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

Anonymous ACL submission

Abstract

Checkpoint selection in out-of-distribution (OOD) studies is only as auditable as the validation signal used to choose the checkpoint. Exact match (EM) can be too sparse for that role: many checkpoints remain tied or all-zero while probability mass shifts toward the correct outputs. This creates a hidden model-selection degree of freedom: a paper can appear to select by OOD EM even when OOD EM cannot distinguish candidate checkpoints, leaving the reported score dependent on an unreported tie-breaker. We study this selection-resolution problem with dense checkpoints and labeled OOD references in semantic parsing and morphological reinflection. Our claim is a reporting protocol: when hard OOD EM is sparse or tied, report hard-EM resolution, latest-among-ties, and teacher-forced OOD token-likelihood tie-breaks; once hard EM is non-sparse, use direct OOD-EM selection. Across 30 scratch-trained runs, 8 OOD references leave the early hard-EM window all-zero in 48% of run/reference draws and the hard-EM maximum tied in 57%. Inside these plateaus, token likelihood reduces random tie-breaking regret from 0.063 to 0.050 over tied draws and from 0.059 to 0.044 in early all-zero windows, with hierarchical run/draw confidence intervals and leave-reference-out checks preserving the effect. Latest-among-ties is stronger on average in mostly monotone trajectories, so the contribution is not a new universal selector; we propose a reporting standard for making OOD checkpoint results auditable.

1 Introduction

Out-of-distribution (OOD) generalization is usually reported as exact match (EM): a model either matches the target sequence or it does not. EM is appropriate for final evaluation, but checkpoint selection asks a different question: did the validation signal have enough resolution to identify the selected checkpoint? A model may increase the probability of many correct OOD tokens before greedy decoding produces a fully correct sequence, leaving hard EM tied or all-zero while token likelihood moves. If a paper reports one checkpoint without showing whether hard EM could distinguish candidates, the result mixes two quantities: model quality and selection visibility. Two checkpoints can have the same labeled-OOD EM but different full generalization scores, so the reported score can depend on an implicit tie-breaker such as earliest, latest, random selection, or an unreported training convention. Our community-practice claim is simple: delayed-transition OOD studies should audit checkpoint-selection resolution before presenting a final OOD score.

We study teacher-forced likelihood of gold OOD target tokens, computed by scoring each gold output while conditioning on previous gold tokens. Because this requires labeled OOD references, it is not an unlabeled deployment signal. Given such references, the rule is direct: when hard OOD EM is all-zero or tied, report token-likelihood and latest-checkpoint tie-breaks; once hard OOD EM is non-sparse, use hard OOD EM directly.

The protocol changes what should be visible in OOD reports: not only the selected score, but also the labeled-reference budget, early all-zero rate, tied-maximum rate, token-likelihood tie-break, and latest-among-ties baseline. We also treat repeated reference draws from the same training run as correlated and use a leave-reference-out check that removes the selected tiny references from the reference-pool evaluation slice. Reporting hard-EM resolution lets readers distinguish model quality from checkpoint-selection visibility.

The main empirical object is the low-reference, pre-transition plateau. Across 30 runs and 20 reference-subset draws per run, 8 OOD references leave early hard EM all-zero in 48% of run/draws and the best hard-EM checkpoint tied in 57%. To-

ken likelihood improves over random hard-EM tie-breaking inside these plateaus, but latest-among-ties is stronger in mostly monotone trajectories and hard OOD EM becomes stronger once 16–64 references make EM less sparse. The fixed-count budgets expose this crossover from sparse hard signals to hard-EM selection. They are not meant to claim that every study selects from exactly eight references; they model early benchmark development, pilot OOD annotation, and low-budget validation audits where labeled OOD references exist but are scarce.

We evaluate COGS, a COGS capacity perturbation, SLOG variable-free semantic parsing, English/Turkish SIGMORPHON, and a pretrained T5-small/BART-base COGS bridge (Kim and Linzen, 2020; Li et al., 2023; Cotterell et al., 2017; Raffel et al., 2020; Lewis et al., 2020). These regimes test where hard-EM plateaus occur, whether token likelihood adds information inside them, and when the phase-dependent rule stops applying.

The resulting contribution is a selection audit rather than another leaderboard metric. Our contributions are:

1. We formalize the granularity mismatch between sequence EM and token likelihood, explaining why a dense soft signal can move inside an EM plateau.

2. We identify a selection-resolution failure mode: under tiny labeled-OOD budgets, many checkpoints share hard EM but yield different full-GEN regret.

3. We show that token likelihood improves over random hard-EM tie-breaking inside plateaus, while latest-among-ties is stronger in mostly monotone trajectories and must be reported.

