Learning, or Just Retrieving? A Copy-Baseline Audit of Zero-Shot Foundation Models in Time-Series and Single-Cell Tasks

Ali Asaria
Transformer Lab

Tony Salomone
Transformer Lab

Deep Gandhi*
Transformer Lab

Abstract

Zero-shot foundation models are increasingly applied far from language, in time-series forecasting and single-cell genomics, on the promise that large-scale pretraining lets them solve new tasks in context. We ask how much of that apparent skill, measured at the score level, a cheap copy already accounts for. We introduce a modality-agnostic audit that places three predictors on one held-out metric: a non-copying floor (climatology mean, or majority class), a retrieve-and-copy baseline (seasonal or last-value copy, or a nearest-neighbor analog for time series; PCA plus nearest-neighbor vote for single cells), and the foundation model itself. The audit reports a single diagnostic, the Retrieval-Explained Fraction (REF), the share of the model's above-floor skill that the cheap copy already achieves. We summarize a family of cheap copies by its strongest single member (the strongest simple copy in a small pre-registered family), an upper bound on what one cheap rule achieves rather than an unbiased expected value. Across four established forecasting series, for Chronos-Bolt-base the strongest simple copy matches most of the model's above-floor skill on these metrics: the mean strongest-copy REF is about 0.72 (range 0.57 to 0.99). On ETTm1 both foundation-model families are statistically indistinguishable from a last-value copy: the binding-copy REF confidence interval includes 1 for both Chronos-Bolt (0.987, 95% CI [0.906, 1.050]) and TimesFM (0.999, 95% CI [0.952, 1.046]), so near-total parroting is significance-backed for two independent families, while on the other series the REF CI excludes 1 and the model adds skill. A similar pattern appears in a very different modality: on pbmc3k cell-type annotation a plain PCA copy is significantly better than the scGPT foundation model on macro-F1 (0.905 versus 0.813, gap 95% CI excludes 0) and tied on accuracy (McNemar exact p=0.59). Read at the score level, the answer to the memorable question "is your foundation model learning, or just retrieving?" is, in the settings audited here, that a trivial copy matches most of the model's above-floor skill on these metrics. Confidence intervals are now reported (moving-block bootstrap for the autocorrelated forecasting series; cell bootstrap plus exact McNemar for single cells), so the small-margin claims are tested rather than hedged. We package the audit as a reusable harness so that any model and task can be checked the same way.

1 Introduction

Zero-shot "foundation models" now reach well beyond language. In time-series forecasting, models such as Chronos [1] and TimesFM [2] are presented as general forecasters that require no task-specific training; in single-cell genomics, transcriptomic models such as scGPT [3] annotate cells from a

*Corresponding author: deep@lab.cloud

reference set in the same zero-shot spirit. The shared promise is that scale of pretraining substitutes for task-specific learning, letting a single model solve a new task in context.

A recurring but scattered set of observations complicates that promise. In forecasting, "context parroting", retrieving the nearest past motif and copying its continuation, is a trivial rule that is hard to beat [4]; in transcriptomics, "one PCA still rules them all" reports that plain linear baselines match or beat pretrained models on several tasks [5]. Each finding is well supported inside its own field, yet the two have not, to our knowledge, been unified, and it is not known whether they are a general property of the zero-shot foundation-model paradigm or a set of disconnected quirks.

This work turns those isolated observations into a single, portable test. The prediction unit is one zero-shot query per modality: a forecasting context window and horizon, or a query cell to be annotated from a reference set. For each query we score three predictors on the same held-out metric, a non-copying floor, a retrieve-and-copy baseline, and the foundation model under audit, and summarize the comparison with one normalized diagnostic, the Retrieval-Explained Fraction (REF): the fraction of the model's above-floor skill that the cheap copy already reproduces. REF near one means the model mostly mimics a trivial copy; REF well below one means it genuinely adds skill; REF above one means copying is better than the model outright. The baseline is fixed before the model numbers are seen, and its strength is reported as a band, so a high REF is a finding rather than an artifact of a hand-tuned baseline.

Our contributions are:

1. A modality-agnostic copy-baseline audit and the REF metric, defined purely from floor, copy, and model scores on a shared metric and therefore comparable across very different domains (§3).

