A FRAMEWORK AND CHECKLIST FOR VALIDITY THREATS IN FOUNDATION MODEL RESEARCH

Anonymous authors
Paper under double-blind review

ABSTRACT

Controlled experiments are the backbone of machine learning research, but at the scale of modern foundation models, they have become prohibitively expensive. Instead, the community increasingly relies on research strategies that approximate the ideal experiment at a fraction of the cost: proxy experiments and scaling laws, observational studies with publicly available models, and single-run designs that leverage variation within individual training runs. In this work, we show that these research strategies perform causal inference in disguise. Specifically, they aim to measure treatment effects conditional on a specific training recipe. To estimate this treatment effect, efficient research strategies replace various parts of the ideal experiment with cheaper substitutes. Based on this analysis, we draw on concepts from the empirical social sciences and propose a unified framework to identify the implicit assumptions made by different research strategies. We find that each comes with a characteristic validity profile: proxy experiments gain statistical and internal validity at the expense of external and construct validity; observational studies face confounding and effect heterogeneity; and single-run designs are strained by interference between treated units. In case studies with recent papers, we find that these assumptions often receive insufficient attention. To overcome this problem, we propose a checklist that can help authors and reviewers identify threats to the validity of foundation model research designs.

1 INTRODUCTION

With the advent of foundation models, the computational cost of training state-of-the-art machine learning models has increased dramatically (Bommasani et al., 2021; Cottier et al., 2024). In particular, training a frontier language model now requires tens of thousands of GPUs (Brown et al., 2020; Dubey et al., 2024; Sevilla et al., 2024). This trend has profound implications for both model developers, who need to make design decisions at a scale so large that models can only be trained once, and for researchers, who aim to answer research questions about models that they cannot afford to train themselves.

Most research questions about foundation models are interventional: They ask what would happen to model behavior if some aspect of the training recipe (the data, the architecture, the optimizer) were changed. The ideal way to answer such research questions would be a controlled experiment requiring several full training runs (Figure 1). At target scale, meaning for frontier models, such a controlled experiment is, however, computationally infeasible.

To work around this constraint, the community has developed various approximation strategies that study model behavior on a compute budget. For example, model developers use scaling laws to predict optimal hyperparameters from small-scale experiments (Kaplan et al., 2020; Hoffmann et al., 2022; Tao et al., 2024; Bergsma et al., 2025; Shukor et al., 2025), and researchers analyze the meta-data of publicly available models to determine factors that matter for downstream performance (Thrush et al., 2025; Liu et al., 2025). What these strategies have in common is that they estimate the relevant *treatment effect* without performing the ideal experiment. Estimating treatment effects in this way is, however, generally only possible under assumptions (Holland, 1986). For example, when we use experiments on smaller models to make claims about large models, we must assume that the causal effect generalizes across scale. Similarly, when we analyze the meta-data of publicly available models, we must assume that there are no unaccounted confounders.

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

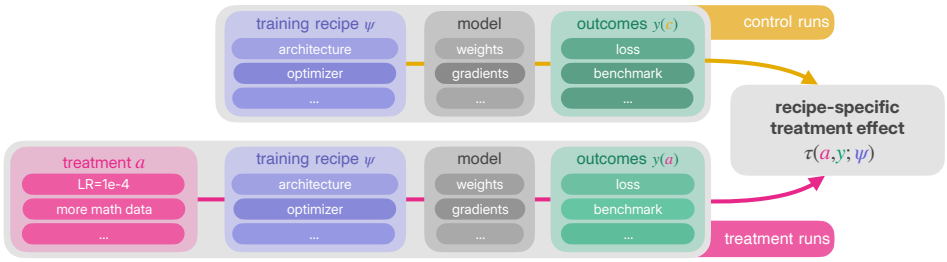

Figure 1: **The ideal but infeasible experiment.** To understand and improve foundation models, one has to answer questions about how changes to a training recipe affect outcomes—for example, how adding math data affects reasoning benchmark scores. Ideally, this **treatment effect** is identified by a controlled experiment with several treatment and control training runs. At frontier scale, the experiment is infeasible and must be approximated, yielding a causal inference problem.

A framework to identify and discuss validity threats. Crucially, violations of necessary assumptions in the estimation of treatment effects pose *threats to the validity* of research results. For example, if the results from smaller proxy experiments do not generalize across scale, then it is invalid to conclude from small-scale experiments about large-scale behavior. For this reason, it is important that the assumptions that underlie different research strategies are *explicitly* discussed. As we show in case studies with recent papers, this is, however, not always the case (Section 6). To overcome this problem, we propose a formal framework that captures the inferential steps that must be taken to draw conclusions about the targeted research question, starting from the experiment that is actually conducted (Section 2). Within this framework, we then discuss the three most important research strategies in the current literature: proxies, observational methods, and single-run approaches (Sections 3–5). For each research strategy, we identify the formal assumptions that are required for the validity of research results. We also derive *validity profiles* that state at an intuitive level which types of assumptions are most at risk in different approaches (the validity profiles are summarized in Appendix Table 1).

Checklist. To operationalize our results, we distill our framework into a checklist for authors and reviewers (Section 6). With the proposed checklist, one can quickly identify the approximation strategies that a given paper employs, and determine whether it appropriately discusses the necessary assumptions that are required for the validity of its research results. In case studies, we then demonstrate that the checklist can be used to identify numerous validity threats that have been missed in a number of top-tier research papers. In summary, our contributions are as follows.

- We propose a **unified framework** (Section 2) that formalizes foundation model research as the attempt to estimate the recipe-specific treatment effect and evaluates research designs according to four types of validity (Shadish et al., 2002): statistical, internal, external, and construct validity.
- We analyze three popular research strategies: the proxy, the observational, and the single-run approach (Sections 3-5). For every strategy, we derive the **formal assumptions that are required for validity** (Appendix C), and derive a characteristic **validity profile** that describes which types of validity are most at risk. We also perform experiments to illustrate key threats.
- We propose a **checklist** that turns validity profiles into concrete questions for authors and reviewers (Section 6). Furthermore, we perform **case studies** of recent papers, showing that assumptions that are required for the validity of research results often receive insufficient attention, and that our checklist can help overcome this problem (Section 6).

Related work is discussed in Appendix A.

2 Foundation Model Research as Causal Inference

In this section, we formalize empirical foundation model research as a causal inference problem and identify four types of validity that every research design must establish.

Training recipe and outcome measures. A foundation model is trained according to a particular **training recipe** ψ, which encompasses all controllable parts of model training (the architecture, the training data, the optimizer, and so on). Up to some uncontrollable random variation, the training

recipe determines the **foundation model** along with a range of **outcome measures** y (including upstream outcomes like the training loss, and downstream outcomes such as benchmark scores). When models share central parts of their recipe, we say that they belong to the same **family**.

Foundation model research as causal inference. We would like to develop a training recipe that yields high-quality outcomes. Suppose that there exists a fixed training recipe ψ that we want to improve. Now, we need to answer questions about how changes in the training recipe would affect the outcomes, for example:

1. Would decreasing the learning rate improve downstream performance?
2. Would adding a data source during mid-training improve reasoning behavior?
3. Would increasing the size of the model improve post-train-ability?

These are *causal* questions: We aim to quantify the effect of a **treatment** a on the outcomes y for a particular recipe ψ. In Definition 1 we make this precise. To do so, we adopt the potential outcomes framework (Rubin, 1974) and write $y(a)$ for the treatment outcome and $y(\emptyset)$ for the control outcome, that is, the outcome for the unmodified recipe ψ. We visualize Definition 1 in Figure 1.

Definition 1 (Recipe-specific treatment effect). Let ψ be a training recipe, a a treatment, and $y(a)$ and $y(\emptyset)$ the treatment and control outcomes. Then the recipe-specific treatment effect is defined as:

$$\tau(a, y; \psi) = \mathbb{E}[y(a) \mid \psi] - \mathbb{E}[y(\emptyset) \mid \psi]\,.$$

At frontier model scale, the ideal experiment that identifies Definition 1 is computationally prohibitive—every practical research design must approximate it, introducing assumptions whose violation can invalidate the results. Formally, for an estimate $\hat{\beta}$ to approximate the recipe-specific treatment effect τ, it must be established that:

$$\hat{\beta} \overset{\text{SV}}{\approx} \beta \overset{\text{IV}}{=} \tau(a', y'; \psi') \overset{\text{EV}}{=} \tau(a', y'; \psi) \overset{\text{CV}}{=} \tau(a, y; \psi). \tag{1}$$

The identities in Equation 1 correspond to the following types of validity:

SV **Statistical Validity**: Each approach must estimate some parameter β from a finite, noisy sample. Statistical validity asks whether the estimate $\hat{\beta}$ reliably reflects the true β.

IV **Internal Validity**: Granted a reliable estimate $\hat{\beta}$, does it reflect a genuine *causal* effect τ? In well-designed experiments, association and causation coincide, however, this is not the case in observational settings. For example, when recent models with a new architecture achieve better outcomes, that does not mean that the architecture caused the improvement.

EV **External Validity**: Granted a genuine causal effect, does it transfer to the recipe ψ that we care about? For example, external validity asks whether the treatment effect for a small model ψ' transfers to the target model ψ.[1]

CV **Construct Validity**: A genuine causal effect τ is only informative if the operationalization of treatment a and outcome y actually reflects the constructs in the research question. For example, changing the treatment from "pre-training on private data" a to "fine-tuning on private data" a' may threaten construct validity.

In the following sections, we discuss three broad research strategies for foundation models and identify which steps in Equation 1 are threatened by each approach.

3 The Proxy Approach

In this section, we begin our discussion of the validity profiles of different research strategies with the proxy approach. In the proxy approach, a controlled experiment is performed, however some part of the experiment—the model, the treatment, or the outcome—is replaced with a proxy to make it computationally feasible (Figure 2).

