al., 2023). We evaluate its variable free condition, whose target programs contain no explicit variables. The English and Turkish tracks of the CoNLL SIGMORPHON 2017 shared task provide gradual morphology settings (Cotterell et al., 2017).

The main checkpoint selection audit contains 30 semantic parsing runs trained from scratch, comprising 10 COGS baseline runs, 10 COGS capacity runs, and 10 SLOG variable free runs. Models in all three regimes are randomly initialized encoder and decoder Transformers (Vaswani et al., 2017), trained for 60 epochs with batch size 32, and saved every 1,000 steps. Representation diagnostics are computed at every third saved checkpoint. The COGS baseline and SLOG models have three layers each in the encoder and decoder, hidden size 256, four attention heads, and feed-forward size 512. The COGS capacity model increases the encoder and decoder to four layers, the number of heads to eight, and the feed-forward size to 1,024, while retaining hidden size 256. The SIGMORPHON experiments use three seeds per language, two layers each in the encoder and decoder, hidden size 128, four attention heads, feed forward size 256, batch size 32, 100 training epochs, and a 100 step evaluation interval. Appendix Table D.2 gives the capacity configuration, and Appendix Table C.1 summarizes the evidence and scope of each regime.

Figure 1 and Table 1 report the central checkpoint selection audit with 30 runs. Section 5.2 relates this audit to COGS pre-generalization dynamics. Sections 5.5 and 5.6 report the morphology boundary and pretrained comparisons, respectively. The appendix reports results for each setting. Appendix B contains the secondary forecasting analyses.

Pretrained comparisons. To test the protocol beyond random initialization, we fine-tune T5-small (Raffel et al., 2020) and BART-base (Lewis et al., 2020) on COGS for three seeds each. Each run contributes a trajectory of 12 checkpoints, which we analyze with the same hard EM sparsity and tie breaking summary used for the semantic parsing audit with 30 runs. These comparisons test whether the same sparsity and tie breaking patterns persist when checkpoint trajectories begin from pretrained parameters. We report Pythia (Biderman et al., 2023) as an exploratory decoder

only comparison. Appendix Table D.1 records the status and scope of all pretrained protocols.

Reproducibility materials. The public repository at <https://github.com/rgh112/oood-checkpoint-selection-resolution> provides configurations, training and analysis code, and the CSV and JSON artifacts used for the reported results. Its scripts reproduce the semantic parsing cohort with 30 runs and the corresponding summaries. The runs with densely saved checkpoints used for the COGS and SIGMORPHON analyses before generalization were executed on one NVIDIA RTX A5000 GPU with 24 GB of video memory. Because hardware was not controlled across all runs, the reproducibility record focuses on model configurations, seeds, checkpoint cadence, and executable scripts.

5 Results

We first quantify how often hard OOD EM fails to distinguish checkpoints. Within those ties, we compare token likelihood with alternative tie breakers and then identify the settings in which the reporting protocol applies. Figure 1 gives the visual overview, and Table 1 reports the plateau cases used to evaluate the tie breakers.

5.1 Token Likelihood Ranks Inside Hard EM Plateaus

The central experiment compares selection by hard OOD exact match and soft OOD token likelihood on the same labeled OOD reference subsets. Across the 600 combinations of runs and draws with eight references, hard OOD EM is sparse. The early window remains at zero in 48%, and the maximum across checkpoints is tied in 57%. Under either condition, hard EM leaves multiple checkpoints unresolved and therefore requires a tie breaker. Across the same combinations, an average of 6.7 checkpoints are distinguished only by token likelihood because hard OOD EM remains zero while normalized token likelihood has moved by at least 10% of its range within a run.

Clustered uncertainty and reference leakage checks. The 20 draws within each run are correlated, so we use a hierarchical bootstrap.

Table 1 reports intervals that resample runs and draws within runs. For the central tied and early zero cases with 8 references, the differences in regret between soft and random selection remain

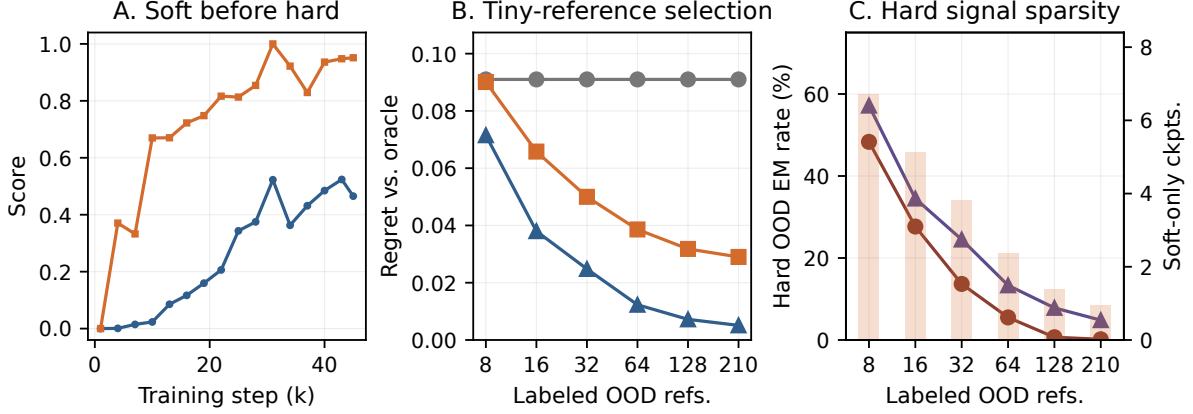


Figure 1: Soft signals before hard OOD generalization. The left panel shows normalized OOD token log probability (orange squares) rising while exact match GEN EM (blue circles) remains near zero in a representative COGS baseline trajectory. The middle panel shows checkpoint selection regret with very small labeled OOD budgets, where development EM is gray, hard OOD EM is blue, and soft OOD token likelihood is orange. The right panel shows that hard exact match is frequently zero throughout the early window (brown) or tied (purple) at tiny budgets, whereas checkpoints distinguished only by token likelihood remain visible as orange bars.

negative under this bootstrap ($[-0.022, -0.005]$ and $[-0.029, -0.004]$, respectively). We also check whether including the sampled references in the evaluation pool favors token likelihood. Removing them from the OOD reference pool evaluation slice with 2,100 examples leaves the corresponding deltas nearly unchanged at -0.012 and -0.015 .