2. Time-series evidence, across four established forecasting series and two independent model families (Chronos-Bolt[1] and TimesFM), that at the score level a trivial copy matches most of these models' above-floor skill: for Chronos-Bolt-base the mean strongest-copy REF is about 0.72 (range 0.57 to 0.99), with the copy matching nearly all of it on ETTm1 for both families (§5).

3. A similar pattern in a very different modality, single-cell genomics, where a plain PCA copy is at least as good as the scGPT foundation model on cell-type annotation (§5).

4. A reusable audit harness, a fixed scoring evaluator plus swappable model predictors, proven on three foundation models across two modalities (§8).

2 Related Work

Copy and retrieval baselines for foundation models. The closest prior work is context parroting, which shows that copying the nearest past motif is a simple but tough-to-beat baseline for scientific foundation models [4]. In forecasting specifically, Toner et al. [6] report that zero-shot time-series models lose to a naive seasonal-copy baseline on cloud-demand data, and Tire et al. [7] build a retrieval-augmented forecaster that matches and copies past windows with a leakage-safe retrieval bank and a matched-budget control, both of which we adopt in spirit. Nearest-neighbor forecasting [8] is a retrieval-augmentation method whose retrieval mechanics (neighbor weighting and sensitivity to the number of neighbors) we borrow as the pre-registered hyperparameters of

[1] An efficient variant of Chronos [1].

the copy baseline. In genomics, Bendidi et al. [5] show that a single PCA still rules perturbation analysis over pretrained transcriptomic models, and Mejia et al. [9] show that a trivial baseline matches or beats several single-cell models under standard metrics, and warn that metric choice, not just the model, decides whether a trivial baseline appears to win. Our audit unifies these threads: rather than proposing a better forecaster or annotator, it puts copy and model on one axis in both modalities and reports the resulting REF.

Why a copy baseline should track model behavior. Theory offers a mechanism. Li et al. [10] prove that a single softmax-attention layer converges to a one-nearest-neighbor predictor in context, and Park et al. [11] show that in-context learning contains competing retrieval and inference solutions in which the copy-like solution often supersedes the generalizing one, a view that has been informally linked to an implicit nearest-neighbor reading of attention [12]. This is why a cheap copy is expected to shadow a zero-shot model, and it frames our empirical measurement rather than being tested directly.

The optimistic counter-view, and leakage. A competing account argues that time-series models perform implicit reasoning and generalize on synthetic out-of-distribution probes [13], but without a copy baseline on the same axis; our contribution is precisely to supply that axis, corroborating the optimistic view where REF is low and challenging it where REF is high. Separately, several benchmarks warn that pretraining leakage can inflate zero-shot scores [14, 15, 16]; we therefore use strictly-past retrieval and reference-only splits, and note that a high REF under possible leakage is, if anything, more damning. We further build on normalized foundation-model-versus-baseline comparisons and memorization-versus-generalization analyses [17, 18], on the nearest-neighbor vote of zero-shot annotation methods [19] (adapted here into a copy that retrieves strictly from a labeled reference split so that it remains a true copy floor), and on recent single-cell modeling, interpretability, and forecasting-observability work [20, 21, 22].

3 Method

The audit scores three predictors on one held-out metric per query, arranged in three tiers.

Floor (non-copying). The floor establishes the "do nothing" score and must not itself be a copy, so that the copy baseline and the model are both measured against a genuinely trivial reference. For forecasting it is the climatology mean (the in-context mean continued over the horizon); for single-cell annotation it is the majority class. Neither reuses a retrieved target, which is what keeps REF interpretable.

Copy family (retrieve and copy). The copy baseline retrieves the nearest example in a cheap input representation and copies its target. For time series we consider a small pre-registered family: a seasonal copy (repeat the last seasonal period), a last-value copy (persist the final observation over the horizon), and a nearest-neighbor analog copy (retrieve the most similar strictly-past window and copy its continuation). Two of these rules are the classical naive forecasting baselines: last-value is persistence and the seasonal copy is seasonal-naive, so REF measures skill above a no-structure climatology floor rather than above these naive baselines; the analog/kNN retrieval copy is the pure retrieval mechanism. For single cells the copy is a PCA embedding of log-normalized expression followed by a nearest-neighbor vote over the labeled reference set. Because more than one trivial rule is a copy, and because which rule binds depends on the structure of the series (in our series a smooth, highly autocorrelated target tends to be best copied by last-value, a periodic one by the seasonal rule), REF is reported as a band over the copy family. The headline is the strongest single copy:

the family member that reproduces the most above-floor skill. This is an upper bound on what one cheap rule achieves (a best-case, maximum-over-family selection statistic), not an unbiased expected REF, and the band is a design range over the pre-registered family, not a statistical confidence interval.