[1] As later sections show, external validity is not restricted to the transfer from one recipe to another, but also concerns, for example, the transfer from an average effect over a distribution of recipes to the target recipe ψ.

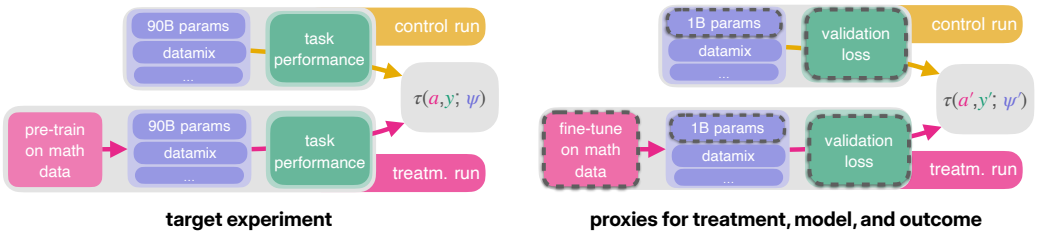

Figure 2: **The proxy approach.** We conduct an experiment but replace one or several parts with proxies (**---**). For example, we conduct the experiment with a proxy model (1B instead of 90B parameters), a proxy treatment (fine-tuning instead of pre-training on math data), or a proxy outcome (validation loss instead of task performance).

3.1 Proxy Models and Scaling Laws

Proxy models replace the target recipe ψ with a proxy recipe ψ' that is cheaper to train. The proxy recipe can have fewer parameters, fewer training tokens (*under-training*; Zhang et al., 2025c; Bordt & Pawelczyk, 2026; O'Brien et al., 2026), or both. The controlled experiment is then performed with ψ', that is we estimate the treatment effect $\tau(a, y; \psi') = \mathbb{E}[y(a) \mid \psi'] - \mathbb{E}[y(\emptyset) \mid \psi']$. Because the proxy model is cheap to train, $\tau(a, y; \psi')$ can be estimated well.

Scaling Laws. In its simplest form, the proxy approach assumes that the treatment effect does not depend on the considered scale, that is, $\tau(a, y; \psi') = \tau(a, y; \psi)$ (Magnusson et al., 2025; Thudi et al., 2025). The more interesting scenario, however, is when the treatment effect depends on the scale in a way that is predictable (Hestness et al., 2017). In this case, treatment effects and outcomes can be estimated with *scaling laws*: Instead of a single proxy recipe, we consider multiple proxy recipes $\psi_1, \ldots, \psi_n$ of various scale $s(\psi_i)$. We then measure $\tau(a, y; \psi_i)$ for every proxy recipe and fit a parametric function f_θ such that $\tau(a, y; \psi_i) \approx f_\theta(s_i)$. The treatment effect at the target scale can be predicted as $f_\theta(s(\psi))$. In practice, f_θ is often fit to outcomes, and treatment effects are the estimated difference between outcomes (as in Figure 3). In recent works, scaling laws have been used to predict the validation loss as a function of model and dataset size (Kaplan et al., 2020; Hoffmann et al., 2022), optimal hyperparameters such as the learning rate and batch size (Bi et al., 2024; Bergsma et al., 2025; Tao et al., 2024), and the effect of the training data mixture (Ye et al., 2024; Shukor et al., 2025). A detailed overview of literature on scaling laws is in Appendix D.1.

Validity Profile. Whether $\tau(a, y; \psi') = \tau(a, y; \psi)$, or whether f_θ extrapolates to the target scale, is a question of *external* validity. We can distinguish between external validity *across scale*, when the proxy is a scaled-down version of the target recipe, and *across training recipes*, when the two also differ in other ways. Importantly, predictability across scale and training recipes can *not* be determined based on experiments with proxy models alone—it has to be verified by conducting additional experiments at a larger scale. As such, the statement that there is a "scaling law" should first of all be seen as a *claim* for the external validity of proxy experiments. Appendix C.3.1 states this assumption formally (Assumption 4).

Proxy Models – Validity Threats

(1) **External validity across scale:** Do the effects estimated with proxy models, or the estimated scaling laws, accurately describe treatments and outcomes at target scale?
(2) **External validity across training recipes:** Do the effects estimated with proxy models, or the estimated scaling laws, also hold for different training recipes?

Experiment. For a concrete example of the success and failure of scaling laws, consider the experiment depicted in Figure 3. We estimate scaling laws for 304M parameter OLMo 2 models (OLMo et al., 2024) across the number of training tokens, fitting on up to 3B tokens, and extrapolating up to 30B tokens (details in Appendix D.2). We consider two different training recipes: One where we train on fresh tokens of unfiltered web data (*unfiltered, fresh*), and one where a small pool of quality-filtered data is increasingly repeated as the token budget grows (*filtered, repeated*). In both cases, we fit a scaling law of the form $L(D) = E + B\,D^{-\beta}$ to the proxy models, where D is the number of training tokens.

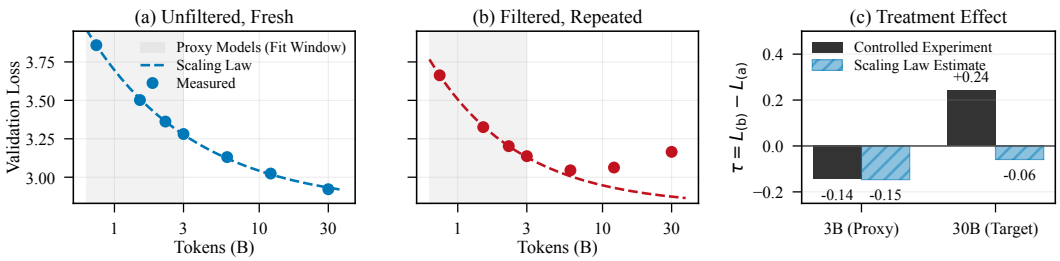

Figure 3: **Success and failure of scaling laws.** (a) and (b) depict the validation loss as we scale the number of training tokens, with a scaling law fit to proxy models of up to 3B tokens (shaded) and extrapolated to 30B tokens. (a) Training on fresh unfiltered tokens (b) Repeatedly training on 0.75B filtered tokens (c) The treatment effect of switching from (a) to (b) at the proxy and target scale, measured and estimated from the two scaling laws. Further details in Appendix D.2.

For the unfiltered, fresh recipe, the scaling law is valid across scale (Figure 3a): The loss extrapolated to 30B tokens is within 0.01 nats of the measured loss. For the filtered, repeated recipe, external validity does not hold (Figure 3b): The estimated scaling law predicts a monotonically decreasing loss, but in fact the loss increases after more than eight repetitions. Taken together, this leads to an invalid estimate of the treatment effect of switching to the filtered, repeated data at the target scale of 30B training tokens (Figure 3c). For a detailed overview of the success and failure of scaling laws in the recent literature, see Appendix D.1.

3.2 Proxy Treatments and Outcomes

Instead of replacing the model with a proxy, we can replace the treatment or outcome. One example of a proxy treatment is to fine-tune a model instead of re-training it from scratch, an approach common in the privacy literature (Maini et al., 2024; Pawelczyk et al., 2025). One example of a proxy outcome is to use a subset of a benchmark to estimate the full benchmark score (Polo et al., 2024; Pacchiardi et al., 2024). Formally, we keep the target recipe ψ but estimate the effect of a proxy treatment a', $\tau(a', y; \psi)$, or the effect on a proxy outcome y', $\tau(a, y'; \psi)$, and use it to approximate $\tau(a, y; \psi)$. Because of space constraints, we defer the majority of the discussion of proxy treatments and outcomes to Appendices D.3 and D.5. Here we provide the validity profile and an experiment.

Validity Profile. Proxy treatments and outcomes threaten *construct* validity. This is because they employ different constructs than the original training recipe, meaning that we risk measuring the *wrong* causal effect $\tau(a', y'; \psi) \neq \tau(a, y; \psi)$. For example, multi-epoch fine-tuning as a proxy for single-epoch pretraining overestimates privacy leakage (see the experiment below and Appendix D.3), and benchmark prediction methods fail for the most performant models (Zhang et al., 2025b) (Appendix D.5). Proxy treatments and outcomes do not threaten *statistical*, *internal*, or *external* validity.

Experiment. To demonstrate the threats of proxy treatments, we show that using fine-tuning as a proxy can severely overestimate privacy risks. Specifically, we compare a target recipe, which pretrains an OLMo2 546M model on 216B tokens with canaries, against a proxy recipe that fine-tunes a clean 546M model on 1B tokens containing the canaries (the setup is similar to Panda et al. (2025)). Figure 4 depicts the result of this experiment. From Figure 4, we see that while the target pretraining maximum AUC reaches 0.768 for heavily repeated random token canaries, the fine-tuning proxy reaches near-perfect memorization (AUC ≈ 1.0) within just two epochs. In other words, the fine-tuning proxy overestimates privacy leakage. Details on this experiment are in Appendix D.4

4 The Observational Approach

The observational approach avoids training new models altogether and instead analyzes the meta-data of existing training runs. Formally, every existing run is one observation (ψ', y) of a recipe and its outcome, where we regard the recipes of the existing runs as drawn from a distribution Π. We write $A \in \{a, \emptyset\}$ for the recipe component that corresponds to the treatment. The observational approach estimates the *association* between treatment and outcome across runs, in the simplest case

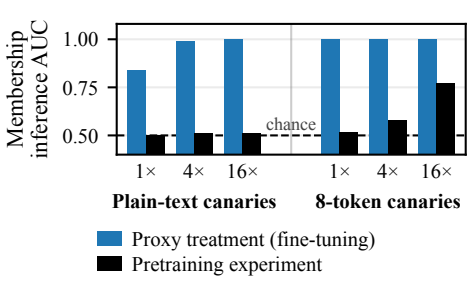

Figure 4: **Proxy treatments distort construct validity.** We visualize the difference between true privacy leakage from pretraining and the overestimated leakage measured via a fine-tuning proxy. Fine-tuning as a substitute misrepresents the underlying construct.