Decoded partial credit and beam controls. We compare two decoded partial credit signals within the same plateaus formed with 8 references. Token F1 measures overlap between predicted and gold whitespace token multisets with multiplicity. Normalized token edit similarity applies Levenshtein distance (Levenshtein, 1966) to tokens, divides it by the longer sequence length, and subtracts the result from one. Table 2 reports the greedy results and a beam search spot check with beam size 4. Across all 30 runs, both decoded signals have slightly higher mean regret than random tie breaking, and both hierarchical intervals for the regret difference relative to random include zero. In the check across three COGS seeds, beam search with size 4 changes the selected checkpoint relative to greedy decoding in 26% of token F1 cases and 24% of edit similarity cases. The corresponding intervals for the regret difference relative to random also include zero. Among the partial credit signals tested with 8 references, token likelihood under teacher forcing is the only one whose hierarchical interval supports lower regret than random

tie breaking.

Comparison with the latest tied checkpoint.

The latest tied checkpoint is a strong baseline because many trajectories are mostly monotone. Averaged over tied draws with 8 references, it has lower regret than token likelihood (0.037 vs. 0.050). Accordingly, the reporting protocol includes both latest and token likelihood. The two tie breakers diverge in difficult cases. In 120/343 tied draws with 8 references, the latest checkpoint has at least 0.05 GEN EM regret, and within that subset token likelihood reduces average regret from 0.088 to 0.061 and beats the latest checkpoint in 47% of draws (Appendix Table A.3). Latest is stronger on average, while token likelihood supplies an independent dense ordering in the minority of hard EM plateaus where the latest checkpoint remains suboptimal. The modest absolute regret reductions quantify the practical role of the audit by revealing when a reported OOD score depends on an unresolved selection choice.

Across budgets, Table 1 shows the complete phase pattern. With 8 references, token likelihood gives a statistically stable improvement over random hard EM tie breaking. With 16 to 64 references, the advantage narrows as hard EM becomes less sparse. The latest tied checkpoint remains the strongest average heuristic in the mostly monotone trajectories studied here. Appendix Tables A.1 and A.2 provide the full budget grid and a complete conditional audit. The larger refer-

Table 1: Tie break regret when hard OOD EM lacks resolution. Soft is token likelihood under teacher forcing restricted to the same hard EM plateau. Random averages uniformly over tied hard EM checkpoints. Latest picks the latest tied checkpoint. n counts combinations of runs and draws on a plateau, with the number of distinct runs in parentheses. Δ is Soft minus Random, so negative values favor token likelihood. Confidence intervals resample runs and then reference draws within runs. LRO removes the selected tiny reference set from the 2,100 example OOD reference pool evaluation slice. With 8 references, token likelihood improves over random tie breaking, while latest is the stronger average heuristic in these mostly monotone trajectories.

Refs	Plateau condition	n (runs)	Soft	Random	Δ cluster CI		LRO Δ CI	Latest
8	tied hard EM max	343 (30)	0.050	0.063	-0.013	$[-0.022, -0.005]$	-0.012 $[-0.021, -0.004]$	0.037
8	early zero	161 (29)	0.044	0.059	-0.015	$[-0.029, -0.004]$	-0.015 $[-0.028, -0.003]$	0.035
16	tied hard EM max	207 (30)	0.038	0.045	-0.007	$[-0.016, +0.001]$	-0.007 $[-0.015, +0.001]$	0.030
32	tied hard EM max	147 (29)	0.032	0.035	-0.002	$[-0.010, +0.005]$	-0.002 $[-0.010, +0.006]$	0.024
64	tied hard EM max	80 (27)	0.022	0.026	-0.005	$[-0.013, +0.004]$	-0.005 $[-0.013, +0.004]$	0.023

Table 2: Decoded partial credit tie breaks for the plateaus with 8 references and a tied maximum in Table 1. Token F1 measures whitespace token overlap with multiplicity, and edit similarity is one minus token level Levenshtein distance divided by the longer sequence length. The beam search uses size 4 and length penalty 1.0 in a spot check on COGS seeds 1, 2, and 3. n counts combinations of runs and draws that meet the plateau condition. Δ is the tie break regret minus random regret, so negative values favor the decoded signal. Confidence intervals come from a hierarchical bootstrap that resamples runs and then draws within runs. Lower regret is better.

Scope	Decode	Tie break	n	Regret	Random	Δ (95% CI)
All 30 runs	Greedy	Token F1	343	0.066	0.063	+0.004 $[-0.004, 0.011]$
All 30 runs	Greedy	Edit similarity	343	0.066	0.063	+0.003 $[-0.004, 0.011]$
COGS seeds 1 to 3	Beam 4	Token F1	38	0.044	0.039	+0.005 $[-0.009, 0.020]$
COGS seeds 1 to 3	Beam 4	Edit similarity	38	0.043	0.039	+0.004 $[-0.011, 0.017]$

ence scaling analysis from 1% to 100% is reported in Appendix Table A.5. Together, these results place token likelihood within the reporting protocol by supplying a dense ordering while hard exact match remains uninformative, alongside recency and hard EM selection as resolution increases.

Task specificity control. On three COGS baseline seeds, we compare the gold OOD targets with permuted and mismatched OOD targets at early, middle, and final checkpoints. The mean gap between gold and permuted targets grows from 2.598 at step 7,000 to 3.251 at step 45,000. The gap between gold and mismatched targets similarly grows from 2.601 to 3.254. The widening gaps show that Equation 1 becomes increasingly specific to the outputs structured for the correct task over training. Because this control measures association rather than a causal intervention, it does not identify the underlying mechanism.

5.2 Relation to Dynamics before Generalization

COGS exhibits a window before generalization in which soft OOD likelihood changes before OOD exact match. Across all seeds, exact match on the in distribution development and test sets rises first,

leaving GEN EM low while token likelihood continues to move. The timing and future gain correlations vary substantially across seeds, but the validation consequence is consistent. Token likelihood separates checkpoints that hard EM still ties.

5.3 What Does Token Likelihood Track?

To characterize what token likelihood tracks, we separate structural and lexical generalization outcomes. For the COGS style splits, *structural GEN EM* covers three categories comprising CP recursion (nested sentential complements), PP recursion (nested prepositional phrases), and transfer from object PP to subject PP. *Lexical GEN EM* covers all remaining labeled generalization cases. Across the 20 COGS runs (10 baseline and 10 capacity), the token likelihood slope in the early window aligns more strongly with lexical than structural GEN EM. In the capacity regime, it correlates with maximum lexical GEN EM at Spearman +0.879 and with maximum structural GEN EM at -0.188. In the baseline regime, the corresponding values are +0.273 and -0.091.