4. We give a phase-dependent reporting rule that makes OOD checkpoint selection auditable, and characterize where it applies with regime controls, mechanism analysis, and pretrained bridge evidence.

2 Related Work and Positioning

Compositional-generalization benchmarks such as the Simplified version of the CommAI Navigation tasks (SCAN), Compositional Freebase Questions (CFQ), COGS, ReCOGS (a revised COGS-style benchmark), and SLOG show that high in-distribution accuracy does not guarantee systematic OOD behavior, and that conclusions depend on split structure, target formalism, architecture, and dataset artifacts (Lake and Baroni, 2018; Loula et al., 2018; Keysers et al., 2020; Kim and Linzen, 2020; Oren et al., 2020; Shaw et al., 2021; Csordás et al., 2021; Ontañón et al., 2022; Wu et al., 2023; Li et al., 2023; Hupkes et al., 2023). We use these benchmarks as temporal testbeds rather than proposing another final OOD score. SIGMORPHON reinflection supplies a contrasting morphology regime in which learning is more gradual (Cotterell et al., 2017).

Work on example forgetting, dataset cartography, grokking, double descent, and mechanistic progress measures shows that hard accuracy can hide important training dynamics (Toneva et al., 2019; Swayamdipta et al., 2020; Power et al., 2021; Liu et al., 2022; Belkin et al., 2019; Nakkiran et al., 2021; Olsson et al., 2022; Nanda et al., 2023). Our contribution is more operational: we ask when a labeled OOD validation signal provides a better checkpoint ranking than hard exact match or in-distribution dev accuracy.

OOD and domain-generalization work emphasizes that model selection under distribution shift must be specified explicitly (Arjovsky et al., 2019; Gulrajani and Lopez-Paz, 2021; Koh et al., 2021; Geirhos et al., 2020). This is directly relevant here because the same labeled OOD examples can support either a hard exact-match selector or a soft likelihood selector. A fair account must therefore state when each selector has resolution. Our contribution is a resolution-first reporting protocol for that decision: before comparing final OOD scores, first report whether hard OOD EM could actually distinguish the candidate checkpoints.

Finally, probing and representation-similarity work provides tools for auditing hidden representations, while warning that probes and similarity scores require controls (Alain and Bengio, 2017; Conneau et al., 2018; Hewitt and Manning, 2019; Hewitt and Liang, 2019; Tenney et al., 2019; Raghu et al., 2017; Morcos et al., 2018; Belinkov, 2022; Kornblith et al., 2019). We keep structural probes and nearest-neighbor alignment as calibrated auxiliary diagnostics. The load-bearing soft signal in this paper is OOD token likelihood; the load-bearing protocol is the joint report of hard-EM resolution, latest tie-breaking, and token-likelihood tie-breaking.

3 Soft and Hard OOD Checkpoint Signals

Let A_t denote OOD exact-match accuracy at checkpoint t. For a horizon h, define the future OOD gain as

$$\Delta_h A_t = A_{t+h} - A_t. \tag{1}$$

This quantity is useful for forecasting analyses, but the central selection problem is simpler: choose a checkpoint before knowing the full OOD trajectory.

OOD reference-token likelihood. The main soft signal is the teacher-forced mean log-likelihood of gold OOD output tokens, where teacher forcing means that each next-token probability is evaluated while conditioning on the gold previous tokens:

$$R_t^{\text{tok}} = \frac{1}{|\mathcal{D}_{ood}|} \sum_{(x,y)\in\mathcal{D}_{ood}} \frac{1}{|y|} \sum_{i=1}^{|y|} \log p_{\theta_t}(y_i \mid x, y_{<i}). \tag{2}$$

Equation 2 is available only when labeled OOD references are available. It is not a replacement for final exact-match evaluation. Its value is that it is additive over positions, whereas exact match is an all-token conjunction. A checkpoint can improve many gold-token probabilities without yet changing the decoded output enough to obtain a full exact match.

Selection definitions. Oracle regret is the full-GEN EM gap between the best evaluated checkpoint in a run and the checkpoint selected by a rule. For the tiny-reference analyses, the early window is the first 30% of evaluated checkpoints. A hard-EM plateau is either an all-zero early window or a tied hard-EM maximum across evaluated checkpoints on the same labeled OOD references. A token-likelihood-only checkpoint is one where hard OOD EM is still zero but normalized token likelihood has moved by at least 10% of its within-run range. All exact-match values use deterministic greedy decoding; beam search and decoded-output soft metrics are left as limitations.