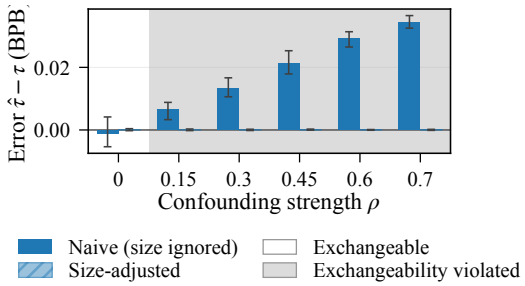

Figure 5: **Confounding threatens the internal validity of observational analyses.** We subsample 1,536 training runs such that model size correlates with data quality (strength ρ). The estimated effect of data quality is then biased unless we adjust for the confounder model size.

$\beta(a, y) = \mathbb{E}[y \mid A = a] - \mathbb{E}[y \mid A = \emptyset]$, and uses the estimate $\hat{\beta}$ as an approximation of $\tau(a, y; \psi)$, that is,

$$\hat{\beta} \overset{\text{SV}}{\approx} \beta(a, y) \overset{\text{IV}}{=} \mathop{\mathbb{E}}_{\psi' \sim \Pi}[\tau(a, y; \psi')] \overset{\text{EV}}{=} \tau(a, y; \psi), \tag{2}$$

where the middle term is the average effect over the existing recipes rather than the recipe-specific effect of Definition 1.

In recent works, the observational approach has been employed to derive scaling laws (Owen, 2024; Ruan et al., 2024; Choshen et al., 2024; Dominguez-Olmedo et al., 2025; Zhang et al., 2025a; Lourie et al., 2025; Ho et al., 2024; Xiao et al., 2025), to improve the pretraining data mix (Thrush et al., 2025), and to predict benchmark performance from fewer samples (Li et al., 2024a; Pacchiardi et al., 2024; Zhang et al., 2025b), to name only a few (Liu et al., 2025; Mertens et al., 2026; Li, 2026).

Internal Validity and Confounding. In a controlled experiment, internal validity holds by design, because we assign the treatment. In observational data, the treatment was chosen by the model developers together with the rest of the recipe. Thus, we have to argue whether the association β equals the average treatment effect over the existing runs (step IV in Equation 2). Therefore, we have to establish several assumptions: exchangeability, positivity, and SUTVA. The most prominent of them is *exchangeability*, $y(a) \perp\!\!\!\perp A$: The runs that received the treatment must not differ systematically from those that did not. Exchangeability can *not* be tested with observational data alone. It has to be justified with experiments, or with knowledge about how model developers choose their recipes, which is rarely available because much of the meta-data is undisclosed. As such, the statement that a recipe component "explains" an outcome across existing models should first of all be seen as a *claim* of exchangeability. Appendix C.2.1 states the three assumptions formally (Assumptions 1–3).

Experiment. Figure 5 illustrates how a violation of exchangeability can bias the observational approach. In the example, the goal is to quantify the effect of data domains on downstream performance by regressing benchmark scores on the data mixes of existing models (Thrush et al., 2025; Liu et al., 2025). We start from 1,536 runs of 30M and 60M models with random data mixes over 100 web domains, a controlled experiment, and emulate observational samples by subsampling such that smaller models are more often observed with high-quality mixes (correlation ρ), mimicking the scarcity of high-quality data at scale (Muennighoff et al., 2023; Hoffmann et al., 2022; Dubey et al., 2024). The model size then confounds the data mix and the outcome. For $\rho = 0.7$, a naive estimator puts the effect of high-quality data at $+0.033$ bits-per-byte instead of -0.001. Adjusting for the confounder model size recovers the experimental estimate (details in Appendix E.2).

Further Threats. Three further steps deserve attention. *External validity*: Even if the average effect over the existing recipes is identified it might not apply to the target recipe (step EV), for example, scaling law coefficients change with the pretraining data composition (Bi et al., 2024, Section 3.3), and data selection conclusions can flip when other hyperparameters are varied (Wang et al., 2025). *Statistical validity* (step SV): As of today, the number of fully documented training runs is small (Thrush et al., 2025; Liu et al., 2025), and at frontier scale one cannot generate more, so the estimate

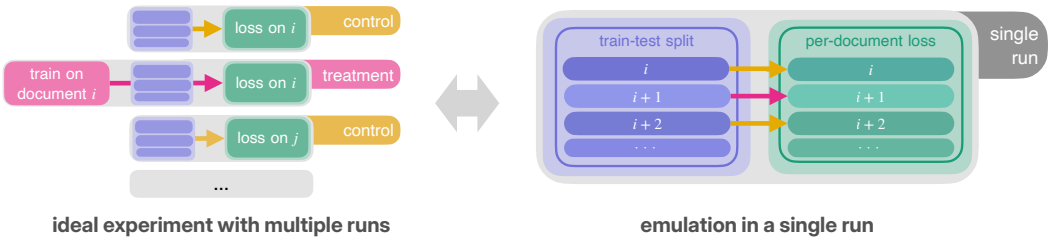

Figure 6: **Analyzing a single training run.** Single-run analysis emulates multiple training runs (left) from one (right). Here, documents are regarded as independent "units", their assignment to the training set as the "treatment", and their respective loss as the "outcome".

$\hat{\beta}$ of β may be too noisy to be conclusive. *Construct validity*: Even though observational analysis per se does not threaten construct validity the treatment and outcome of interest are often not documented. As a result, the analysis often falls back on proxies and then inherits their threats (Section 3.2). Appendix C.3.2 states the external validity assumption formally (Assumption 5).

Validity Profile. The key threat of the observational approach is internal validity, which rests on untestable assumptions. External and statistical validity require careful consideration, and construct validity is affected when the approach is combined with proxies. The observational approach thus faces threats on all fronts. We refer to Appendix E.1 for a more thorough discussion of the threats with examples from the literature.

Observational Approach – Validity Threats

(1) **Internal validity:** Are the runs with and without the treatment exchangeable, or does the association reflect confounding or selection bias?
(2) **External validity across training recipes:** Does the average effect across the pooled model families hold for the target recipe?
(3) **Statistical validity:** Do the few documented runs suffice to estimate the association?

5 The Single-Run Approach

Single-run approaches approximate the treatment effect with a single training run, by emulating several treatment and control runs from it. One such approach, which we from here on call the *within-run* approach, regards the training run as containing a set of independently treatable units with separate outcomes. For example, in data memorization (Feldman & Zhang, 2020; Zhang et al., 2023), each document i in the training data can be regarded as a separate unit, meaning that we can form treatment and control populations in a single run (Figure 6). This has, for example, been leveraged to study data memorization (Lesci et al., 2024), tokenization bias (Lesci et al., 2025), and the effects of pretraining co-occurrence statistics (Elazar et al., 2022). A second approach, the *theoretical control* approach, derives the control outcomes analytically instead; we discuss it in Appendix F.1.

More formally, in the within-run approach we would ideally estimate the effect of applying the treatment a to unit i of the recipe ψ, that is, $\tau(a, y; \psi, i) := \mathbb{E}[y_i(a_i) \mid \psi] - \mathbb{E}[y_i(\emptyset_i) \mid \psi]$, where $y_i(a_i)$ and $y_i(\emptyset_i)$ are the outcomes of unit i if it does and does not receive the treatment. In its simplest form, the within-run approach approximates τ by forming treatment and control groups within the units of a single run and then comparing their average outcomes, formally,

$$\hat{\beta} \overset{\text{SV}}{\approx} \beta(a, y; \psi) \overset{\text{IV}}{=} \mathop{\mathbb{E}}_{i}[\tau(a, y; \psi, i)] \overset{\text{EV}}{=} \tau(a, y; \psi, i), \tag{3}$$

where $\beta(a, y; \psi) = \mathbb{E}[y_i \mid A_i = a, \psi] - \mathbb{E}[y_i \mid A_i = \emptyset, \psi]$ is the association between treatment and outcome across the units of the run (Appendix F).

Internal Validity: Exchangeability and Interference. For the association β to equal the average causal effect $\mathbb{E}_i[\tau(a, y; \psi, i)]$ (step IV), several assumptions must be met; particularly relevant here are exchangeability and no interference. *Exchangeability* requires that the treated and the control units do not differ systematically, formally $(y_i(a_i), y_i(\emptyset_i)) \perp\!\!\!\perp A_i$. Exchangeability fails when the groups are selected differently, for example when deduplication differs between the training and

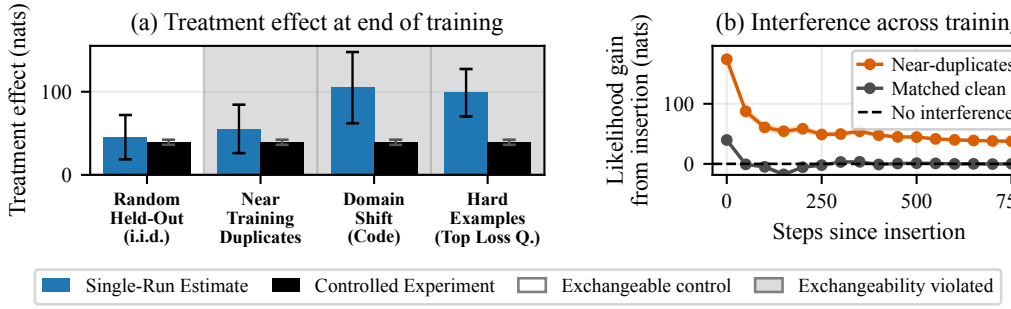

Figure 7: **Violated exchangeability and interference threaten the internal validity of single-run inference.** (a) **Exchangeability**: In the experiment, the within-run estimate overestimates memorization when exchangeability is violated, specifically, when the treated units S are sampled randomly, and the untreated units reflect near training duplicates, a particular domain (code), or especially hard examples. (b) **Interference**: Training on S also changes the outcomes of similar, untreated sequences. The plot shows the effect of training on S on 100 untreated near-duplicates of S and on 100 untreated, unrelated sequences with the same baseline loss; for the near-duplicates, the effect persists until the end of training.

the validation set; difference-in-differences and related estimators replace it with more plausible assumptions (Lesci et al., 2024). *No interference*, part of SUTVA, requires that the outcome of a unit depends on its own assignment only, formally y_i depends on A_i but not on the assignments A_j of the other units. In the within-run approach, this assumption is particularly threatened: all units share the same model parameters, so training on document i changes the parameters that also process document j. Formally, the potential outcome of unit i depends on the entire assignment, $y_i(A_1, \ldots, A_n)$ for the n units of the run, and a control unit that resembles a treated unit is partially treated, which can bias the estimate even under random assignment. For a more formal treatment of these assumptions, we refer to Appendix C.2.1.