Foundation model. The model under audit is run zero-shot on the identical context and query, and scored by the identical metric.

The REF metric. Let s_x be the score of predictor x and s_{floor} the floor score. Define the above-floor skill $I(x)$ in the favorable direction of the metric, and the Retrieval-Explained Fraction as the ratio of the copy's skill to the model's:

$$\text{REF} = \frac{I(\text{copy})}{I(\text{FM})}, \qquad I(x) = \begin{cases} s_{\text{floor}} - s_x & \text{MASE (lower is better)}, \\ s_x - s_{\text{floor}} & \text{macro-F1 (higher is better)}. \end{cases} \tag{1}$$

REF near one means the copy reproduces almost all of the model's above-floor skill; REF above one means the copy is better than the model. Defining REF purely from floor, copy, and model scores on a shared metric makes it modality-agnostic by construction, which is what lets the same number be compared across forecasting and genomics. REF is a score-level diagnostic, not a mechanistic one: equal scores do not establish that the model internally copies, because two predictors can reach the same MASE with different per-item errors. The per-item shared-failure test (whether the model and the copy fail on the same queries), which we defer, is what would license a mechanistic reading. REF is also a ratio estimator that becomes unstable as $I(\text{FM})$ approaches zero, so REF values for models only slightly above the floor should be read with caution.

Leakage-safe splits. No post-target information enters any predictor. Time-series retrieval and scoring use only strictly-past windows relative to each forecast origin, so the retrieval bank never contains the horizon it is used to predict. Single-cell retrieval draws only from the labeled reference split and never from the query cell's own label. This keeps the copy baseline a true copy baseline rather than a transductive shortcut.

4 Experimental Setup

Data. For time series we use the ETT [23] benchmark, forecasting the oil-temperature (OT) target of ETTh1, ETTh2, ETTm1, and ETTm2. The hourly series ETTh1 has 17,420 timesteps and the 15-minute series ETTm1 has 69,680; ETTh1 shows strong daily structure, with autocorrelation 0.941 at the 24-hour lag. For single cells we use the pbmc3k [24] annotation dataset (2,638 cells, 1,838 processed genes, 8 cell-type classes), with a majority-class fraction of 0.434 (CD4 T cells). The single-cell reference and query sets are a stratified 50/50 split (1,318 reference and 1,320 query cells).

Models. We audit three zero-shot foundation models across two modalities: Chronos-Bolt (base and tiny) and TimesFM-2.5-200m for forecasting, and scGPT-human for single-cell annotation. No model is trained or fine-tuned; the audit is evaluation-only.

Protocol. Forecasting uses MASE (lower is better) as the point metric, with a context length of 512, a horizon of 24, and 64 evaluation windows per series; the analog copy uses k=15 neighbors. Single-cell annotation uses macro-F1 (primary) and accuracy, with a k=15 nearest-neighbor vote and a 50-component PCA copy embedding. All runs use seed 0. The complete audit consumed 0.97 GPU-hours against a planned budget of 10, on a single A10-class GPU.

5 Results

Time series: a trivial copy matches most of the model's above-floor skill. Table 1 reports, per series and model, the floor and model MASE, the binding-copy REF with its 95% confidence interval, and which copy is the strongest. The strongest-copy REF is strongly series-dependent, ranging from 0.57 to 0.99, and the mean strongest-copy REF over the four base ETT series for Chronos-Bolt-base is about 0.72. This mean averages across different binding rules (last-value on ETTh1, ETTm1, and ETTm2; seasonal on ETTh2), which are the classical naive baselines. The strongest rule is structure-specific: last-value copying is strongest on the smooth ETTh1, ETTm1, and ETTm2 targets, whereas the seasonal copy is strongest on ETTh2. The analog/kNN copy, the one variant that actually implements retrieval, never binds and is often worse than the floor: its ETTh1 MASE of 1.665 is worse than the floor, and its ETTm1 MASE of 1.094 is worse than last-value. This partly qualifies a "retrieving" reading, since the pure retrieval mechanism is not what matches the model; we report it as part of the band but not as the headline.