Auxiliary representation diagnostics. We also compute structural probe accuracy and structural nearest-neighbor alignment:

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

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

Here $z_t(x)$ is the checkpoint representation, $g_{t,p}$ is a lightweight probe, $N_k^t(x)$ is the set of nearest training examples in representation space, and $c(x)$ is a structural class. These diagnostics help audit what the model represents, but they are not the main selector. Several probe targets are near ceiling, and COGS structural classes are imbalanced, so nearest-neighbor scores require random and class-frequency baselines.

For completeness, we also define a normalized composite score

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

where μ_m and σ_m are computed over checkpoints within the same seed and regime. The composite in Equation 5 is a summary diagnostic, not the load-bearing claim. When results are driven by Equation 2, we report them as token-likelihood results. The forecasting coefficient used in supporting analyses is

$$\rho_h = \operatorname{corr}(R_t, \Delta_h A_t)\,, \tag{6}$$

computed over checkpoints for which $t + h$ is observed.

Phase-dependent selection rule. The empirical results motivate the following labeled-OOD validation rule. At a fixed compute budget, evaluate each checkpoint on the same labeled OOD reference subset. If hard OOD EM is all-zero or tied, report both the Equation 2 tie-break and the latest-among-ties checkpoint. If hard OOD EM has a unique nonzero maximum, select by hard OOD EM. This rule encodes the paper's main claim: token likelihood is useful before hard exact match is reliable, but latest is an important monotone-training heuristic and hard EM is preferred once it provides a clean ranking. The intended output is an audit table rather than a new leaderboard number. For each regime, the report should state the reference budget and sampling scheme, the all-zero and tied-maximum rates of hard OOD EM, the checkpoint selected by hard EM, the latest tied checkpoint, and the token-likelihood tie-break. This makes the checkpoint-selection degree of freedom visible without claiming that any one selector is universally best.

4 Experimental Regimes

We choose regimes by the learning pattern they expose. COGS is the primary delayed-transition semantic-parsing case (Kim and Linzen, 2020). It contains a long interval in which in-distribution exact match is high while OOD exact match remains low. The Conference on Computational Natural Language Learning (CoNLL)–SIGMORPHON English and Turkish reinflection tasks provide gradual morphology regimes (Cotterell et al., 2017). The SLOG variable-free split is the SLOG semantic-parsing benchmark evaluated on examples whose target programs do not require explicit variables; it is a harder structural generalization split, not a new dataset.

All controlled scratch regimes use randomly initialized encoder–decoder Transformers with dense checkpointing every 1000 steps and diagnostics every third checkpoint. The compact COGS/SLOG layout is 3 layers, d_{model}=256, FFN 512 (COGS ∼4.6M parameters; batch 32; 60 epochs); the capacity variant changes layers, heads, and FFN size (Appendix Table C.2). The 30-run expanded cohort covers the COGS baseline (n=10), a COGS capacity perturbation (n=10), and the SLOG variable-free split (n=10). Appendix Table B.1 summarizes the regime-level evidence and scope. Generalization exact match (GEN EM) denotes greedy-decoding exact-match accuracy on the held-out generalization split, and in-distribution development exact match (dev-EM) denotes greedy-decoding exact-match accuracy on the in-distribution development split.

Pretrained bridge status. Pythia adaptation is retained as an exploratory decoder-only bridge protocol. We also run T5-small and BART-base encoder–decoder bridges on COGS (Raffel et al., 2020; Lewis et al., 2020), using the same dense-checkpoint logs and the same hard-EM sparsity/tie-break summary used for the scratch experiments. Because pretrained adaptation can reflect elicitation of existing structure rather than formation from scratch, these bridges test external validity rather than replace the controlled scratch evidence. Appendix Table C.1 records the status of these boundary and bridge protocols.

Reproducibility materials. We release code, configurations, generated table/JSON summaries, and scripts for the reported tables, so numerical claims are traceable to configurations, cached summaries, and analysis scripts. The original COGS and SIGMORPHON dense-checkpoint experiments were run on one NVIDIA RTX A5000 graphics processing unit (GPU; 24GB video random-access memory, VRAM); expanded-cohort summaries can be regenerated by rerunning the included training and analysis scripts. Because hardware was not controlled for wall-clock benchmarking, we report model configuration, seeds, checkpoint cadence, and analysis scripts rather than treating runtime as an evaluation metric.

5 Results

The Results section follows the audit hierarchy: first show that hard OOD EM often lacks checkpoint-selection resolution, then ask whether token likelihood supplies an independent ordering inside those plateaus, and finally delimit the regimes where the protocol applies. Figure 1 gives the visual intuition, and Table 1 isolates the conditional plateau cases that carry the main selection claim.