Further Threats. *Statistical validity* (step SV) holds as long as the population of units is large. *External validity across units* (step EV): The average effect must transfer to the individual units, or the subpopulation of units, we care about; since the approach averages across units, it does not quantify the effect for any single unit. Appendix C.3.3 states this assumption formally (Assumption 6).

Validity Profile. The key threat of the within-run approach is internal validity, through exchangeability and interference, and neither can be tested from the run alone. External validity across units requires care, whereas external validity across scale and recipes is not at stake as long as the analyzed run is the recipe we care about. Statistical validity is usually less of a concern, because the number of units is large, and construct validity is not threatened by the approach itself, as long as the treatment is applied as stated rather than through a proxy.

Within-Run Approach – Validity Threats

(1) **Internal validity, exchangeability:** Are the treated and control units exchangeable, or were they selected into the training set or the vocabulary in systematically different ways?
(2) **Internal validity, interference:** Do the treated units change the outcomes of the control units, for example through near-duplicates? This cannot be tested from the run alone.
(3) **External validity across units:** Does the average effect across units hold for the individual units, or the subpopulation, of interest?

Experiment. The experiments depicted in Figure 7 illustrate how violations of exchangeability and no interference threaten the validity of the within-run approach for memorization, where the goal is to estimate the effect of training on a text sequence on its likelihood under the trained model.

In Figure 7a, we illustrate the issue of exchangeability. We train an OLMo2 179M model from scratch for 1,500 steps (run T) and an identical run C that differs only at step 750, where 100 sequences S are swapped for 100 others. The units are sequences, the treatment $A_i = a$ is being part of the batch at step 750, and the outcome y_i is the log-likelihood of sequence i at a checkpoint c. Because training is deterministic, for $i \in S$ run T yields $y_i(a_i)$ and run C yields $y_i(\emptyset_i)$, and the ground-truth effect $\tau(a, y; \psi, i) = y_i(a_i) - y_i(\emptyset_i)$ is known exactly, as is its average over S (Appendix F.2). Since

the within-run approach only has access to run T, it takes a control set V of held-out sequences that were never trained on and compares the mean log-likelihood of S with that of V, that is, it estimates $\beta(a, y; \psi)$. The experiment demonstrates that the estimate is unbiased when the treated sequences S and the untreated sequences V do not differ systematically; but when the control set differs systematically, for example because it contains particularly hard sequences, exchangeability is violated, and the estimate is misleading.

In Figure 7b, we turn to interference, that is, to y_j depending on the assignments A_i of other units. Let S' be 100 synthetic near-duplicates of the sequences in S and V 100 unrelated held-out sequences with the same baseline loss, neither trained on in either run, so $A_j = \emptyset$ for all $j \in S' \cup V$. Without interference, y_j would be identical in both runs. Instead, right after S is inserted, the sequences in S' become about 135 nats more likely under run T than under run C, relative to V, and the gap plateaus at about 40 nats until the end of training. Learning a sequence thus permanently makes near-duplicate sequences more likely.

6 Checklist and Case Studies

Checklist. To operationalize the findings from the Sections 3–5, we distill our framework into a checklist for authors and reviewers. The first part of the checklist asks about the research question and identifies the approximation strategies of the central experiment. The second part of the checklist then asks whether the given work *properly discusses* the validity threats that come with the chosen research strategies. The checklist provides guidelines to the user and we aim it to be self-contained: authors and reviewers should be able to work with the proposed checklist without having read this paper. The checklist is inspired by existing approaches to improve research practices, most importantly the NeurIPS paper checklist (Pineau et al., 2021b; Beygelzimer et al., 2021). The full checklist is provided in Appendix H.

Case studies. To demonstrate the practical benefits of the proposed checklist, we perform case studies where we apply the checklist to recent top-tier research papers (Khaddaj et al., 2025; Muennighoff et al., 2023; Dong et al., 2024; Chang et al., 2024; Ruan et al., 2024; Dominguez-Olmedo et al., 2025; Lesci et al., 2024). The case studies are performed from the perspective of a reviewer, which means that we aim to identify the approximation strategies that a given paper uses, and whether the corresponding validity threats are appropriately discussed and addressed. For four papers, we find that at least one validity threat is entirely unaddressed. The full case studies are provided in Appendix G, and we encourage the reader to take a look.

Beyond human reviewers, we hope that the proposed checklist may also help LLM agents to evaluate and review papers. In preliminary experiments, we tested whether Claude Fable 5.1 (Anthropic, 2026) could complete the checklist. When prompted with the papers from the case studies, the model identified the approximation strategies of Part A correctly in all seven cases, but the answers in Part B were less reliable.

7 Discussion

In this work, we provide a unified framework for validity threats in foundation model research designs. In practice, researchers may often want to know how to address particular validity threats in their research designs—can clever experiments, better theoretical controls, or additional data help address the validity threats? One answer to this question is that we can always perform the ideal experiment, and see if results agree (for example, by training a large model we can verify the external validity of a scaling law across scale). There are other answers as well, for example, a scientific literature can over time accumulate knowledge about the practical importance of different threats to different research questions and designs (e.g., we know that scaling laws for the cross-entropy loss after pretraining work pretty well, but that this does not necessarily hold for all other outcomes). It could also be the case that we have theory which supports the validity of the experimental design (Yang et al., 2021; Shulgin et al., 2026). Open-science initiatives like the Marin Project that openly document training recipes and meta-data can also help. At any rate, however, the necessary first step is always to be *aware* of potential validity threats, and to have a common language to discuss them—and this is exactly what we provide with our framework.

Ethics Statement

This work is a conceptual and methodological analysis of research practices. It does not involve human subjects and we do not collect novel datasets. We do not believe that this work raises significant ethical concerns.

Reproducibility Statement

The experiments in this paper rely on the Olmo 2 family of fully open models (OLMo et al., 2024) and on publicly available datasets. We will release our code upon publication.

AI Use Statement

During the writing of the paper, we use generative AI tools to suggest individual sentences and paragraphs, draft figures, and to give feedback on the paper. We have reviewed all AI-assisted suggestions and take full responsibility for the submitted paper. We used Claude and other LLMs to search for related works, but all of these works were then reviewed by us, and no part of the discussion of related works was drafted by LLMs. The experiments in this paper were designed by us, and partly executed by Claude Code. The theory in this paper and the checklist were developed without AI assistance. The style-template for the checklist was generated by Claude Code.

References

Galen Andrew, Peter Kairouz, Sewoong Oh, Alina Oprea, H Brendan McMahan, and Vinith M Suriyakumar. One-shot empirical privacy estimation for federated learning. In *International Conference on Learning Representations*, 2024.

Anthropic. System card: Claude fable 5.1 & claude mythos 5.1. Technical report, Anthropic, 2026. URL https://www-cdn.anthropic.com/0339e6a7c5c7b87f5c07798616dc32c215d14235/Claude%20Fable%205.1%20&%20Claude%20Mythos%205.1%20System%20Card.pdf.

Juhan Bae, Nathan Ng, Alston Lo, Marzyeh Ghassemi, and Roger B Grosse. If influence functions are the answer, then what is the question? *Advances in Neural Information Processing Systems*, 35:17953–17967, 2022.

Juhan Bae, Wu Lin, Jonathan Lorraine, and Roger Grosse. Training data attribution via approximate unrolled differentiation. *arXiv preprint arXiv:2405.12186*, 2024.

Andrew M Bean, Ryan Othniel Kearns, Angelika Romanou, Franziska Sofia Hafner, Harry Mayne, Jan Batzner, Negar Foroutan, Chris Schmitz, Karolina Korgul, Hunar Batra, et al. Measuring what matters: Construct validity in large language model benchmarks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Shane Bergsma, Nolan Dey, Gurpreet Gosal, Gavia Gray, Daria Soboleva, and Joel Hestness. Power lines: Scaling laws for weight decay and batch size in llm pre-training. *arXiv preprint arXiv:2505.13738*, 2025.

Alina Beygelzimer, Yann Dauphin, Percy Liang, and Jennifer Wortman Vaughan. Introducing the NeurIPS 2021 paper checklist. NeurIPS Blog, March 2021. URL https://blog.neurips.cc/2021/03/26/introducing-the-neurips-2021-paper-checklist/. Accessed 2026-09-25.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International conference on machine learning*, pp. 2397–2430. PMLR, 2023.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Sebastian Bordt and Martin Pawelczyk. Train once, answer all: Many pretraining experiments for the cost of one. In *ICLR*, 2026.