On ETTm1 both foundation-model families are statistically indistinguishable from a last-value copy: the binding-copy REF is 0.987 (95% CI [0.906, 1.050]) for Chronos-Bolt-base and 0.999 (95% CI [0.952, 1.046]) for TimesFM-2.5-200m, and both intervals include 1. Near-total parroting is therefore significance-backed, and holds for two independently trained families rather than as a single-model quirk. This equivalence reading is floor-robust: swapping the climatology floor for a seasonal-naive floor leaves the ETTm1 binding REF essentially unchanged (0.984 for Chronos-Bolt and 0.999 for TimesFM), so the parroting conclusion does not hinge on the floor choice. The pattern is series-dependent as a tested claim: on ETTh1 (REF 0.695, 95% CI [0.370, 0.853]), ETTh2 (Chronos 0.634, 95% CI [0.376, 0.778]; TimesFM 0.712, 95% CI [0.405, 0.898]), and ETTm2 (0.574, 95% CI [0.372, 0.697]) the REF confidence interval excludes 1, so the model adds skill beyond the binding copy there. Model size does not change the picture: scaling Chronos-Bolt from tiny to base on ETTh1 leaves the strongest-copy REF essentially unchanged (0.71 versus 0.70).

Single cell: PCA is significantly better on macro-F1 and tied on accuracy. On pbmc3k cell-type annotation (Table 2), a plain PCA-plus-nearest-neighbor copy is at least as good as the scGPT-human foundation model and, on the rare-class-sensitive metric, significantly better: macro-F1 0.905 (95% CI [0.862, 0.936]) for the copy versus 0.813 (95% CI [0.763, 0.865]) for scGPT, against a majority-class floor of 0.076. The scGPT macro-F1 varied run-to-run (0.813 to 0.854) whereas the PCA copy was deterministic at 0.905; we report the 0.813 CI-round value. The macro-F1 gap (PCA minus scGPT) is 0.092 with a 95% CI [0.034, 0.150] that excludes 0, and the binding REF is 1.125 with a 95% CI [1.044, 1.215] that excludes 1, so on macro-F1 the copy is significantly better than the model. On accuracy the two are a statistical tie (0.940 versus 0.936; McNemar exact p=0.59, discordant pairs b=46, c=40). The macro-F1 gap is driven by the rare cell types, to which macro-F1 is sensitive and accuracy is not. All single-cell intervals are from a bootstrap over the 1,320 query cells (4000 resamples) with an exact McNemar test on accuracy. This echoes the "one PCA still rules them all" [5] observation inside the same audit that measured the forecasting models: the copy-baseline pattern is not confined to time series, though a single dataset and split is a narrow basis for any transfer claim.

6 Discussion

Read together, the two modalities give, at the score level, a consistent reading of "is your foundation model learning, or just retrieving?": a cheap copy matches most of the model's skill on these

Table 1: Time-series copy-baseline audit (MASE, lower is better; floor is the climatology mean; L=512, H=24, 64 windows, seed 0). REF is the share of the model's above-floor skill reproduced by each structural copy (Eq. 1); the strongest (binding) copy (bold) is the family member reproducing the most above-floor skill, a maximum-over-family best case. The binding-REF column adds a 95% confidence interval from a moving-block bootstrap (4000 resamples, block 8); a CI that includes 1 means the model is statistically indistinguishable from that copy, and a CI that excludes 1 means the model adds skill beyond it. The Chronos-Bolt-tiny row has no CI from this round and its binding REF is shown as a point estimate. The analog/kNN retrieval copy is omitted from the table because it never binds (it is often worse than the floor) and is reported in text. Dashes mark copy variants not run for that model.

Series	Model	Floor	Model	Binding REF (95% CI)	Binding
ETTh1	Chronos-Bolt-base	1.462	0.819	**0.695** [0.370, 0.853]	last-value
ETTh1	Chronos-Bolt-tiny	1.462	0.831	**0.709** (point est.)	last-value
ETTh2	Chronos-Bolt-base	1.631	0.906	**0.634** [0.376, 0.778]	seasonal
ETTh2	TimesFM-2.5-200m	1.631	0.986	**0.712** [0.405, 0.898]	seasonal
ETTm1	Chronos-Bolt-base	1.409	0.418	**0.987** [0.906, 1.050]	last-value
ETTm1	TimesFM-2.5-200m	1.409	0.430	**0.999** [0.952, 1.046]	last-value
ETTm2	Chronos-Bolt-base	1.475	0.398	**0.574** [0.372, 0.697]	last-value