5.1 Token Likelihood Ranks Inside Hard-EM Plateaus

The central experiment compares hard OOD exact-match selection and soft OOD token-likelihood selection on the same labeled OOD reference subsets. At 8 references, hard OOD EM is sparse: the early window is all-zero in 48% of run/draws, and the across-checkpoint hard-EM maximum is tied in 57%. In those cases, a hard selector cannot distinguish many checkpoints without an arbitrary tie-breaker. Token likelihood remains dense: it provides an average of 6.7 token-likelihood-only checkpoints where hard OOD EM is still zero but normalized token likelihood has moved by at least 10% of its within-run range.

Clustered uncertainty and reference leakage checks. The 20 reference draws from a single run are correlated, so Table 1 reports hierarchical bootstrap intervals that resample runs and then draws within runs, rather than treating all run/draw pairs as independent. For the central 8-reference tied and early-all-zero cases, the soft-minus-random regret differences remain negative under hierarchical bootstrap ($[-0.022, -0.005]$ and $[-0.029, -0.004]$). Removing the selected tiny references from the 2100-example OOD

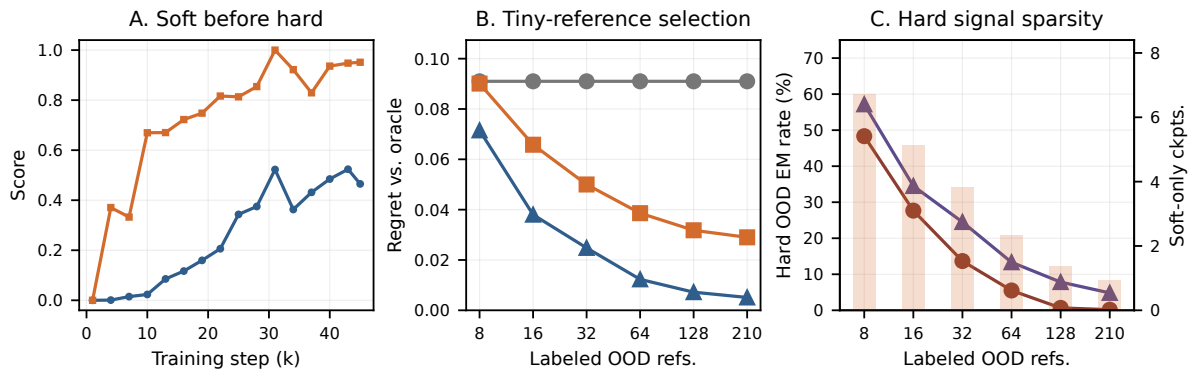

Figure 1: Soft signals before hard OOD generalization. Left: in a representative COGS baseline trajectory, normalized OOD token log-probability (orange squares) rises while exact-match GEN EM (blue circles) remains near zero. Middle: checkpoint-selection regret under ultra-small labeled-OOD budgets: dev EM is gray, hard OOD EM is blue, and soft OOD token likelihood is orange. Right: hard exact match is frequently all-zero (brown) or tied (purple) at tiny budgets, whereas token-likelihood-only checkpoints remain visible as orange bars.

reference-pool evaluation slice leaves the deltas nearly unchanged (-0.012 and -0.015).

What latest-tie does and does not settle. Latest-among-ties is a strong baseline because many trajectories are mostly monotone; averaged over 8-reference tied draws, latest has lower regret than token likelihood (0.037 vs. 0.050). This is why our recommendation is to report latest and token likelihood together, not to replace latest. However, latest is not a complete audit: in $120/343$ tied 8-reference draws, latest still has at least 0.05 GEN-EM regret, and in that subset token likelihood reduces average regret from 0.088 to 0.061 and beats latest in 47% of draws (Appendix Table A.3). Thus latest is stronger on average, while token likelihood exposes an independent dense ordering in the minority of hard-EM plateaus where recency is not safe. The absolute regret reductions are modest, but the purpose of the audit is not to create a larger leaderboard gain; it is to reveal when a reported OOD score depends on an unresolved selection degree of freedom.

Table 1 is the main plateau result. It shows the complete phase pattern: at 8 references, token likelihood gives a statistically stable improvement over random hard-EM tie-breaking; at 16–64 references the advantage narrows as hard EM becomes less sparse; and latest-among-ties 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 self-contained conditional audit. The larger 1%–100% reference scaling analysis is reported in Appendix Table A.5. Together, these results locate the value of token likelihood: it is a dense audit signal when hard exact match is still uninformative, not a universal replacement for recency or hard-EM selection.

Task-specificity control. The token-likelihood signal is not merely generic perplexity drift. 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 gold-minus-permuted log-likelihood gap grows from 2.598 at step 7,000 to 3.251 at step 45,000; the gold-minus-mismatched gap similarly grows from 2.601 to 3.254. This control does not establish a mechanism, but it supports the interpretation that Equation 2 becomes increasingly specific to the correct task-structured outputs rather than only reflecting global likelihood drift.