Wassim Bouaziz, Nicolas Usunier, and El-Mahdi El-Mhamdi. Data taggants: Dataset ownership verification via harmless targeted data poisoning. 2025.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, et al. Accounting for variance in machine learning benchmarks. *Proceedings of machine learning and systems*, 2021.

Samuel Bowman and George Dahl. What will it take to fix benchmarking in natural language understanding? In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4843–4855, 2021.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 2020.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pp. 2633–2650, 2021.

Hoyeon Chang, Jinho Park, Seonghyeon Ye, Sohee Yang, Youngkyung Seo, Du-Seong Chang, and Minjoon Seo. How do large language models acquire factual knowledge during pretraining? *Advances in neural information processing systems*, 37:60626–60668, 2024.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*, 2024.

Leshem Choshen, Yang Zhang, and Jacob Andreas. A hitchhiker's guide to scaling law estimation. *arXiv preprint arXiv:2410.11840*, 2024.

Miranda Christ, Sam Gunn, Tal Malkin, and Mariana Raykova. Provably robust watermarks for open-source language models. *arXiv preprint arXiv:2410.18861*, 2024.

Ben Cottier, Robi Rahman, Loredana Fattorini, Nestor Maslej, Tamay Besiroglu, and David Owen. The rising costs of training frontier ai models. *arXiv preprint arXiv:2405.21015*, 2024.

Mostafa Dehghani, Yi Tay, Alexey A Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. The benchmark lottery. *arXiv preprint arXiv:2107.07002*, 2021.

Jasper Dekoninck, Mark Müller, and Martin Vechev. Constat: Performance-based contamination detection in large language models. *Advances in Neural Information Processing Systems*, 37: 92420–92464, 2024.

Nolan Dey, Gurpreet Gosal, Zhiming Chen, Hemant Khachane, William Marshall, Ribhu Pathria, Marvin Tom, and Joel Hestness. Cerebras-GPT: Open compute-optimal language models trained on the Cerebras wafer-scale cluster. *arXiv preprint arXiv:2304.03208*, 2023.

Nolan Dey, Bin Claire Zhang, Lorenzo Noci, Mufan Li, Blake Bordelon, Shane Bergsma, Cengiz Pehlevan, Boris Hanin, and Joel Hestness. Don't be lazy: Completep enables compute-efficient deep transformers. *arXiv preprint arXiv:2505.01618*, 2025.

Ricardo Dominguez-Olmedo, Florian E Dorner, and Moritz Hardt. Training on the test task confounds evaluation and emergence. In *International Conference on Learning Representations (ICLR)*, 2025.

Jinshuo Dong, Aaron Roth, and Weijie J Su. Gaussian differential privacy. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(1):3–37, 2022.

Yihong Dong, Xue Jiang, Huanyu Liu, Zhi Jin, Bin Gu, Mengfei Yang, and Ge Li. Generalization or memorization: Data contamination and trustworthy evaluation for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12039–12050, 2024.

Vineeth Dorna, Anmol Mekala, Wenlong Zhao, Andrew McCallum, Zachary C Lipton, J Zico Kolter, and Pratyush Maini. Openunlearning: Accelerating llm unlearning via unified benchmarking of methods and metrics. *arXiv preprint arXiv:2506.12618*, 2025.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pp. arXiv–2407, 2024.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Amir Feder, Abhilasha Ravichander, Marius Mosbach, Yonatan Belinkov, Hinrich Schütze, and Yoav Goldberg. Measuring causal effects of data statistics on language model's "factual" predictions. *arXiv preprint arXiv:2207.14251*, 2022.

Epoch AI. Data on AI models, 3 2026. URL https://epoch.ai/data/ai-models. Accessed: 21 Mar 2026.

Vitaly Feldman and Chiyuan Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. *Advances in Neural Information Processing Systems*, 33:2881–2891, 2020.

Clémentine Fourrier, Nathan Habib, Alina Lozovskaya, Konrad Szafer, and Thomas Wolf. Open llm leaderboard v2. https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard, 2024.

Timo Freiesleben and Sebastian Zezulka. The benchmarking epistemology: Construct validity for evaluating machine learning models. *arXiv preprint arXiv:2510.23191*, 2025.

Ryan Giordano, William Stephenson, Runjing Liu, Michael Jordan, and Tamara Broderick. A swiss army infinitesimal jackknife. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1139–1147. PMLR, 2019.

Sachin Goyal, Pratyush Maini, Zachary C Lipton, Aditi Raghunathan, and J Zico Kolter. Scaling laws for data filtering–data curation cannot be compute agnostic. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22702–22711, 2024.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. Olmo: Accelerating the science of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15789–15809, 2024.

Roger Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, et al. Studying large language model generalization with influence functions. *arXiv preprint arXiv:2308.03296*, 2023.

Moritz Haas, Sebastian Bordt, Ulrike von Luxburg, and Leena Chennuru Vankadara. On the surprising effectiveness of large learning rates under standard width scaling. *arXiv preprint arXiv:2505.22491*, 2025.

Frank R Hampel. The influence curve and its role in robust estimation. *Journal of the american statistical association*, 69(346):383–393, 1974.

Tessa Han, Sebastian Bordt, Hanlin Zhang, and Sham Kakade. Weight decay improves language model plasticity. *arXiv preprint arXiv:2602.11137*, 2026.

Moritz Hardt. The emerging science of machine learning benchmarks. *Manuscript. https://mlbenchmarks.org*, 2025.

Miguel A Hernán and James M Robins. Using big data to emulate a target trial when a randomized trial is not available. *American journal of epidemiology*, 183(8):758–764, 2016.

Moritz Herrmann, F Julian D Lange, Katharina Eggensperger, Giuseppe Casalicchio, Marcel Wever, Matthias Feurer, David Rügamer, Eyke Hüllermeier, Anne-Laure Boulesteix, and Bernd Bischl. Position: Why we must rethink empirical research in machine learning. In *International Conference on Machine Learning (ICML)*, 2024.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

Anson Ho, Tamay Besiroglu, Ege Erdil, David Owen, Robi Rahman, Zifan C Guo, David Atkinson, Neil Thompson, and Jaime Sevilla. Algorithmic progress in language models. *Advances in Neural Information Processing Systems*, 37:58245–58283, 2024.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Paul W Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81 (396):945–960, 1986.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*, 2024.

Abigail Z. Jacobs and Hanna Wallach. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 375–385. ACM, 2021. doi: 10.1145/3442188.3445901.

Matthew Jagielski, Om Thakkar, Florian Tramer, Daphne Ippolito, Katherine Lee, Nicholas Carlini, Eric Wallace, Shuang Song, Abhradeep Guha Thakurta, Nicolas Papernot, and Chiyuan Zhang. Measuring forgetting of memorized training examples. In *ICLR*, 2023.

Keller Jordan, Yuchen Jin, Vlado Boza, You Jiacheng, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. *URL https://kellerjordan. github. io/posts/muon*, 2024.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Amit Keinan, Moshe Shenfeld, and Katrina Ligett. How well can differential privacy be audited in one run? *arXiv preprint arXiv:2503.07199*, 2025.

Alaa Khaddaj, Logan Engstrom, and Aleksander Madry. Small-to-large generalization: Training data influences models consistently across scale. In *International Conference on Learning Representations*, volume 2025, pp. 33729–33768, 2025.

Konwoo Kim, Suhas Kotha, Percy Liang, and Tatsunori Hashimoto. Pre-training under infinite compute. *arXiv preprint arXiv:2509.14786*, 2025.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pp. 1885–1894. PMLR, 2017.

Woosung Koh, Juyoung Suk, Sungjun Han, Se-Young Yun, and Jamin Shin. Predicting llm reasoning performance with small proxy model. 2026.

Jakub Krajewski, Amitis Shidani, Dan Busbridge, Sam Wiseman, and Jason Ramapuram. Revisiting the scaling properties of downstream metrics in large language model training. *arXiv preprint arXiv:2512.08894*, 2025.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.

Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. Gaussian membership inference privacy. *Advances in Neural Information Processing Systems*, 36:73866–73878, 2023.

Pietro Lesci, Clara Meister, Thomas Hofmann, Andreas Vlachos, and Tiago Pimentel. Causal estimation of memorisation profiles. *arXiv preprint arXiv:2406.04327*, 2024.

Pietro Lesci, Clara Meister, Thomas Hofmann, Andreas Vlachos, and Tiago Pimentel. Causal estimation of tokenisation bias. *arXiv preprint arXiv:2506.03149*, 2025.

Bojie Li. Incompressible knowledge probes: Estimating black-box llm parameter counts via factual capacity, 2026. URL https://arxiv.org/abs/2604.24827.

Yang Li, Jie Ma, Miguel Ballesteros, Yassine Benajiba, and Graham Horwood. Active evaluation acquisition for efficient llm benchmarking. *arXiv preprint arXiv:2410.05952*, 2024a.

Zhe Li, Wei Zhao, Yige Li, and Jun Sun. Do influence functions work on large language models. *arXiv preprint arXiv:2409.19998*, 3, 2024b.

Thomas Liao, Rohan Taori, Inioluwa Deborah Raji, and Ludwig Schmidt. Are we learning yet? a meta review of evaluation failures across machine learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

Emmy Liu, Amanda Bertsch, Lintang Sutawika, Lindia Tjuatja, Patrick Fernandes, Lara Marinov, Michael Chen, Shreya Singhal, Carolin Lawrence, Aditi Raghunathan, et al. Not-just-scaling laws: Towards a better understanding of the downstream impact of language model design decisions. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 16407–16438, 2025.

Shayne Longpre, Sneha Kudugunta, Niklas Muennighoff, I-Hung Hsu, Isaac Caswell, Alex Pentland, Sercan Arik, Chen-Yu Lee, and Sayna Ebrahimi. Atlas: Adaptive transfer scaling laws for multilingual pretraining, finetuning, and decoding the curse of multilinguality. In *ICLR*, 2026.