The temporal pattern is consistent with this interpretation. In both the COGS baseline and the COGS capacity perturbation, token likelihood

first crosses a $+0.10$ delta from its initial value a median of 7,000 training steps before lexical GEN EM crosses 0.10. On the SLOG variable free split, where some structural generalization occurs, the same token likelihood feature also correlates positively with maximum structural GEN EM ($+0.359$). These results show that token likelihood tracks increasing compatibility with the OOD outputs, while the form of generalization associated with that movement depends on the regime.

5.4 Checkpoint Selection Beyond the Window with Few References

The selection rule should change as the labeled OOD reference budget grows. Table 3 quantifies this transition. With 8 references, hard EM remains zero throughout the early window for 48% of combinations of runs and draws, and its overall maximum is tied for 57%. An average of 6.7 checkpoints have zero hard EM even after token likelihood has moved by at least 10% of its range within a run. At 64 references, the corresponding resolution failures fall to 6% and 13%, and at 210 references they fall to 0% and 5%. Over the same range, hard EM selection regret decreases from 0.072 to 0.005, whereas token likelihood regret decreases from 0.090 to 0.029. As the reference set resolves more checkpoint differences, the target hard metric becomes the better unconditional selector, while token likelihood remains available inside unresolved plateaus. Reference budget and split hardness jointly determine the failure mode. T5-small retains the plateau pattern at a tiny budget, while BART-base remains near zero exact match despite improving likelihood (Section 5.6). This regime with a small budget can arise during pilot benchmark development, in newly annotated shift domains, and on hard structural tasks or tasks with program equivalence wherever only a small labeled OOD set is available. The protocol assumes labeled OOD validation references. Once a representative reference set resolves checkpoint differences, selection returns to the target hard metric.

Token likelihood using all references remains useful in these regimes as a practical alternative to selection by in distribution development EM. Across the semantic parsing cohort with 30 runs, replacing development EM selection with Equation 1 reduces oracle regret from 0.068 to 0.025 on the COGS baseline, from 0.065 to 0.013 on

Table 3: Resolution handover as the labeled OOD reference budget grows. Each row pools 30 runs and 20 deterministic subset draws per run. Regret is measured against the best evaluated checkpoint in the run. Lower regret is better.

OOD refs.	Early zero	Maximum tied	Hard EM regret	Token likelihood regret
8	48%	57%	0.072	0.090
16	28%	35%	0.038	0.066
32	14%	25%	0.025	0.050
64	6%	13%	0.012	0.039
128	1%	8%	0.007	0.032
210	0%	5%	0.005	0.029

the COGS capacity perturbation, and from 0.140 to 0.021 on the SLOG variable free split. Pooled across runs, the regret reduction is 0.071 GEN EM (95% bootstrap CI [0.041, 0.102]), with 22 improved, 4 tied, and 4 worsened runs. Appendix Table A.4 reports the paired breakdown by run. To distinguish OOD specific information from likelihood based checkpointing generally, we add a baseline using token likelihood on the in domain development set. Selecting by token likelihood under teacher forcing on the development split has pooled regret 0.097, compared with 0.091 for development EM and 0.020 for OOD token likelihood (Appendix Table A.4). Its regret is 0.050 on the capacity perturbation, 0.095 on the COGS baseline, and 0.146 on SLOG variable free. Only the capacity result improves on its development EM counterpart of 0.065. This contrast is consistent with the useful dense signal deriving from the labeled OOD references.

The comparison using all references evaluates OOD token likelihood against standard in distribution development EM selection, whereas Table 1 compares tie breakers using the same OOD references. In the full-reference comparison, soft OOD likelihood outperforms dev-EM selection in the studied regimes. When the same OOD references yield informative exact matches, hard OOD EM becomes the stronger unconditional selector. The latest checkpoint also remains a strong tie breaker for mostly monotone trajectories. The resulting hierarchy is explicit. Report token likelihood as a dense plateau signal, report the latest checkpoint as the monotone tie breaker, and use hard OOD EM when it supplies a reliable ranking.

5.5 SIGMORPHON as a Gradual Boundary Case

The SIGMORPHON experiments test whether the same relation with positive lead time appears in gradual morphology. For each aligned held out subset, we correlate GEN EM gain three checkpoints later with a composite of OOD token likelihood, probe accuracy, and nearest neighbor alignment normalized within each run. Every English and Turkish subset has a negative mean correlation across three seeds. Appendix Table D.4 reports the individual results. These gradual trajectories offer little sustained interval before the transition for the composite used here, although other soft diagnostics may suit this learning pattern.

5.6 Pretrained Encoder and Decoder Comparison

The pretrained comparison evaluates whether the plateau pattern persists after initialization from T5-small and BART-base. We fine-tune each model on COGS for three seeds, recording 12 checkpoints per run, generated outputs, and teacher-forced OOD token likelihood. Before exact-match scoring, we canonicalize COGS punctuation because T5 often decodes `teleport . agent` where the benchmark target writes `teleport . agent`. Canonicalization removes only differences in punctuation spacing. It does not reorder conjuncts or rename variables.

Under this canonical evaluation, T5-small shows the same qualitative pattern as the scratch models. Across three seeds, canonical max GEN EM is 0.361 ± 0.006 , final GEN EM is 0.361 ± 0.006 , and selecting by OOD token likelihood reaches 0.353 ± 0.010 GEN EM with mean regret 0.008. The same phase pattern appears with a tiny budget. With 8 references, the early hard EM window is zero throughout in 0.317 ± 0.153 of draws and the hard EM maximum is tied in 0.633 ± 0.058 of draws. BART-base provides a contrasting result in which canonical maximum GEN EM remains 0.004 ± 0.001 even though OOD token likelihood improves early by 0.114 ± 0.021 . Together, T5-small and BART-base establish a mixed pretrained boundary. T5-small preserves the plateau behavior after canonicalization, whereas the likelihood movement of BART-base does not translate into correct OOD outputs. Because the models differ in architecture and output formatting, we treat this comparison as evi-

dence of external validity for these specific adaptation settings. Appendix Table D.3 summarizes the comparison.