5.2 COGS: The Delayed Window Exists but Its Timing Is Seed-Sensitive

COGS provides the clearest delayed-transition case. In all seeds, in-distribution dev/test exact match rises earlier than OOD exact match, creating a window in which exact-match GEN remains low despite continued changes in model state. This is the setting where the plateau analysis in Section 5.1 is meaningful.

The predictive statistics are weaker than a broad readiness claim would require, and we report that directly. At horizon $h = 3$, the original three-seed composite correlations are positive but heterogeneous: 0.578, 0.288, and 0.086. The stronger association is phase-local rather than trajectory-universal: before transition (GEN EM < 0.20),

Table 1: Tie-breaking regret when hard OOD EM lacks resolution. Soft is teacher-forced OOD token likelihood restricted to the same hard-EM plateau; Random averages uniformly over tied hard-EM checkpoints; Latest picks the latest tied checkpoint. Δ is Soft–Random, so negative values favor token likelihood. Confidence intervals resample runs and then reference draws within runs. LRO removes the selected tiny references from the 2100-example OOD reference-pool evaluation slice. The table supports an audit protocol, not a replacement selector: token likelihood beats random tie-breaking at 8 references, 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 all-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

the association is $r = 0.539$, while later saturation weakens or reverses it. We therefore interpret COGS as evidence for a measurable delayed window, not as evidence that a single readiness score robustly predicts future OOD gains in every seed.

Transition-window metrics are the right abstraction for this result. The diagnostic is useful when it detects movement before hard exact match becomes measurable: first soft-likelihood movement, first nonzero GEN EM, first GEN EM above a threshold, plateau length, and tie/all-zero rate. The released threshold and horizon analyses show that the precise transition step depends on seed and threshold, but the qualitative condition needed for the main claim is stable: there is a low-hard-EM interval in which Equation 2 can move.

5.3 What Does Token Likelihood Track?

The mechanism result further narrows the claim. OOD token likelihood does not prove that a model has learned structural generalization. It tracks the generalization mode that the regime actually supports. Across the 20 COGS-like expanded runs (COGS baseline plus capacity), early-window slope of Equation 2 is more aligned with lexical GEN EM than with structural GEN EM: in the COGS capacity perturbation, it correlates with max lexical GEN EM at Spearman $+0.879$ but with max structural GEN EM at -0.188; in the COGS baseline 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 max structural GEN EM ($+0.359$). Thus the diagnostic is behavior-adjacent: it tracks emerging compatibility with the OOD outputs, but the type of generalization being tracked depends on the regime.

5.4 Checkpoint Selection Beyond the Tiny-Reference Window

Full-reference token likelihood remains useful in these regimes as a practical alternative to in-distribution dev-EM selection. Across the 30 expanded-cohort runs, selecting by Equation 2 reduces oracle regret, the gap to the best evaluated checkpoint in the same run, from 0.068 to 0.025 on the COGS baseline, from 0.065 to 0.013 on 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 run-level breakdown. An in-domain dev token-likelihood baseline does not explain the effect: selecting by dev-split teacher-forced token likelihood has pooled regret 0.097, similar to or worse than dev-EM regret 0.091, and far above OOD token-likelihood regret 0.020 (Appendix Table A.4). It helps on the capacity perturbation (0.050 versus dev-EM 0.065), but not on the COGS baseline (0.095) or SLOG variable-free (0.146). This supports the interpretation that the useful dense signal is OOD-reference-specific rather than a generic preference for likelihood-based checkpointing.

This result answers a different question from Table 1. It shows that soft OOD likelihood can outperform standard dev-EM selection when labeled OOD references are available in the studied

regimes. It does not show that soft likelihood is always better than hard OOD EM or latest-among-ties on the same references. The same-reference hard-EM baseline shows the opposite once exact matches are non-sparse, and Table 1 shows that latest is a strong heuristic when trajectories are mostly monotone. The resulting practical hierarchy is therefore explicit: report token likelihood as a dense plateau signal, report latest as the monotone tie-breaker, and use hard OOD EM when it supplies a reliable ranking.

5.5 Forecasting Across Regimes Is Supporting but Mixed

The broader forecasting analysis supports the diagnostic but should not be the headline. In the pooled 30-run cohort, early OOD token log-probability achieves Spearman $\rho{=}0.793$ with max 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 versus 0.713. The COGS capacity perturbation is the strongest per-regime case: early token-likelihood slope reaches $\rho{=}0.867$ with max GEN EM, while the best progress baseline is -0.127. SLOG variable-free shows a smaller but positive final-GEN margin of $+0.267$.