metrics. Whether the model internally retrieves is a hypothesis this score-level audit cannot settle. In forecasting the copy reproduces roughly 57 to 100 percent of the model's above-floor skill, and on ETTm1 both foundation-model families are statistically indistinguishable from a last-value copy (their binding-REF confidence intervals include 1), while on ETTh1, ETTh2, and ETTm2 the intervals exclude 1 and the model adds skill; in single-cell annotation a 50-component PCA is significantly better than a pretrained transcriptomic model on macro-F1 and tied on accuracy. That the same identically-defined REF is high in two very different domains, under a baseline defined independently of the model scores, is the central point. On the two settings audited here the same pattern appears; we do not claim a general law, and the copy-baseline behavior we report is a score-level match rather than evidence that the model copies internally. The audit also localizes where a model does add skill: ETTh2, where a last-value copy fails (REF 0.06) and the model most clearly clears a persistence copy, is the clearest such case in our set, so the tool is a map of matches-a-copy versus adds-skill rather than a blanket dismissal.

An honest reading tempers the headline. Much of the high REF reflects that the strongest copies are the classical naive baselines (persistence and seasonal-naive), which are hard to beat on persistent, highly autocorrelated series; part of the result therefore restates that naive baselines are strong in this regime. The contribution is the unified, pre-registered, cross-task accounting on one axis, the finding that the audited foundation models barely clear these baselines in the settings tested, and the observation that the pure retrieval copy (the analog/kNN variant) itself fails. Practically, REF is a cheap pre-deployment check, and a model that a last-value copy can match on this point metric is, tentatively and pending probabilistic scores, not obviously worth its inference cost on that task.

Table 2: Single-cell copy-baseline audit on pbmc3k annotation (macro-F1 and accuracy, higher is better; floor is the majority class; PCA(50)+kNN copy, k=15, seed 0). Confidence intervals are from a bootstrap over the 1,320 query cells (4000 resamples); the accuracy comparison uses an exact McNemar test. The copy is significantly better than the scGPT foundation model on macro-F1 (binding REF 1.125, 95% CI [1.044, 1.215] excludes 1; gap 95% CI [0.034, 0.150] excludes 0) and tied on accuracy (McNemar exact p=0.59). The scGPT macro-F1 varied run-to-run (0.813 to 0.854); the CI-round value 0.813 is shown.

Predictor	macro-F1 (95% CI)	Accuracy
Floor (majority class)	0.076	0.433
Copy (PCA(50)+kNN)	**0.905** [0.862, 0.936]	**0.940**
Foundation model (scGPT-human)	0.813 [0.763, 0.865]	0.936
Binding REF (macro-F1)	1.125 [1.044, 1.215]; accuracy tie (McNemar p=0.59)	

7 Limitations

Several boundaries should temper how far these results are read.

Uncertainty. Confidence intervals are now reported, so the small-margin claims are tested rather than hedged. Because the evaluation windows are autocorrelated (ETTh1 target autocorrelation 0.941 at the daily lag), naive per-window intervals would be too tight, so the forecasting CIs use a moving-block bootstrap (4000 resamples, block 8) over well-separated origins; the single-cell CIs use a bootstrap over the 1,320 query cells (4000 resamples) with an exact McNemar test for the accuracy tie. The two previously open small-margin cases are now resolved: on ETTm1 both families' binding-REF intervals include 1 (0.987, 95% CI [0.906, 1.050]; 0.999, 95% CI [0.952, 1.046]), a tested near-equivalence to a last-value copy, while the single-cell copy is significantly above the model on macro-F1 (REF 1.125, 95% CI [1.044, 1.215], gap 95% CI [0.034, 0.150]) and tied on accuracy (McNemar p=0.59). Two caveats remain. REF is floor-dependent in general: it is measured relative to a chosen floor and would shift under a different defensible floor; while the ETTm1 parroting conclusion is floor-robust (binding REF 0.984 to 0.999 under both the climatology and the seasonal-naive floor), under a seasonal-naive floor the ETTh2 last-value copy falls below the floor, so the floor choice still moves individual cells. And the single-cell verdict rests on one dataset and split (see single-cell fairness below).