Nicholas Lourie, Michael Y Hu, and Kyunghyun Cho. Scaling laws are unreliable for downstream tasks: A reality check. *arXiv preprint arXiv:2507.00885*, 2025.

Anton Lozhkov, Loubna Ben Allal, Leandro von Werra, and Thomas Wolf. Fineweb-edu: the finest collection of educational content, 2024. URL https://huggingface.co/datasets/HuggingFaceFW/fineweb-edu.

Lovish Madaan, Aaditya K. Singh, Rylan Schaeffer, Andrew Poulton, Sanmi Koyejo, Pontus Stenetorp, Sharan Narang, and Dieuwke Hupkes. Quantifying variance in evaluation benchmarks. *arXiv preprint arXiv:2406.10229*, 2024.

Ian Magnusson, Nguyen Tai, Ben Bogin, David Heineman, Jena D Hwang, Luca Soldaini, Akshita Bhagia, Jiacheng Liu, Dirk Groeneveld, Oyvind Tafjord, et al. Datadecide: How to predict best pretraining data with small experiments. *arXiv preprint arXiv:2504.11393*, 2025.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C Lipton, and J Zico Kolter. Tofu: A task of fictitious unlearning for llms. *arXiv preprint arXiv:2401.06121*, 2024.

Prasanna Mayilvahanan, Thaddäus Wiedemer, Sayak Mallick, Matthias Bethge, and Wieland Brendel. Llms on the line: Data determines loss-to-loss scaling laws. In *ICML*, 2025.

Matthias Mertens, Natalia Fischl-Lanzoni, and Neil Thompson. Is there "secret sauce" in large language model development? *arXiv preprint arXiv:2602.07238*, 2026.

Evan Miller. Adding error bars to evals: A statistical approach to language model evaluations. *arXiv preprint arXiv:2411.00640*, 2024.

Bruno Mlodozeniec, Isaac Reid, Sam Power, David Krueger, Murat Erdogdu, Richard E Turner, and Roger Grosse. Distributional training data attribution: What do influence functions sample? 2025.

Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A Raffel. Scaling data-constrained language models. *Advances in Neural Information Processing Systems*, 36:50358–50376, 2023.

Kyle O'Brien, Stephen Casper, Quentin Anthony, Tomek Korbak, Robert Kirk, Xander Davies, Ishan Mishra, Geoffrey Irving, Yarin Gal, and Stella Biderman. Deep ignorance: Filtering pretraining data builds tamper-resistant safeguards into open-weight llms. In *ICLR 2026*, 2026.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, et al. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. *NeurIPS*, 2022.

David Owen. How predictable is language model benchmark performance? *arXiv preprint arXiv:2401.04757*, 2024.

Lorenzo Pacchiardi, Lucy G Cheke, and José Hernández-Orallo. 100 instances is all you need: predicting the success of a new llm on unseen data by testing on a few instances. *arXiv preprint arXiv:2409.03563*, 2024.

Ashwinee Panda, Xinyu Tang, Milad Nasr, Christopher A Choquette-Choo, and Prateek Mittal. Privacy auditing of large language models. In *ICLR*, 2025.

Martin Pawelczyk, Jimmy Z Di, Yiwei Lu, Gautam Kamath, Ayush Sekhari, and Seth Neel. Machine unlearning fails to remove data poisoning attacks. In *International Conference on Learning Representations (ICLR)*, 2025.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program). *Journal of machine learning research*, 22(164):1–20, 2021a.

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. Improving reproducibility in machine learning research (a report from the NeurIPS 2019 reproducibility program). *Journal of Machine Learning Research*, 22(164):1–20, 2021b. URL http://jmlr.org/papers/v22/20-303.html.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinybenchmarks: evaluating llms with fewer examples. *arXiv preprint arXiv:2402.14992*, 2024.

Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. AI and the everything in the whole wide world benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

Anka Reuel, Amelia Hardy, Chandler Smith, Max Lamparth, Malcolm Hardy, and Mykel J. Kochenderfer. BetterBench: Assessing AI benchmarks, uncovering issues, and establishing best practices. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Yangjun Ruan, Chris J Maddison, and Tatsunori B Hashimoto. Observational scaling laws and the predictability of langauge model performance. *Advances in Neural Information Processing Systems*, 37:15841–15892, 2024.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, and Hervé Jégou. Radioactive data: tracing through training. In *International Conference on Machine Learning*, pp. 8326–8335. PMLR, 2020.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? *Advances in neural information processing systems*, 36:55565–55581, 2023.

Andrea Schioppa, Katja Filippova, Ivan Titov, and Polina Zablotskaia. Theoretical and practical perspectives on what influence functions do. *Advances in Neural Information Processing Systems*, 36:27560–27581, 2023.

Jaime Sevilla, Tamay Besiroglu, Ben Cottier, Josh You, Edu Roldán, Pablo Villalobos, and Ege Erdil. Can AI scaling continue through 2030? https://epoch.ai/blog/can-ai-scaling-continue-through-2030, 2024. Epoch AI.

William R. Shadish, Thomas D. Cook, and Donald T. Campbell. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton Mifflin, 2002.

Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A Smith, and Chiyuan Zhang. Muse: Machine unlearning six-way evaluation for language models. *arXiv preprint arXiv:2407.06460*, 2024.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pp. 3–18. IEEE, 2017.

Mustafa Shukor, Louis Bethune, Dan Busbridge, David Grangier, Enrico Fini, Alaaeldin El-Nouby, and Pierre Ablin. Scaling laws for optimal data mixtures. *arXiv preprint arXiv:2507.09404*, 2025.

Egor Shulgin, Dimitri von Rütte, Tianyue H Zhang, Niccolò Ajroldi, Bernhard Schölkopf, and Antonio Orvieto. Deriving hyperparameter scaling laws via modern optimization theory. *arXiv preprint arXiv:2603.15958*, 2026.

Shivalika Singh, Yiyang Nan, Alex Wang, Daniel D'souza, Sayash Kapoor, Ahmet Üstün, Sanmi Koyejo, Yuntian Deng, Shayne Longpre, Noah A Smith, et al. The leaderboard illusion. *arXiv preprint arXiv:2504.20879*, 2025.

Liwei Song and Prateek Mittal. Systematic evaluation of privacy risks of machine learning models. In *30th USENIX security symposium (USENIX security 21)*, pp. 2615–2632, 2021.

Thomas Steinke, Milad Nasr, and Matthew Jagielski. Privacy auditing with one (1) training run. *Advances in Neural Information Processing Systems*, 36:49268–49280, 2023.

Jasper Tan, Daniel LeJeune, Blake Mason, Hamid Javadi, and Richard G Baraniuk. A blessing of dimensionality in membership inference through regularization. In *International Conference on Artificial Intelligence and Statistics*, pp. 10968–10993. PMLR, 2023.

Chaofan Tao, Qian Liu, Longxu Dou, Niklas Muennighoff, Zhongwei Wan, Ping Luo, Min Lin, and Ngai Wong. Scaling laws with vocabulary: Larger models deserve larger vocabularies. *Advances in Neural Information Processing Systems*, 37:114147–114179, 2024.

Tristan Thrush, Christopher Potts, and Tatsunori Hashimoto. Improving pretraining data using perplexity correlations. In *International Conference on Learning Representations*, 2025.

Anvith Thudi, Evianne Rovers, Yangjun Ruan, Tristan Thrush, and Chris J Maddison. Mixmin: Finding data mixtures via convex minimization. *arXiv preprint arXiv:2502.10510*, 2025.

Rajan Vivek, Kawin Ethayarajh, Diyi Yang, and Douwe Kiela. Anchor points: Benchmarking models with much fewer examples. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1576–1601, 2024.

Hanna Wallach, Meera Desai, A Feder Cooper, Angelina Wang, Chad Atalla, Solon Barocas, Su Lin Blodgett, Alexandra Chouldechova, Emily Corvi, P Alex Dow, et al. Position: Evaluating generative ai systems is a social science measurement challenge. In *International Conference on Machine Learning (ICML)*, 2025.

Jiachen T Wang, Tong Wu, Kaifeng Lyu, James Zou, Dawn Song, Ruoxi Jia, and Prateek Mittal. Can small training runs reliably guide data curation? rethinking proxy-model practice. *arXiv preprint arXiv:2512.24503*, 2025.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

Alexander Wettig, Kyle Lo, Sewon Min, Hannaneh Hajishirzi, Danqi Chen, and Luca Soldaini. Organize the web: Constructing domains enhances pre-training data curation. In *International Conference on Machine Learning*, 2025.

Parker Whitfill. Note on selection bias in observational estimates of algorithmic progress. *arXiv preprint arXiv:2508.11033*, 2025.

Chaojun Xiao, Jie Cai, Weilin Zhao, Biyuan Lin, Guoyang Zeng, Jie Zhou, Zhi Zheng, Xu Han, Zhiyuan Liu, and Maosong Sun. Densing law of llms. *Nature Machine Intelligence*, 7(11): 1823–1833, 2025.

Ge Yang, Edward Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tuning large neural networks via zero-shot hyperparameter transfer. *Advances in Neural Information Processing Systems*, 34:17084–17097, 2021.

Jiasheng Ye, Peiju Liu, Tianxiang Sun, Jun Zhan, Yunhua Zhou, and Xipeng Qiu. Data mixing laws: Optimizing data mixtures by predicting language modeling performance. *arXiv preprint arXiv:2403.16952*, 2024.

Chiyuan Zhang, Daphne Ippolito, Katherine Lee, Matthew Jagielski, Florian Tramèr, and Nicholas Carlini. Counterfactual memorization in neural language models. *Advances in Neural Information Processing Systems*, 36:39321–39362, 2023.

Guanhua Zhang, Ricardo Dominguez-Olmedo, and Moritz Hardt. Train-before-test harmonizes language model rankings. *arXiv preprint arXiv:2507.05195*, 2025a.