However, the COGS baseline alone does not show a cross-run readiness advantage at $n{=}10$: progress baselines beat readiness for both max and final GEN EM. After z-scoring features and outcomes within regime, Equation 2 only slightly beats progress baselines for max GEN EM (0.306 vs. 0.283) and loses for final GEN EM (0.277 vs. 0.422). Appendix Table B.2 gives the regime-normalized and leave-one-regime-out checks. Leave-one-regime-out checks are also mixed. These results are useful boundary evidence, but they reinforce the narrower framing: token likelihood is not a regime-invariant forecasting law.

A negative-control analysis mitigates the concern that Equation 2 merely reflects generic perplexity drift. The gap between gold OOD targets and permuted or mismatched targets grows across early, middle, and late checkpoints, supporting the interpretation that the signal becomes increasingly specific to task-structured outputs. This control strengthens the plateau-ranking result without turning it into a universal predictor.

5.6 SIGMORPHON: A Gradual Boundary Case

The SIGMORPHON experiments test whether the same lead-time diagnostic appears in morphology. It does not. Subset-aligned English and Turkish analyses have negative mean horizon-3 correlations (-0.9466 and -0.9325), with every aligned subset/seed cell negative. Representative aligned subset results are reported in Appendix Table C.4. We interpret this as a boundary condition rather than as evidence that morphology lacks useful soft signals. The specific hard-sparse/soft-dense window studied here is most useful when the trajectory contains delayed OOD exact-match emergence. Gradual morphology offers little sustained pre-transition plateau of that kind.

5.7 Pretrained Encoder–Decoder Bridge

The pretrained bridge is deliberately not the main evidence, but it addresses whether the phenomenon is tied to the compact scratch Transformer. We fine-tune T5-small and BART-base on COGS for three seeds each, logging dense checkpoints, generated outputs, and teacher-forced OOD token likelihood. For pretrained models, exact-match evaluation needs COGS punctuation canonicalization: T5 often decodes `teleport. agent` where the benchmark target writes `teleport . agent`; canonicalization removes only punctuation-spacing differences and does not reorder conjuncts or rename variables.

Under this canonical evaluation, T5-small provides a positive bridge. Across three seeds, canonical max GEN EM is $0.361{\pm}0.006$, final GEN EM is $0.361{\pm}0.006$, and selecting by OOD token likelihood reaches $0.353{\pm}0.010$ GEN EM with mean regret 0.008. The tiny-budget phase pattern also appears: with 8 references, the early hard-EM window is all-zero in $0.317{\pm}0.153$ of draws and the hard-EM maximum is tied in $0.633{\pm}0.058$ of draws. BART-base is a boundary case: canonical max GEN EM remains $0.004{\pm}0.001$ even though OOD token likelihood improves early by $0.114{\pm}0.021$. Thus the bridge supports the plausibility of the dense-soft/hard-sparse hierarchy beyond the scratch model, but it also reinforces the paper's conservative framing: pretrained architecture and formatting choices matter, and bridge evidence is mixed rather than universal. Appendix Table C.3 gives the bridge summary.

6 Discussion

The main result is that hard-EM plateaus are not just low performance; they are low-resolution validation signals. They require an auditable report rather than a hidden tie-breaker. Once hard exact match becomes non-sparse, direct hard-OOD exact-match selection becomes the appropriate selector for the reported exact-match metric. This phase-dependent hierarchy turns the label dependency of Equation 2 into a protocol rather than a hidden weakness.

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 be Equation 2, and let $\mathcal{A} = \arg\max_{t \in \mathcal{T}} H_t$. If no labeled OOD references are available, Equation 2 is unavailable. If $\max_t H_t = 0$, hard EM cannot rank checkpoints; report both $\arg\max_t L_t$ and the latest checkpoint. 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.

What the evidence supports. Equation 2 is labeled-OOD validation likelihood; the contribution is a resolution comparison under the same references. Table 1 isolates the setting where this distinction matters: hard EM is tied or all-zero, while token likelihood lowers regret relative to random/dev tie-breaking inside the plateau. Latest-among-ties is stronger in many monotone trajectories, so token likelihood is not the default replacement for recency. Its role is to supply an independent reference-conditioned ordering for auditing cases where exact match lacks resolution. The full-reference selection and forecasting results are supporting evidence; the conceptual center is checkpoint ranking when exact match cannot rank.

Scope and reporting recommendation. The boundary results specify when the protocol applies. SIGMORPHON shows that gradual morphology need not contain a COGS-like lead-time window; regime-normalized forecasting prevents the pooled 30-run correlation from being read as a regime-invariant law; and BART-base shows that pretrained adaptation can remain a low-EM boundary case even when token likelihood moves. For delayed-transition benchmarks, we recommend reporting checkpoint-selection evidence separately from final OOD accuracy: the reference budget and sampling scheme, early all-zero rate, tied-maximum rate, same-reference hard-EM selector, latest tie-break, and token-likelihood selector. This convention separates final generalization quality from the visibility of progress during training and keeps the mechanism claim bounded: Equation 2 measures compatibility with labeled OOD outputs, not a fixed sub-skill or causal mechanism.