Floor sensitivity. REF is measured relative to a chosen floor and would shift under a different defensible floor. We ran a partial floor sweep on ETTm1: under a seasonal-naive floor the binding REF is essentially unchanged (0.984 for Chronos-Bolt, 0.999 for TimesFM), so the ETTm1 parroting conclusion is floor-robust. But the floor still moves individual cells: under a seasonal-naive floor the ETTh2 last-value copy falls below the floor, so REF there is floor-dependent, and a full sweep over at least two defensible floors for every series is still deferred. This matters because the strongest copies coincide with the standard naive baselines (persistence and seasonal-naive), so REF partly reflects the distance between a no-structure climatology floor and those naive baselines rather than a property of the model alone.

Single-cell fairness. The single-cell comparison is not representation-fair: the PCA copy uses the curated, HVG-selected log-normalized matrix, whereas scGPT ingests raw counts, so preprocessing is not matched. Moreover pbmc3k is a single donor, so reference and query cells share batch structure, the single-donor split removes scGPT's batch-integration regime, and the

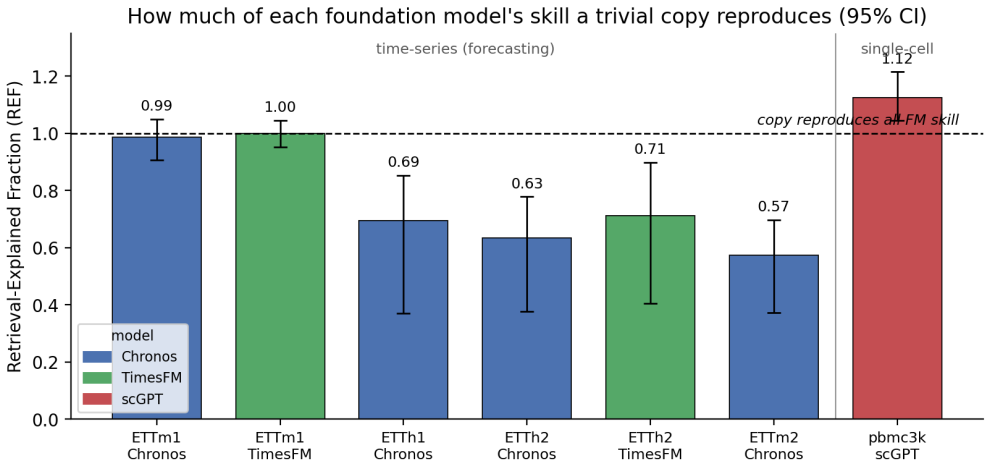

Figure 1: Strongest-copy REF with 95% confidence intervals across the audited (series or dataset, model) pairs in both modalities. Time-series CIs are a moving-block bootstrap (4000 resamples, block 8); the single-cell CI is a cell bootstrap (4000 resamples). Error bars that cross the dashed REF=1 line (both ETTm1 bars) mark models statistically indistinguishable from a last-value copy; bars whose interval clears 1 (ETTh1, ETTh2, ETTm2) mark series where the model adds skill, and the single-cell bar sits significantly above 1 (its interval excludes 1).

donor-grouped leakage-safe split is not exercised; the leakage guarantee we provide is label-level, not batch-level. A matched-preprocessing, multi-donor comparison against a second transcriptomic model is needed before any transfer claim, and the evidence here rests on one dataset (pbmc3k) and one model (scGPT). The defensible claim is "PCA is at least as good as scGPT here" [5], not that the model is useless in general.

Metric. The headline uses point metrics. MASE in particular is normalized to the naive baseline and is structurally sympathetic to persistence-style copies, so a copy-matches result on MASE is partly a metric artifact; probabilistic scores (WQL/CRPS) and, for single cells, clustering-quality measures are deferred, which makes "not worth its inference cost" style readings tentative. A per-item analysis of whether the model and the copy fail on the *same* queries is also deferred and would strengthen the construct validity of REF.

Selection and regime. The headline REF is a maximum over the copy family, a best-case selection statistic rather than an unbiased expected REF. Separately, all four forecasting series are the ETT oil-temperature target, which is persistent and highly autocorrelated; a low-seasonality, low-persistence contrast (for example an exchange-rate or traffic series), where copying should fail and a foundation model should shine, is not yet run, so the high time-series REF is demonstrated only on a favorable regime for copying. We therefore state the result as "a cheap copy matches most of the model's above-floor skill on these metrics in the settings audited here" and hedge any parity, superiority, or mechanistic claim accordingly.