6 Discussion

The main result is that a hard EM plateau leaves checkpoint identity unresolved even when its validation score is reported. Such plateaus call for an auditable tie break rather than an implicit choice. Once hard exact match becomes informative, direct selection by hard OOD EM becomes appropriate for the reported exact match metric. The token likelihood branch of this rule requires labeled OOD references.

Practical protocol. For a labeled OOD reference set S and evaluated checkpoints $\mathcal{T}$, let H_t be hard OOD EM on S , let $L_t = R_t^{\text{tok}}$ from Equation 1, and let $\mathcal{A} = \arg \max_{t \in \mathcal{T}} H_t$. If no labeled OOD references are available, Equation 1 is unavailable. If $\max_t H_t = 0$, hard EM cannot rank checkpoints. Report both $\arg \max_t L_t$ and the latest checkpoint. Otherwise, if $|\mathcal{A}| > 1$, restrict attention to $\mathcal{A}$ and report both the token likelihood tie break $\arg \max_{t \in \mathcal{A}} L_t$ and the latest checkpoint in $\mathcal{A}$. If $|\mathcal{A}| = 1$ and $\max_t H_t > 0$, select by hard OOD EM directly, while still reporting the reference budget and sampling uncertainty for tiny reference sets.

Practitioner checklist. A minimal checkpoint selection report should contain the following information.

1. state the labeled OOD reference budget, sampling scheme, and whether the same references were used for every checkpoint.
2. quantify hard EM resolution with the early zero rate, tied maximum rate, and maximum tie size.
3. when hard EM is zero throughout or tied, report both the latest tied checkpoint and token likelihood choices, including checkpoint steps and validation scores.
4. when hard EM has a unique nonzero maximum, select by hard EM and state that selection uses the target metric.
5. report test EM on the full split separately from every audit quantity on the validation set.

In the concrete record, hard EM ties three checkpoints. The latest checkpoint and token likelihood both choose 45k, whereas the two decoded signals choose different plateau members. Reporting the full split values separately exposes the conse-

Table 4: Audit record combining resolution across the cohort with one traceable COGS baseline example (seed 1, draw 1). Resolution rates use all 600 combinations of runs and draws. The concrete row uses the same 8 labeled OOD references for every checkpoint.

Audit field	Example entry
Cohort sampling	30 runs with 20 uniform draws per run and 8 references per draw
Cohort resolution	early zero rate 48%, maximum tied rate 57%, mean maximum multiplicity 2.4
Example identity	COGS baseline, seed 1, draw 1
Hard EM tie	steps 31k, 43k, and 45k with validation EM = 0.625 for all three
Latest / likelihood	latest at 45k (EM 0.625), likelihood at 45k (mean log probability -0.210)
Decoded metrics	token F1 at 43k (0.951), edit similarity at 31k (0.905)
Final GEN EM	31k = 0.523, 43k = 0.524, 45k = 0.465

quence without using test performance to choose the checkpoint. Checkpoint identities necessarily depend on the run, while the cohort rates show how often such a record is needed.

Interpretation of the evidence. The central evidence compares checkpoint resolution under identical labeled OOD references. Table 1 isolates the relevant cases in which hard EM is tied or zero throughout, while token likelihood lowers regret relative to random tie breaking inside the plateau. The latest tied checkpoint remains stronger in many monotone trajectories, so the protocol reports both orderings. Token likelihood supplies an independent ordering conditioned on the references when exact match lacks resolution. The selection and forecasting analyses with all references show how this ranking result changes across budgets and regimes.

Scope and reporting recommendation. The comparison settings define the protocol’s empirical scope. SIGMORPHON shows that gradual morphology need not exhibit the positive lead time relation found in COGS. Forecasting normalized by regime varies across settings and limits the pooled correlation across 30 runs. BART-base shows that token likelihood can move even when pretrained adaptation produces almost no correct OOD outputs. For benchmarks with delayed transitions, we recommend reporting checkpoint selection evidence separately from final OOD accuracy. The report includes the reference budget and sampling scheme, early zero rate, tied maximum

rate, hard EM selector on the same references, latest tied checkpoint, and token likelihood selector. This convention separates final generalization quality from the ability of the validation metric to identify a checkpoint. Equation 1 measures compatibility with labeled OOD outputs. Its relation to particular subskills and causal mechanisms remains unspecified. The claim is specific to validation based on exact match. Genuinely graded target metrics may already provide sufficient checkpoint resolution, and how the protocol changes for such metrics remains open.

7 Conclusion

This paper argues that sparse OOD EM creates a resolution problem for checkpoint selection. Labeled OOD token likelihood is useful when hard EM cannot distinguish checkpoints. With 8 OOD references, hard EM is often zero throughout or tied, and token likelihood improves over random tie breaking inside those plateaus under intervals clustered by run and checks that exclude selected references. The latest tied checkpoint is stronger on average in the mostly monotone trajectories studied here, so reporting both orderings reveals when they agree or diverge. With larger reference budgets, hard EM gains enough resolution and should become the direct selector. Mixed results from SIGMORPHON, BART base, and forecasting normalized by regime, together with the lexical and structural breakdown, define the scope as labeled OOD checkpoint selection when exact match is too sparse to reveal partial progress. The practical report is small. It states whether hard EM had a unique nonzero maximum on the labeled OOD references, lists the latest and token likelihood choices when the maximum was tied or zero throughout, and keeps final test EM separate from the selection audit. This report distinguishes model performance from the validation rule used to select the reported checkpoint. A checkpoint selected by OOD EM should therefore be accompanied by evidence that OOD EM had enough resolution to select it.

8 Limitations

The diagnostic in Equation 1 requires gold OOD reference outputs under teacher forcing and therefore applies to labeled OOD validation. It cannot be computed in an unlabeled deployment setting. The reference set must represent the target

shift. Otherwise, hard EM, the latest tied checkpoint, and token likelihood can all select for the wrong subproblem. Uniform sampling, stratified sampling, and sampling targeted by phenomenon can answer different questions, and the sampling scheme should be reported with the reference budget. Likelihood under teacher forcing is also sensitive to output format and canonicalization, as the T5-small bridge illustrates. Our beam search control is a spot check across three seeds rather than a full decoding study, and we do not evaluate smoothed hard EM, entropy or margin selectors, or likelihood with mixed prefixes. The dense checkpoint requirement confines the protocol to controlled training studies and labeled OOD validation analyses. Arbitrary pretrained systems without comparable logs fall outside the evaluated scope. Structural generalization also remains difficult across the studied settings. We exclude CFQ because we could not establish a reproducible protocol that passed basic validation checks.