Guanhua Zhang, Florian E Dorner, and Moritz Hardt. How benchmark prediction from fewer data misses the mark. *arXiv preprint arXiv:2506.07673*, 2025b.

Hugh Zhang, Jeff Da, Dean Lee, Vaughn Robinson, Catherine Wu, Will Song, Tiffany Zhao, Pranav Raja, Charlotte Zhuang, Dylan Slack, et al. A careful examination of large language model performance on grade school arithmetic. *Advances in Neural Information Processing Systems*, 37: 46819–46836, 2024a.

Rongting Zhang, Martin Andres Bertran, and Aaron Roth. Order of magnitude speedups for LLM membership inference. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4431–4443, 2024b.

Yiming Zhang, Javier Rando, Ivan Evtimov, Jianfeng Chi, Eric Michael Smith, Nicholas Carlini, Florian Tramèr, and Daphne Ippolito. Persistent pre-training poisoning of llms. In *ICLR*, 2025c.

Table 1: **Validity profiles of different research strategies for foundation models.** This table summarizes the discussion of validity profiles in Sections 3-5 at an intuitive level. Each cell encodes how the respective validity type (column, Section 2) fares under the research strategy (row, Sections 3–5). Each cell summarizes how the strategy affects the corresponding validity type: ● rather not affected; ● careful consideration required; ● key threat.

Strategy	Statistical	Internal	External	Construct
Proxy approach *(Section 3)*				
Proxy models	●	●	●	●
Proxy treatments and outcomes	●	●	●	●
Observational approach *(Section 4)*	●	●	●	●
Single-run approach *(Section 5)*				
Within-run control outcomes	●	●	●	●
Theoretical control outcomes	●	●	●	●

A Related Work

Traditionally, machine learning relied on the ability to conduct (repeated) controlled experiments at the target scale. Within this paradigm, discussions of validity have usually concentrated on the *outcome end* of the pipeline. In particular, a large number of works discuss the (construct) validity of benchmarks (Jacobs & Wallach, 2021; Raji et al., 2021; Bowman & Dahl, 2021; Dehghani et al., 2021; Wallach et al., 2025; Freiesleben & Zezulka, 2025; Bean et al., 2025), the (statistical) validity of model comparisons (Bouthillier et al., 2021; Miller, 2024; Madaan et al., 2024; Herrmann et al., 2024), and more generally the evaluation of machine learning systems (Liao et al., 2021; Pineau et al., 2021a; Reuel et al., 2024; Hardt, 2025). The validity challenges discussed in these works are acutely relevant for foundation model research and often persist even when we are able to perform the ideal experiment. As such, they are mostly orthogonal to the challenges that we discuss in this work.

Taken one at a time, several of the validity threats that we discuss in this paper have been raised before. Among the most-examined threats in the existing literature is the external validity of scaling laws (Choshen et al., 2024), which has been examined for downstream tasks and reasoning benchmarks (Lourie et al., 2025; Koh et al., 2026), for data selection (Wang et al., 2025; Khaddaj et al., 2025), and across training recipes (Muennighoff et al., 2023; Kim et al., 2025; Mayilvahanan et al., 2025; Longpre et al., 2026). The construct validity of fine-tuning as a stand-in for pretraining has similarly been questioned by prior works (Song & Mittal, 2021; Dong et al., 2022; Leemann et al., 2023; Tan et al., 2023; Zhang et al., 2024b). With respect to observational studies, confounding by training on the test task (Dominguez-Olmedo et al., 2025) and selection bias (Singh et al., 2025; Whitfill, 2025) have been documented, among others. For proxy outcomes, Zhang et al. (2025b) show the failure of benchmark prediction on the most performant models. For theoretical controls, the gap between influence functions and retraining has been discussed (Bae et al., 2022; Schioppa et al., 2023), and for within-run designs, interference between the units of a single training run is discussed in Bordt & Pawelczyk (2026).

What sets this work apart from previous discussions of validity threats in foundation model research designs is that previous works are usually limited to a single method and setting, for example, the validity of fine-tuning as a proxy for pretraining in privacy settings. In contrast, what this work offers is a comprehensive framework in which many different approximation strategies can be reasoned about at the same time. Our framework also applies to combinations of strategies and to strategies that have yet to be developed. To the best of our knowledge, this paper offers the first comprehensive treatment of the general problem of the validity challenges of efficient foundation model research designs.

B Glossary

Table 2 summarizes the central notation used throughout the paper; this section gives intuitive definitions for the most important terms.

Table 2: Overview of the notation used in the paper.

Notation	Terminology	Meaning
ψ	training recipe	all controllable parts of training (architecture, optimizer, training data, ...)
y	outcome	any property we care about (training loss, benchmark, downstream performance, ...)
a	treatment	a change to the recipe
τ	treatment effect	change in outcome induced by the treatment
$\tau(a, y; \psi)$	recipe-specific treatment effect	see Definition 1
ψ'	proxy recipe	e.g. with smaller model size
a'	proxy treatment	e.g. fine-tuning instead of pre-training
y'	proxy outcome	e.g. a cheap benchmark approximation
β	statistical estimand	a parameter of the data distribution

Causal effect (τ). The change in the outcome induced by the treatment, holding the rest of the recipe fixed. The causal effect is what we ultimately care about—classically obtained by means of a controlled experiment, and increasingly by means of an approximation strategy (Sections 3–5). Eventually we are interested in the recipe-specific average treatment effect (Definition 1). When we want to emphasize the inferential role of τ in contrast to the data-distributional β below, we call it the *causal estimand*.

Causal identification. The process of translating a *causal estimand* (the treatment effect τ) into a statistical estimand (in the best case, β). To do so, causal identification must rely on assumptions about the underlying causal structure. We explain one example in Appendix C.2.

Confounding. A non-causal statistical dependence between treatment and outcome induced by a common cause of both. Confounding is one mechanism through which exchangeability (Assumption 1) can fail; we discuss it in Section 4 and illustrate it in Figure 8.

Ideal experiment. The hypothetical experiment that identifies the causal effect of interest by training many models with and without the treatment and comparing average outcomes. At frontier scale this is infeasible because we cannot afford to train many models; Sections 3–5 discuss the approximation strategies that researchers use instead.

Proxy model (ψ'). A model trained according to a scaling recipe ψ' that follows the original training recipe ψ but uses substantially less compute (Section 3.1).

Proxy treatment (a') and proxy outcome (y'). Computationally cheaper substitutes for the target treatment or outcome (Section 3.2). A proxy treatment may be qualitatively different from the original (e.g. fine-tuning in place of pre-training); a proxy outcome may be an approximation of the target (e.g. a benchmark subset) or altogether different (e.g. pretraining loss in place of post-training benchmark scores).

Scaling recipe. A specification of how the training recipe changes as it is scaled up or down: how the learning rate, architecture, and other hyperparameters are adjusted as the model and training-data sizes change. The GPT-3 paper provides a canonical example (Brown et al., 2020, Table 2.1); many scaling-law papers use slightly customized recipes (Kaplan et al., 2020; Hoffmann et al., 2022).

Selection bias. A non-causal statistical dependence between treatment and outcome induced by the process by which observations enter the sample, e.g. the decision to publish a model. In line with recent causal inference literature (Hernán & Robins, 2016), we use "selection bias" in the narrow sense where conditioning on the selection introduces a noncausal association between treatment and outcome; see Appendix C.2 for an extended discussion and Figure 8 for the corresponding causal graph.

Statistical estimand (β). A parameter of the observable data distribution—e.g. a conditional expectation, a regression coefficient, a difference of group means.

Training recipe (ψ). The collection of all controllable parts of model training—architecture, optimizer, learning-rate schedule, training data, random seeds, and so on. The training recipe contains everything required to replicate an *individual* training run.

Treatment (a) and outcome (y). A treatment is the modification of interest—typically a change to the training recipe, e.g. adding math data to the pre-training mix or increasing the learning rate. The outcome is the property we wish to influence—e.g. benchmark performance, training loss, or a privacy metric. We adopt the potential-outcomes framework (Rubin, 1974) to formalize the outcomes under alternative treatments (but equivalent formulations using Pearl's do-operator are possible (Pearl, 2009)).

Validity. Conclusions can become invalid for various reasons. Based on a framework from the causal inference literature in the social sciences (Shadish et al., 2002) we distinguish four types of validity: *statistical*, *internal*, *external*, and *construct* (Section 2). A more formal introduction of the validity types is provided in Appendix C.

Validity-profile bullets (●, ●, ●). In Sections 3–5 and Table 1 we mark each (strategy, validity type) pair with a coloured bullet: ● indicates that the strategy per se does not threaten the validity type; ● that the strategy raises threats requiring careful consideration in practice; and ● that the strategy poses a key threat to the validity type.

C Extended Discussion of the Four Validity Types

In this section we extend our discussion of the four validity types; in particular, we discuss the assumptions required for internal validity in more detail, and formally distinguish between different forms of external validity and construct validity.

C.1 Statistical Validity

Every research strategy eventually approximates some population parameter β from a finite sample. The form of β depends on the strategy: in a controlled experiment with one treatment and one control run, the estimate $\hat{\beta}$ could be the difference between the two outcomes, and the estimand the *expected* difference between the two outcomes. In an observational analysis, the estimate may be a regression coefficient $\hat{\beta}$, and the estimand the 'true' population parameter. In all cases, statistical validity asks whether the estimate $\hat{\beta}$ from the finite sample reliably reflects the underlying population parameter,

$$\hat{\beta} \approx \beta.$$

This approximation can rely on strong assumptions; for example a linear model assumes a linear relationship between treatment and outcome. Since this is standard knowledge in the field, we do not discuss these assumptions in more detail here.