8 Availability

All artifacts are available from the authors on request: the copy-baseline audit harness (a fixed scoring evaluator that computes the floor, copy, and model scores and the REF), the swappable model predictors, and the compute-task configurations used for every reported run. Because the

audit is evaluation-only, there are no model checkpoints or trained weights. The datasets are public: the ETT forecasting benchmark, and pbmc3k as distributed with scanpy.

9 Conclusion and Future Work

We introduced a modality-agnostic copy-baseline audit and the Retrieval-Explained Fraction, and used them to show that, at the score level and on the settings audited here, a trivial retrieve-and-copy baseline matches most of a zero-shot foundation model's above-floor skill, across two modalities and three models, with the small-margin cases now tested: on ETTm1 both forecasting families are statistically indistinguishable from a last-value copy (binding-REF confidence intervals include 1), and in single cells a PCA copy is significantly better than scGPT on macro-F1 and tied on accuracy. The finding is score-level and scoped: the same pattern appears in both settings audited and across two independent forecasting model families, but we do not claim a general law. The ranked next steps are: add a low-persistence forecasting contrast to probe where models genuinely outperform copying; extend the single-cell audit to a batched, donor-grouped dataset and a second foundation model; complete the multi-metric robustness band and the per-item shared-failure analysis; and broaden the harness to further modalities so that any zero-shot model can be audited the same way.

References

[1] Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the Language of Time Series. arXiv:2403.07815 [cs.LG], 2024. URL https://arxiv.org/abs/2403.07815.

[2] Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A Decoder-Only Foundation Model for Time-Series Forecasting. arXiv:2310.10688 [cs.LG], 2023. URL https://arxiv.org/abs/2310.10688.

[3] Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scGPT: Toward Building a Foundation Model for Single-Cell Multi-Omics Using Generative AI. *Nature Methods*, 21, 2024. URL https://doi.org/10.1038/s41592-024-02201-0.

[4] Yuanzhao Zhang and William Gilpin. Context parroting: A simple but tough-to-beat baseline for foundation models in scientific machine learning. arXiv:2505.11349 [cs.LG], 2025. URL https://arxiv.org/abs/2505.11349.

[5] Ihab Bendidi, Shawn Whitfield, Kian Kenyon-Dean, Hanene Ben Yedder, Yassir El Mesbahi, Emmanuel Noutahi, and Alisandra K. Denton. Benchmarking Transcriptomics Foundation Models for Perturbation Analysis: one PCA still rules them all. arXiv:2410.13956 [cs.LG], 2024. URL https://arxiv.org/abs/2410.13956.

[6] William Toner, Thomas L. Lee, Artjom Joosen, Rajkarn Singh, and Martin Asenov. Performance of Zero-Shot Time Series Foundation Models on Cloud Data. arXiv:2502.12944 [cs.LG], 2025. URL https://arxiv.org/abs/2502.12944.

[7] Kutay Tire, Ege Onur Taga, Muhammed Emrullah Ildiz, and Samet Oymak. Retrieval Augmented Time Series Forecasting. arXiv:2411.08249 [cs.LG], 2024. URL https://arxiv.org/abs/2411.08249.

[8] Huiliang Zhang, Ping Nie, Lijun Sun, and Benoit Boulet. Nearest Neighbor Multivariate Time Series Forecasting. arXiv:2505.11625 [cs.LG], 2025. URL https://arxiv.org/abs/2505.11625.

[9] Gabriel M. Mejia, Henry E. Miller, Francis J. A. Leblanc, Bo Wang, Brendan Swain, and Lucas Paulo de Lima Camillo. Diversity by Design: Addressing Mode Collapse Improves scRNA-seq Perturbation Modeling on Well-Calibrated Metrics. arXiv:2506.22641 [q-bio.GN], 2025. URL https://arxiv.org/abs/2506.22641.

[10] Zihao Li, Yuan Cao, Cheng Gao, Yihan He, Han Liu, Jason M. Klusowski, Jianqing Fan, and Mengdi Wang. One-Layer Transformer Provably Learns One-Nearest Neighbor In Context. arXiv:2411.10830 [cs.LG], 2024. URL https://arxiv.org/abs/2411.10830.

[11] Core Francisco Park, Ekdeep Singh Lubana, Itamar Pres, and Hidenori Tanaka. Competition Dynamics Shape Algorithmic Phases of In-Context Learning. arXiv:2412.01003 [cs.LG], 2024. URL https://arxiv.org/abs/2412.01003.