9 Ethical Considerations

This work uses public benchmark datasets and synthetic examples or examples provided by benchmarks for semantic parsing and morphological reinflection. It involves no data from human subjects, private information, or sensitive annotation. The main interpretive risk is applying a diagnostic based on labeled OOD validation as though it were an unlabeled deployment detector or evidence of solved structural generalization. We therefore state the label requirement, report contrasting results, and treat pretrained findings as exploratory until they undergo the same resolution, leakage, and uncertainty checks as the primary evidence.

Acknowledgments

This research was supported by the Basic Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education, Science and Technology (2710108964). It was also supported in part by a grant from the Institute for AI and Social Innovation at Yonsei University, by the Yonsei University Office of Research Affairs through the Yonsei T.R.U.S.T. (Yonsei Transdisciplinary Research Union for a Sustainable Tomorrow) program (Project Y, project no. 2025-22-0461), and by the Institute of Convergence Science (ICONS), Yonsei University.

AI assistants supported editorial revisions, formatting and checklist preparation, and coding and debugging. The authors reviewed and verified all text, code, analyses, and scientific claims.

References

- Guillaume Alain and Yoshua Bengio. 2017. [Understanding intermediate layers using linear classifier probes](#). In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24–26, 2017, Workshop Track Proceedings*. OpenReview.net.
- Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. [Invariant risk minimization](#). *arXiv preprint arXiv:1907.02893*.
- Yonatan Belinkov. 2022. [Probing classifiers: Promises, shortcomings, and advances](#). *Computational Linguistics*, 48(1):207–219.
- Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. 2019. [Reconciling modern machine-learning practice and the classical bias–variance trade-off](#). *Proceedings of the National Academy of Sciences*, 116(32):15849–15854.
- Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O’Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. [Pythia: A suite for analyzing large language models across training and scaling](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 2397–2430. PMLR.
- Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. [What you can cram into a single \$\\$ \& \! \#^*\$ vector: Probing sentence embeddings for linguistic properties](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.
- Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2017. [CoNLL-SIGMORPHON 2017 shared task: Universal morphological reinflection in 52 languages](#). In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 1–30, Vancouver. Association for Computational Linguistics.
- Róbert Csordás, Kazuki Irie, and Juergen Schmidhuber. 2021. [The devil is in the detail: Simple tricks improve systematic generalization of transformers](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages

- 619–634, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. [Shortcut learning in deep neural networks](#). *Nature Machine Intelligence*, 2(11):665–673.
- Ishaan Gulrajani and David Lopez-Paz. 2021. [In search of lost domain generalization](#). In *International Conference on Learning Representations*.
- John Hewitt and Percy Liang. 2019. [Designing and interpreting probes with control tasks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.
- John Hewitt and Christopher D. Manning. 2019. [A structural probe for finding syntax in word representations](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.
- Dieuwke Hupkes, Mario Giulianelli, Verna Dankers, Mikel Artetxe, Yanai Elazar, Tiago Pimentel, Christos Christodoulopoulos, Karim Lasri, Naomi Saphra, Arabella Sinclair, Dennis Ulmer, Florian Schottmann, Khuyagbaatar Batsuren, Kaiser Sun, Koustuv Sinha, Leila Khalatbari, Maria Ryskina, Rita Frieske, Ryan Cotterell, and Zhijing Jin. 2023. [A taxonomy and review of generalization research in NLP](#). *Nature Machine Intelligence*, 5(10):1161–1174.
- Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, Dmitry Tsarkov, Xiao Wang, Marc van Zee, and Olivier Bousquet. 2020. [Measuring compositional generalization: A comprehensive method on realistic data](#). In *International Conference on Learning Representations*.
- Najoung Kim and Tal Linzen. 2020. [COGS: A compositional generalization challenge based on semantic interpretation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9087–9105, Online. Association for Computational Linguistics.
- Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanus Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M. Beery, Jure Leskovec, Anshul Kundaje, and 4 others. 2021. [WILDS: A benchmark of in-the-wild distribution shifts](#). In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5637–5664. PMLR.
- Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. [Similarity of neural network representations revisited](#). In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3519–3529. PMLR.
- Brenden M. Lake and Marco Baroni. 2018. [Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks](#). In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2873–2882. PMLR.
- Vladimir I. Levenshtein. 1966. [Binary codes capable of correcting deletions, insertions, and reversals](#). *Soviet Physics Doklady*, 10(8):707–710.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. [BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.
- Bingzhi Li, Lucia Donatelli, Alexander Koller, Tal Linzen, Yuekun Yao, and Najoung Kim. 2023. [SLOG: A structural generalization benchmark for semantic parsing](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3213–3232, Singapore. Association for Computational Linguistics.
- Ziming Liu, Ouail Kitouni, Niklas S. Nolte, Eric J. Michaud, Max Tegmark, and Mike Williams. 2022. [Towards understanding grokking: An effective theory of representation learning](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 34651–34663.
- João Loula, Marco Baroni, and Brenden Lake. 2018. [Rearranging the familiar: Testing compositional generalization in recurrent networks](#). In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 108–114, Brussels, Belgium. Association for Computational Linguistics.
- Ari S. Morcos, Maithra Raghu, and Samy Bengio. 2018. [Insights on representational similarity in neural networks with canonical correlation](#). In *Advances in Neural Information Processing Systems*, volume 31.
- Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. 2021. [Deep](#)

- double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003.
- Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. 2023. [Progress measures for grokking via mechanistic interpretability](#). In *International Conference on Learning Representations*.
- Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, and 7 others. 2022. [In-context learning and induction heads](#). Transformer Circuits Thread.
- Santiago Ontañón, Joshua Ainslie, Zachary Fisher, and Václav Cvicek. 2022. [Making transformers solve compositional tasks](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3591–3607, Dublin, Ireland. Association for Computational Linguistics.
- Inbar Oren, Jonathan Herzig, Nitish Gupta, Matt Gardner, and Jonathan Berant. 2020. [Improving compositional generalization in semantic parsing](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2482–2495, Online. Association for Computational Linguistics.
- Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. 2021. [Grokking: Generalization beyond overfitting on small algorithmic datasets](#). In *1st Mathematical Reasoning in General Artificial Intelligence Workshop, ICLR 2021*.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *Journal of Machine Learning Research*, 21(140):1–67.
- Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. 2017. [SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability](#). In *Advances in Neural Information Processing Systems*, volume 30.
- Peter Shaw, Ming-Wei Chang, Panupong Pasupat, and Kristina Toutanova. 2021. [Compositional generalization and natural language variation: Can a semantic parsing approach handle both?](#) In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 922–938, Online. Association for Computational Linguistics.
- Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. [Dataset cartography: Mapping and diagnosing datasets with training dynamics](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.
- Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019. [What do you learn from context? probing for sentence structure in contextualized word representations](#). In *International Conference on Learning Representations*.
- Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. 2019. [An empirical study of example forgetting during deep neural network learning](#). In *International Conference on Learning Representations*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems*, volume 30.
- Zhengxuan Wu, Christopher D. Manning, and Christopher Potts. 2023. [ReCOGS: How incidental details of a logical form overshadow an evaluation of semantic interpretation](#). *Transactions of the Association for Computational Linguistics*, 11:1719–1733.