7 Conclusion

This paper argues that sparse OOD EM creates a resolution problem for checkpoint selection. Labeled OOD token likelihood is useful as part of a phase-dependent report, especially inside hard-EM plateaus: with 8 OOD references, hard EM is often all-zero or tied, and token likelihood improves over random/dev tie-breaking under run-clustered and leave-reference-out checks. But latest-tie is stronger on average in the mostly monotone trajectories studied here, so token likelihood should be reported as dense diagnostic resolution rather than sold as the default selector. With larger reference budgets, hard EM gains enough resolution and should become the direct selector. Boundary cases in SIGMORPHON, BART-base, regime-normalized forecasting, and lexical/structural analyses show where the protocol applies: labeled-OOD checkpoint selection when exact match is too sparse to reveal partial progress. The practical report is small: state whether hard EM had a unique nonzero maximum on the labeled OOD references, list latest and token-likelihood tie-breaks when it did not, and keep final test EM separate from the selection audit. Making that resolution visible also changes how low OOD scores are read: they may reflect model failure, or they may reflect a plateau where hard EM did not identify the reported checkpoint. The recommendation is simple: do not report an OOD EM-selected checkpoint without also reporting whether OOD EM had enough resolution to select it.

8 Limitations

The diagnostic in Equation 2 requires gold OOD reference outputs under teacher forcing, so it is a labeled-OOD validation diagnostic rather than an unlabeled deployment detector. The reference set must represent the target shift; otherwise hard

EM, latest-tie, and token likelihood can all select for the wrong subproblem. Uniform, stratified, and phenomenon-targeted reference sampling can therefore answer different questions, and the sampling scheme should be reported with the reference budget. Teacher-forced likelihood is also sensitive to output format and canonicalization, as the T5-small bridge illustrates, and we do not evaluate beam decoding, smoothed hard EM, decoded-output edit distance, token F1, entropy/margin selectors, or mixed-prefix likelihood. The protocol assumes dense checkpoint logs and is most appropriate for controlled training studies and labeled-OOD validation analyses, not post-hoc claims about arbitrary pretrained systems. Finally, structural generalization remains difficult across settings, and CFQ is excluded because we could not establish a reproducible, sanity-checked protocol.

9 Ethical Considerations

This work uses public benchmark datasets and synthetic or benchmark-provided semantic parsing and morphological reinflection examples; it does not introduce human-subject data, collect private information, or annotate sensitive content. The main risk is interpretive: a labeled-OOD validation diagnostic should not be mistaken for an unlabeled deployment detector or for evidence that structural generalization has been solved. We therefore state the label requirement, report boundary cases, and keep pretrained-system claims exploratory unless the bridge experiments pass the same promotion checks as the main evidence.

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.

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*.

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 Lanas 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.

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.

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 Vaclav 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*.

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: first the low-reference hard-EM sparsity analysis, then conditional tie-breaking inside hard-EM plateaus, then full-reference paired selection statistics, and finally the larger small-reference scaling study. The goal is to make the proposed reporting protocol auditable: 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-run generalization exact-match log. "Early hard all-zero" is the fraction of run/draws 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/draws 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, a strong heuristic in mostly monotone trajectories; "Dev" breaks ties by in-distribution development exact match. The soft likelihood improvement over random hard-EM tie-breaking is clearest in the 8–16 reference plateau regime, but latest is the stronger selector in these mostly monotone trajectories. The final column shows that token likelihood still wins in a nonzero minority of tied draws, which is why the protocol reports both signals instead of hiding the tie-breaker. The appendix reports the row-level diagnostic confidence intervals from the full table; Table 1 reports the run-clustered intervals used for 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 still has at least 0.05 GEN-EM regret against the run oracle. This is the setting where the monotone recency heuristic is not sufficient. Within this subset, token likelihood reduces average regret for the main 8-reference plateaus and remains competitive for larger tied plateaus, which supports reporting token likelihood alongside latest rather than replacing latest on average. 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), but selecting by in-domain dev token likelihood does not reproduce the effect. This separates the OOD-reference signal from a generic preference for likelihood-based checkpointing.

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 against the per-run oracle generalization exact match over evaluated checkpoints. The result shows that the practical checkpoint-selection gain does not require the full OOD validation set, although it still requires labeled OOD references.