[12] Gautham Bekal, Ashish Pujari, and Scott David Kelly. Continual Learning with Query-Only Attention. arXiv:2510.00365 [cs.LG], 2025. URL https://arxiv.org/abs/2510.00365.

[13] Willa Potosnak, Cristian Challu, Mononito Goswami, Michał Wiliński, Nina Żukowska, and Artur Dubrawski. Implicit Reasoning in Deep Time Series Forecasting. arXiv:2409.10840 [cs.LG], 2024. URL https://arxiv.org/abs/2409.10840.

[14] Taha Aksu, Gerald Woo, Juncheng Liu, Xu Liu, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. GIFT-Eval: A Benchmark For General Time Series Forecasting Model Evaluation. arXiv:2410.10393 [cs.LG], 2024. URL https://arxiv.org/abs/2410.10393.

[15] Yunshi Wen, Wesley M. Gifford, Chandra Reddy, Lam M. Nguyen, Jayant Kalagnanam, and Anak Agung Julius. Revisiting the Generic Transformer: Deconstructing a Strong Baseline for Time Series Foundation Models. arXiv:2602.06909 [cs.LG], 2026. URL https://arxiv.org/abs/2602.06909.

[16] Kavin Soni, Debanshu Das, and Vamshi Guduguntla. Assessing the Operational Viability of Foundation Models for Time Series Forecasting. arXiv:2605.24381 [cs.LG], 2026. URL https://arxiv.org/abs/2605.24381.

[17] Zongzhe Xu, Ritvik Gupta, Wenduo Cheng, Alexander Shen, Junhong Shen, Ameet Talwalkar, and Mikhail Khodak. Specialized Foundation Models Struggle to Beat Supervised Baselines. arXiv:2411.02796 [cs.LG], 2024. URL https://arxiv.org/abs/2411.02796.

[18] Xinyi Wang, Antonis Antoniades, Yanai Elazar, Alfonso Amayuelas, Alon Albalak, Kexun Zhang, and William Yang Wang. Generalization v.s. Memorization: Tracing Language Models' Capabilities Back to Pretraining Data. arXiv:2407.14985 [cs.CL], 2024. URL https://arxiv.org/abs/2407.14985.

[19] Tianxiang Hu, Chenyi Zhou, Jiaxiang Liu, Jiongxin Wang, Ruizhe Chen, Haoxiang Xia, Gaoang Wang, Jian Wu, and Zuozhu Liu. GRIT: Graph-Regularized Logit Refinement for Zero-shot Cell Type Annotation. arXiv:2508.04747 [q-bio.GN], 2025. URL https://arxiv.org/abs/2508.04747.

[20] Haoran Wang, Xuanyi Zhang, Shuangsang Fang, Longke Ran, Ziqing Deng, Yong Zhang, Yuxiang Li, and Shaoshuai Li. Open World Knowledge Aided Single-Cell Foundation Model with Robust Cross-Modal Cell-Language Pre-training. arXiv:2601.05648 [q-bio.GN], 2026. URL https://arxiv.org/abs/2601.05648.

[21] Ihor Kendiukhov. Sparse autoencoders reveal organized biological knowledge but minimal regulatory logic in single-cell foundation models: a comparative atlas of Geneformer and scGPT. arXiv:2603.02952 [q-bio.GN], 2026. URL https://arxiv.org/abs/2603.02952.

[22] Ben Cohen, Emaad Khwaja, Youssef Doubli, Salahidine Lemaachi, Chris Lettieri, Charles Masson, Hugo Miccinilli, Elise Ramé, Qiqi Ren, Afshin Rostamizadeh, Jean Ogier du Terrail, Anna-Monica Toon, Kan Wang, Stephan Xie, Zongzhe Xu, Viktoriya Zhukova, David Asker, Ameet Talwalkar, and Othmane Abou-Amal. This Time is Different: An Observability Perspective on Time Series Foundation Models. arXiv:2505.14766 [cs.LG], 2025. URL https://arxiv.org/abs/2505.14766.

[23] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. arXiv:2012.07436 [cs.LG], 2021. URL https://arxiv.org/abs/2012.07436.

[24] F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis. SCANPY: Large-Scale Single-Cell Gene Expression Data Analysis. *Genome Biology*, 19, 2018. URL https://doi.org/10.1186/s13059-017-1382-0.