A Selection-Protocol Audit

This appendix follows the evidence order used in the main paper. It first presents the hard EM sparsity analysis with few references, then conditional tie breaking inside hard EM plateaus, then paired selection statistics with all references, and finally the larger scaling study with small reference sets. The proposed reporting protocol is auditable because a reader can check whether hard exact match had enough resolution before interpreting a selected OOD checkpoint.

Table A.1: Ultra-small labeled-OOD reference budget analysis. Each row aggregates 30 runs $\times$ 20 same-reference subset draws. Hard OOD exact match and OOD token likelihood selectors use exactly the same labeled OOD examples at each budget. Selected checkpoints are evaluated by the full split GEN EM recorded for each run. “Early hard all-zero” is the fraction of run–draw pairs in which hard OOD EM is zero throughout the first 30% of evaluated checkpoints. “Hard maximum tied” is the fraction whose hard OOD exact-match maximum is shared by multiple checkpoints. “Token-likelihood only ckpts.” counts checkpoints where hard OOD EM is zero while normalized token likelihood has moved by at least 10% of its within-run range.

OOD refs.	Dev regret	Hard OOD exact-match regret	OOD token-likelihood regret	Early hard all-zero	Hard exact-match maximum tied	Token-likelihood only ckpts.
8	0.091	0.072	0.090	48%	57%	6.7
16	0.091	0.038	0.066	28%	35%	5.1
32	0.091	0.025	0.050	14%	25%	3.8
64	0.091	0.012	0.039	6%	13%	2.4
128	0.091	0.007	0.032	1%	8%	1.4
210	0.091	0.005	0.029	0%	5%	0.9

Table A.2: Conditional tie-break analysis for small labeled-OOD references. Rows are restricted to run–draw pairs where hard OOD exact match is tied at its maximum, or where early hard OOD exact match is all-zero. “Soft” selects, among hard exact-match tied checkpoints, the checkpoint with highest OOD token likelihood on the same references. “Random” is the expected regret from uniformly selecting among the tied hard exact-match checkpoints. “Latest” selects the latest tied checkpoint, which is a strong heuristic in mostly monotone trajectories, while “Dev” breaks ties by in distribution development exact match. Here n counts eligible run–draw pairs, lower regret is better, and negative Soft-minus-baseline differences favor Soft. The soft improvement over random tie-breaking is clearest at 8–16 references, while Latest has lower average regret in these mostly monotone trajectories. The final column records the nonzero subset in which Soft beats Latest, motivating the protocol’s joint report of both tie-breakers. The intervals in this table treat eligible combinations of runs and draws as rows. Table 1 reports the hierarchical intervals clustered by run that support the main claim.

OOD refs.	Condition	n	Soft	Random	Latest	Dev	Soft–random (95% CI)	Soft–latest (95% CI)	Soft beats latest
8	hard exact-match maximum tied	343	0.050	0.063	0.037	0.052	−0.013 [−0.018, −0.008]	0.012 [0.006, 0.019]	21.9%
8	early hard exact-match all-zero	161	0.044	0.059	0.035	0.048	−0.015 [−0.023, −0.008]	0.009 [0.000, 0.018]	19.9%
16	hard exact-match maximum tied	207	0.038	0.045	0.030	0.040	−0.007 [−0.013, −0.002]	0.008 [0.001, 0.015]	20.8%
16	early hard exact-match all-zero	45	0.044	0.053	0.035	0.045	−0.009 [−0.022, 0.004]	0.009 [−0.003, 0.021]	15.6%
32	hard exact-match maximum tied	147	0.032	0.035	0.024	0.029	−0.002 [−0.008, 0.003]	0.009 [0.002, 0.016]	16.3%
64	hard exact-match maximum tied	80	0.022	0.026	0.023	0.024	−0.005 [−0.010, 0.001]	−0.001 [−0.010, 0.007]	30.0%

Table A.3: Soft-vs-latest failure-case analysis. Rows condition on the same hard-EM plateaus as Table A.2, then restrict to the subset where latest-among-ties has at least 0.05 GEN EM regret against the run oracle. Within these subsets, token likelihood has lower mean regret in every listed row, motivating its joint report with Latest. Rows with fewer than 10 latest-failure draws are omitted.

OOD refs.	Plateau condition	All tied draws	Latest-failure draws	Soft regret in failures	Latest regret in failures	Soft beats latest
8	hard exact-match maximum tied	343	120	0.061	0.088	46.7%
8	early hard exact-match all-zero	161	52	0.064	0.091	48.1%
16	hard exact-match maximum tied	207	55	0.056	0.080	45.5%
16	early hard exact-match all-zero	45	17	0.073	0.077	29.4%
32	hard exact-match maximum tied	147	28	0.050	0.074	42.9%
64	hard exact-match maximum tied	80	12	0.042	0.085	50.0%

Table A.4: Full-reference checkpoint-selection baselines. The OOD token-likelihood selector improves over standard dev-EM selection in 22/30 runs (pooled sign-test $p = 0.00053$ after excluding ties). The in-domain dev token-likelihood baseline tests whether the gain comes from labeled OOD references or from likelihood-based checkpointing generally.