Scope	Labeled OOD refs.	Dev-EM regret	R_t^{tok} regret	Mean-regret reduction	COGS baseline regret	COGS capacity / SLOG regret
Pooled	1% ($\sim$210)	0.091	0.031 ± 0.047	66%	0.025	0.043 / 0.027
Pooled	5% ($\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 Regime and Forecasting Evidence

This section collects the regime-level evidence behind the paper's scope. These tables are intentionally placed after the selection-protocol audit: they explain where the protocol applies and where forecasting claims should remain secondary.

Table B.1: Regime-level evidence summary. GEN EM is exact-match accuracy on the held-out generalization split. Within-trajectory h=3 is the per-seed Pearson correlation between readiness R_t (Equation 5) and future OOD gain $\Delta_3 A_t$ (Equation 1). Expanded rows are cross-run measurements over 30 seeds total; they support the positive COGS/SLOG cases while preserving SIGMORPHON as a gradual boundary case.

Regime	Model	Max GEN EM	Final GEN EM	Within-trajectory h=3	Interpretation
COGS baseline	scratch seq2seq	0.525 ± 0.041	0.485 ± 0.058	0.317 ± 0.247	positive transition-like case
SIGMORPHON EN/TR	scratch seq2seq	–	–	no COGS-like lead	gradual morphology boundary
Expanded multi-regime extension (n=10 seeds per regime; see §5.5)					
COGS baseline	scratch seq2seq	0.543 ± 0.030	0.493 ± 0.053	$+0.126\pm0.242$	wider replication; trajectory noise visible
COGS capacity	capacity seq2seq	0.482 ± 0.030	0.391 ± 0.048	-0.082 ± 0.204	cross-run max-EM signal strong
SLOG variable-free	scratch seq2seq	0.309 ± 0.041	0.232 ± 0.061	$+0.246\pm0.187$	harder structural; positive trajectory trend

Table B.2: Regime-normalized and leave-one-regime-out checks for expanded-cohort forecasting. Features and targets are z-scored within regime before pooling; mixed results support treating forecasting as secondary.

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

C Boundary and Bridge Protocols

This section separates reported boundary evidence from exploratory bridge protocols, ordered by protocol status, capacity perturbation, pretrained bridge, and subset-aligned SIGMORPHON analysis.

Table C.1: Status of boundary and bridge protocols. Capacity and SLOG rows are reported evidence; pretrained rows test external validity without replacing the controlled scratch-trained evidence.

Protocol	Purpose and status
COGS capacity perturbation	Reported capacity perturbation testing whether the COGS transition-window analysis is tied to one compact scratch-trained architecture.
SLOG variable-free split	Reported harder SLOG structural-generalization condition whose target programs do not require explicit variables.
Pythia COGS adaptation	Exploratory pretrained decoder-only bridge protocol; current artifacts are not promoted into the main evidence.
T5-small/BART-base COGS bridge	Reported pretrained encoder–decoder bridge with 3 seeds per model. T5-small is a positive bridge after COGS punctuation canonicalization; BART-base is a low-EM boundary case.

Table C.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
Original trajectory seeds	3	3
Expanded-cohort seeds	10	10

Table C.3: Pretrained encoder–decoder COGS bridge over 3 seeds per model. Canonical GEN EM removes punctuation-spacing differences only; soft-selected GEN EM is measured at the OOD-token-likelihood checkpoint.

Model	Canonical max 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	positive pretrained encoder–decoder bridge 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. We rerun SIGMORPHON with readiness and target performance computed on the same held-out subsets; the aligned analyses still do not recover a COGS-like positive lead-time relation.

Table C.4: Representative subset-aligned SIGMORPHON predictive correlations. r is the three-seed mean correlation with horizon-3 future gain; all aligned cells remain negative.

Language	Aligned subset	Mean r	Max A_t
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

D Auxiliary Diagnostic Calibration

The main claim does not rely on structural probes or nearest-neighbor alignment. We therefore keep auxiliary checks after the selection and boundary evidence. The checklist below records what must be controlled before these diagnostics are promoted to primary evidence; until then, checkpoint choice remains governed by hard exact-match resolution, token-likelihood tie-breaking, and latest-among-ties reporting.

Calibration checklist.

Calibration item	Purpose
Nearest-neighbor alignment	
Class distribution	Quantify majority-class effects in COGS structural labels.
Random/class-frequency baselines	Separate representation geometry from label imbalance.
k sensitivity	Check whether structural alignment is stable across neighbor counts.
Richer probe targets	
Recursion/count buckets	Test whether probes capture more than shallow split membership.
Length and construction buckets	Check whether probe accuracy has enough dynamic range over checkpoints.
Structural signatures	Evaluate whether richer structural labels remain auxiliary or become predictive.