Scope	Dev-EM regret	Dev token-likelihood regret	OOD token-likelihood regret	OOD-vs-dev reduction (95% CI)	Improved / tied / worsened
Pooled	0.091	0.097	0.020	0.071 [0.041, 0.102]	22 / 4 / 4
COGS baseline	0.068	0.095	0.025	0.043 [−0.002, 0.090]	6 / 2 / 2
COGS capacity	0.065	0.050	0.013	0.052 [0.018, 0.087]	8 / 1 / 1
SLOG variable-free	0.140	0.146	0.021	0.119 [0.057, 0.174]	8 / 1 / 1

Table A.5: Small labeled-OOD reference ablation for checkpoint selection. Dev-EM denotes exact-match selection on the in-distribution development split. For each run, we sample a fraction of generalization-split references, compute mean teacher-forced OOD token log-probability on that subset at each checkpoint, and select the checkpoint with the highest subset mean. Fractions below 100% use 20 deterministic draws per run. Regrets are measured against the oracle generalization exact match for each run over the evaluated checkpoints. Checkpoint selection gains appear with subsets as small as 1%, while every setting still uses labeled OOD outputs.

Scope	Labeled OOD refs.	Dev-EM regret	R_t^{tok} regret	Mean-regret reduction	COGS baseline regret	COGS capacity / SLOG regret
Pooled	1% (~ 210)	0.091	0.031 ± 0.047	66%	0.025	0.043 / 0.027
Pooled	5% ($\sim 1,050$)	0.091	0.025 ± 0.039	73%	0.024	0.024 / 0.026
Pooled	10% ($\sim 2,100$)	0.091	0.022 ± 0.034	76%	0.025	0.018 / 0.023
Pooled	25% ($\sim 5,250$)	0.091	0.021 ± 0.031	77%	0.025	0.013 / 0.024
Pooled	100% ($\sim 21,000$)	0.091	0.020 ± 0.031	78%	0.025	0.013 / 0.021

B Secondary Pre-Generalization and Forecasting Analyses

These analyses examine whether soft diagnostics precede or forecast later OOD gains. Their predictive strength varies across seeds and experimental settings, so we use them as secondary evidence for the validation-resolution audit.

B.1 Diagnostic Definitions

Let A_t denote GEN EM at checkpoint t . For a horizon h , the future OOD gain is

$$\Delta_h A_t = A_{t+h} - A_t. \quad (2)$$

We also compute structural probe accuracy and structural nearest neighbor alignment as follows.

$$R_t^{\text{probe}} = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \frac{1}{|\mathcal{D}_p|} \sum_{x \in \mathcal{D}_p} \mathbf{1}[g_{t,p}(z_t(x)) = p(x)]. \quad (3)$$

$$R_t^{\text{nn}}(k) = \frac{1}{|\mathcal{D}_{\text{ood}}|} \sum_{(x,y) \in \mathcal{D}_{\text{ood}}} \frac{1}{k} \sum_{x' \in N_k^t(x)} \mathbf{1}[c(x') = c(x)]. \quad (4)$$

Here $\mathcal{P}$ is the set of probe targets, $\mathcal{D}_p$ is the evaluation set for target p , $p(x)$ is its gold label, $z_t(x)$ is the checkpoint representation, $g_{t,p}$ is the fitted probe, $N_k^t(x)$ is the set of k nearest training examples in representation space, and $c(x)$ is a structural class. Several probe targets are near ceiling, and COGS structural classes are imbalanced. Appendix E therefore lists the calibration required before treating these diagnostics as primary evidence.

The normalized composite diagnostic is

$$R_t = \frac{1}{|\mathcal{M}_t|} \sum_{m \in \mathcal{M}_t} \frac{R_t^m - \mu_m}{\sigma_m}, \quad (5)$$

where $\mathcal{M}_t$ indexes the available components among OOD token likelihood, structural probe accuracy, and nearest-neighbor alignment. R_t^m is the raw value of component m , and μ_m and σ_m are its mean and standard deviation over checkpoints within the same seed and setting. We use this composite for secondary forecasting within trajectories. Every result labeled token likelihood uses Equation 1 alone. The within-trajectory forecasting coefficient is

$$\rho_h = \text{corr}(R_t, \Delta_h A_t), \quad (6)$$

where corr denotes Pearson correlation computed over checkpoints for which $t + h$ is observed. For cross-run forecasting, the *early soft mean* is the mean OOD token log-probability over the first 30% of checkpoints, and the *early soft slope* is its least-squares slope against training step (measured in thousands). The *last-early* GEN EM and dev-EM values are taken at the final checkpoint of the early window. The pooled *progress baseline* is the strongest Spearman correlate among these two values and negative mean early training loss. The table normalized by regime restricts this comparison to last early GEN EM and development EM.

B.2 COGS Within-Trajectory Results

Across the three COGS baseline seeds, the horizon 3 composite correlations are positive but heterogeneous with values of 0.578, 0.288, and 0.086. Restricting the analysis to checkpoints before GEN EM reaches 0.20 yields $r = 0.539$. Later saturation weakens or reverses the association. Threshold and horizon analyses likewise show that the transition step depends on the seed and threshold. These results identify a measurable but seed-dependent COGS pre-generalization window.

B.3 Cross-Setting Forecasting Results

In the pooled 30-run cohort, early OOD token log-probability achieves Spearman $\rho=0.793$ with maximum GEN EM, exceeding the strongest progress baseline ($\rho=0.684$) by +0.108 with permutation $p=0.001$. For final GEN EM, the corresponding values are 0.759 and 0.713. The COGS capacity setting is the strongest individual case. Early token likelihood slope reaches $\rho=0.867$ with maximum GEN EM, while the best progress baseline is -0.127 . SLOG variable-free shows a smaller but positive margin of +0.267 over its progress baseline for final GEN EM.

At $n=10$, the COGS baseline progress features correlate more strongly with both maximum and final GEN EM than the soft features. After features and outcomes are standardized within each setting, token likelihood reaches 0.306 versus 0.283 for the progress baseline on maximum GEN EM, and 0.277 versus 0.422 on final GEN EM. Leave-one-setting-out performance also varies by the held-out regime. Table C.2 reports the regime-normalized and leave-one-regime-out results. Accordingly, we retain the cross-setting correlations as setting-dependent secondary evidence.

C Regime and Forecasting Evidence

This section summarizes the regime-level evidence that defines the protocol’s empirical scope. Table C.1 contrasts within-trajectory patterns, and Table C.2 tests cross-run associations after regime normalization and held-out-regime evaluation.

Table C.1: Evidence summary by regime. GEN EM is exact match accuracy on the held out generalization split. Within trajectory $h=3$ is the Pearson correlation for each seed between readiness R_t (Equation 5) and future OOD gain $\Delta_3 A_t$ (Equation 2). Values with $\pm$ are the mean and standard deviation across seeds. The dense diagnostic group contains three COGS seeds and three seeds per SIGMORPHON language. The main semantic parsing cohort contains 10 seeds per regime and 30 runs total.

Regime	Model	Maximum GEN EM	Final GEN EM	Within trajectory $h=3$	Interpretation
<i>Dense diagnostic trajectories</i>					
COGS baseline	scratch Transformer	0.525 $\pm$ 0.041	0.485 $\pm$ 0.058	0.317 $\pm$ 0.247	positive mean with substantial seed variation
SIGMORPHON EN/TR	scratch Transformer	–	–	no COGS like lead	gradual morphology boundary
<i>30 run semantic parsing cohort (n=10 per regime, see §B)</i>					
COGS baseline	scratch Transformer	0.543 $\pm$ 0.030	0.493 $\pm$ 0.053	+0.126 $\pm$ 0.242	positive mean with substantial seed variation
COGS capacity	capacity Transformer	0.482 $\pm$ 0.030	0.391 $\pm$ 0.048	–0.082 $\pm$ 0.204	negative mean with strong maximum EM signal across runs
SLOG variable free	scratch Transformer	0.309 $\pm$ 0.041	0.232 $\pm$ 0.061	+0.246 $\pm$ 0.187	structural setting with positive mean

Table C.2: Checks normalized by regime and evaluated with one regime held out for the cohort with 30 runs. Features and targets are standardized within each regime before pooling. ρ_R and ρ_B are Spearman correlations for the selected token likelihood and progress features, respectively. Their performance varies across held out regimes.

Analysis	Held-out regime	Target	R_t^{tok} feature	ρ_R	Progress feature	ρ_B
Regime-normalized pooled	–	maximum GEN EM	early soft mean	0.306	last-early GEN EM	0.283
Regime-normalized pooled	–	final GEN EM	early soft slope	0.277	last-early GEN EM	0.422
Leave-one-regime-out	COGS baseline	maximum GEN EM	early soft slope	0.331	last-early GEN EM	0.361
Leave-one-regime-out	COGS baseline	final GEN EM	early soft slope	0.287	last-early GEN EM	0.367
Leave-one-regime-out	COGS capacity	maximum GEN EM	early soft mean	0.239	last-early GEN EM	0.435
Leave-one-regime-out	COGS capacity	final GEN EM	early soft mean	0.218	last-early GEN EM	0.594
Leave-one-regime-out	SLOG variable-free	maximum GEN EM	early soft mean	0.382	last-early GEN EM	0.051
Leave-one-regime-out	SLOG variable-free	final GEN EM	early soft slope	0.365	last-early GEN EM	0.275

D Boundary and Bridge Protocols

This section records the evidence status of each comparison and then provides the capacity, pretrained, and subset-aligned morphology details.

Table D.1: Status of boundary and bridge protocols. The rows distinguish reported controlled evidence, reported pretrained comparisons, and exploratory protocols.

Protocol	Purpose and status
COGS capacity perturbation	Reported comparison that varies model capacity relative to the COGS baseline.
SLOG split without variables	Reported structural generalization condition whose target programs do not require explicit variables.
Pythia COGS adaptation	Exploratory pretrained decoder only adaptation. Released artifacts remain outside the main evidence.
T5-small and BART-base COGS bridge	Reported pretrained encoder and decoder comparison with 3 seeds per model. T5-small preserves the pattern after punctuation canonicalization. BART-base provides a low EM boundary.

Table D.2: Reported COGS capacity perturbation configuration.

Hyperparameter	COGS baseline	COGS capacity
Encoder layers	3	4
Decoder layers	3	4
Hidden size	256	256
Attention heads	4	8
FFN size	512	1024
Eval/save interval	1k steps	1k steps
Dense diagnostic seeds	3	3
Main cohort seeds	10	10

Table D.3: Pretrained encoder and decoder COGS bridge over 3 seeds per model. Canonical GEN EM removes differences in punctuation spacing only. Soft selected GEN EM is measured at the OOD token likelihood checkpoint. Early soft delta is the maximum OOD token log probability in the first 30% of checkpoints minus its value at the first checkpoint.

Model	Canonical maximum GEN EM	Soft-selected GEN EM	Soft regret	Early soft delta	Interpretation
T5-small	0.361±0.006	0.353±0.010	0.008±0.013	1.008±0.017	plateau pattern after punctuation canonicalization
BART-base	0.004±0.001	0.004±0.001	0.000±0.000	0.114±0.021	low-EM boundary despite early token-likelihood movement

Subset-aligned morphology. For SIGMORPHON, we compute readiness and target performance on the same held-out subsets to remove subset mismatch from the comparison.

Table D.4: Representative subset-aligned SIGMORPHON predictive correlations. Mean r is the mean Pearson correlation across three seeds between composite readiness and GEN EM gain three checkpoints later. All reported correlations are negative. MSD denotes morphosyntactic description.

Language	Aligned subset	Mean r	Maximum GEN EM
English	full dev	-0.9781	0.844
English	low edit	-0.8951	0.888
English	medium edit	-0.9665	0.747
Turkish	rare MSD	-0.9469	0.689
Turkish	common MSD	-0.9348	0.786
Turkish	high edit	-0.9583	0.796
Turkish	medium edit	-0.8621	0.710
Turkish	low edit	-0.9373	0.683

E Auxiliary Diagnostic Calibration

Structural probes and nearest-neighbor alignment appear only in the secondary forecasting analyses. Their interpretation is limited by near-ceiling probe targets and imbalanced COGS structural classes. Table E.1 lists the calibration required before treating these diagnostics as primary evidence. The selection protocol reports token likelihood results separately from the composite diagnostic.

Table E.1: Calibration requirements for auxiliary structural diagnostics.

Calibration item	Required check
<i>Nearest-neighbor alignment</i>	
Class distribution	Report majority-class effects in COGS structural labels.
Random/class-frequency baselines	Separate representation geometry from label imbalance.
k sensitivity	Establish whether structural alignment is stable across neighbor counts.
<i>Richer probe targets</i>	
Recursion/count buckets	Test whether probes capture more than shallow split membership.
Length and construction buckets	Establish whether probe accuracy varies enough across checkpoints.
Structural signatures	Test whether richer structural labels predict later OOD gains.