C.2 Internal Validity

Suppose we have estimated β accurately—for example, we may have found a strong association between using a new architecture and some benchmark score. The question for internal validity is whether this association reflects a genuine causal effect, or, slightly more formally, whether

$$\beta = \tau.$$

In a well-designed controlled experiment, $\beta = \tau$ typically holds *by design*. Ablations, hyperparameter sweeps, and grid searches randomize over the runs that produce treatment and control groups, so systematic differences between the groups vanish in expectation, and the difference of group means identifies the causal effect. [2]

C.2.1 Internal Validity in Section 4

In observational data, the equality $\beta = \tau$ must be *argued for* using a set of structural assumptions about the data-generating process, a step known as *causal identification* (Rubin, 1974; Pearl, 2009). A rich literature provides identification tools—the back-door and front-door criteria, instrumental

[2] As we make formal below, this is because the identifying assumptions—exchangeability, positivity, and SUTVA—are usually satisfied by construction in such experiments.

variables, difference-in-differences, regression discontinuity, do-calculus—each requiring different assumptions and applicable in different settings. Most designs in the observational approach (Section 4) take the simplest stance and assume that the parameters learned by a treatment-outcome association model are themselves causal: that the pairwise association between a and y already coincides with the causal effect. Therefore, three assumptions must hold: exchangeability, positivity, and SUTVA. Since these assumptions did not receive attention in the primary literature, we introduce them more formally below.

Exchangeability. Exchangeability (Assumption 1) requires that the units that receive a particular treatment do not differ systematically from those that do not, so the difference of group means is causally interpretable. The two canonical examples by which exchangeability fails are confounding (a common cause of a and y) and selection bias (conditioning on a collider); both are illustrated in Figure 8.

We illustrated the case of confounding in detail in Section 4. As follows we discuss *selection bias*. While confounding concerns the data-generating process itself, selection bias arises from the process by which observations are *included* in the sample.[3] For example, Singh et al. (2025) show that some labs evaluate many model variants on private benchmark data before publishing the best score. Conditioning on the implicit "decision to publish" opens a non-causal path between provider and benchmark score: a high observed score is more likely to come from a provider that could privately test several variants beforehand.

In Assumption 1 we formally introduce exchangeability. In the simplest case there are two potential outcomes, one for the treatment a and one for the control $\emptyset$; this is the case that we discuss in the main text. Below we use more general definitions that allow a range of possible treatments $a \in \mathcal{A}$.

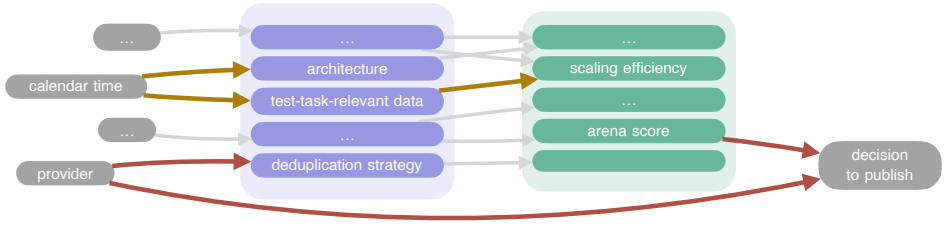

Figure 8: **Confounding and selection bias in the observational approach.** In observational data, recipe components and outcomes can be statistically dependent via non-causal paths. *Confounders* are variables that cause both the treatment and outcome thereby inducing a non-causal dependence between them. In our example, calendar time confounds the relationship between architectural choices and scaling efficiency via training on task-relevant data (brown path). *Selection bias* arises from how observations enter the sample. The decision to publish acts as a *collider*, which opens non-causal paths between recipe components and outcomes. In our example, the training recipe choices of certain providers appear more favorable in part because these providers can pick and publish the best score out of several (noisy and private) benchmark runs (dark red path).

Assumption 1 (Exchangeability)**.** The potential outcome $Y(a)$ under treatment value a is independent of the observed treatment A:

$$Y(a) \perp\!\!\!\perp A \quad \text{for all } a \in \mathcal{A}.$$

Positivity. Even if exchangeability holds, the causal effect of a particular treatment value cannot be estimated unless that value is observed in the sample. In the simple setting with one treatment a and one control $\emptyset$, positivity is typically satisfied. In this setting, the assumption can also easily be verified based on a sample.

Assumption 2 (Positivity)**.** Every treatment value of interest must occur with positive probability in the data:

$$P(A = a) > 0 \quad \text{for all } a \in \mathcal{A}.$$

[3] We use "selection bias" in the narrow Pearl/Hernán sense of conditioning on a collider (Hernán & Robins, 2016); some older causal-inference literature uses the term as an umbrella for any violation of exchangeability, including confounding.

SUTVA. The third assumption is the *stable unit treatment value* assumption, which has two parts: *no interference* between units, and *consistency* between observed and potential outcomes.

Assumption 3 (SUTVA)**.** Let a_i denote the treatment received by unit i and y_i the corresponding observed outcome. SUTVA comprises:

1. *No interference:* $Y_i(a_1, \ldots, a_n) = Y_i(a_i)$ for all $(a_1, \ldots, a_n) \in \mathcal{A}^n$, i.e. the potential outcome of unit i depends only on its own treatment.
2. *Consistency:* $A_i = a \implies Y_i = Y_i(a)$, i.e. the observed outcome equals the potential outcome under the treatment actually received.

In the observational approach, a violation of no interference means that the treatment received by one training run influences the potential outcomes of other runs. For instance, if the fact that one model was trained with Muon (Jordan et al., 2024) changes the potential outcomes (both with and without Muon) for another model, no interference is violated. In the single-run setting the units are not the training runs but the documents. Thus no interference has a slightly different interpretation in this context (see below).

C.2.2 Assumptions Made by Research Strategies in Section 5

The single-run designs in Section 5 relax the unconditional exchangeability of Assumption 1 and replace it with an alternative identifying assumption that is more credible in the single-run setting. Lesci et al. (2024) adopt a difference-in-differences design, which replaces exchangeability with a *parallel outcome trends* assumption: in the absence of treatment, the expected change in outcome over training time is the same in the treated and control groups. Lesci et al. (2025) use a regression discontinuity design, which assumes that potential outcomes vary continuously across the cutoff that determines treatment assignment, so units just above and below the cutoff are exchangeable in the limit. Elazar et al. (2022) instead invoke *conditional exchangeability*: potential outcomes are independent of treatment after conditioning on a sufficient set of observed covariates. We refer to the original papers for the formal definitions.

It is important to note that relaxing exchangeability typically tightens the other assumptions required for identification. For example, conditional exchangeability requires *conditional positivity*: every treatment value must occur with positive probability *within every stratum* of the conditioning covariates—a substantially stronger requirement than unconditional positivity (Assumption 2).

For Section 5, *no interference* (Assumption 3) is particularly relevant. In the single run setting, the "treatments" for the different "units", for example training on a particular document or not, all affect the same training (Sections 5 and F.1): training on data point i updates the parameters that are subsequently used to process data point j, so the potential outcome for j depends on i's treatment. As such, the assumption is strained.

C.3 External Validity

External validity concerns the generalization of a causal effect: even if the effect τ is correctly identified for the recipe under study, does it apply to the recipe (or population, or unit) we ultimately care about? We distinguish several flavors of external-validity claim that recur throughout the paper: external validity *across recipes*, in particular *across scale*, and external validity *across units*.

C.3.1 Assumptions Made by Research Strategies in Section 3.1

External validity across families and scale. The recipe-specific treatment effect (Definition 1) depends on the recipe ψ at which it is evaluated. When the treatment effect is evaluated for a proxy ψ' it is unclear whether the causal effect also applies for our target recipe ψ. We distinguish a couple of special cases. External validity *across scale* considers the case where the proxy ψ' is a scaled-down variant of the original ψ. External validity *across families* is the special case in which ψ and ψ' belong to different model families. In both cases, we want the treatment effect for the proxy and original recipe to be the same.

Assumption 4 (External validity of a proxy recipe). Let ψ be the target training recipe and ψ' a proxy training recipe; let a be the treatment of interest and y the outcome. The proxy recipe is

externally valid with respect to (a, y, ψ) iff

$$\tau(a, y; \psi) = \tau(a, y; \psi').$$

Scaling laws (Section 3.1) make a parametric version of this assumption. For proxy recipes $\psi_1, \ldots, \psi_n$ of increasing scale $s_1 < \cdots < s_n$, they assume that the treatment effects follow a parametric function of the scale, $\tau(a, y; \psi_i) = f_\theta(s_i)$, and that the same function describes the treatment effect at the target scale, $\tau(a, y; \psi) = f_\theta(s(\psi))$. Assumption 4 is the special case of a constant function.

C.3.2 Assumptions Made by Research Strategies in Section 4

External validity across recipes. The observational approach (Section 4) often pools data across many model families to obtain enough statistical power. Even if the resulting estimate is causally identified, it does not target the recipe-specific effect of Definition 1, but rather an *average* effect over the distribution of recipes from which the sample is drawn. External validity then concerns whether this average matches the recipe-specific effect for the target recipe.

Assumption 5 (External validity of an average effect across recipes). Let Π be a distribution over training recipes (e.g. the empirical distribution over publicly documented model families) and let

$$\bar{\tau}(a, y; \Pi) := \mathbb{E}_{\psi \sim \Pi}\big[\tau(a, y; \psi)\big]$$

be the average treatment effect across recipes drawn from Π. The average effect $\bar{\tau}(a, y; \Pi)$ is *externally valid* with respect to the target recipe ψ iff

$$\tau(a, y; \psi) = \bar{\tau}(a, y; \Pi).$$

The average and the recipe-specific effect coincide when the treatment effect is homogeneous across recipes in the support of Π, and may diverge when the effect interacts with other recipe components (cf. the heterogeneity discussion in Section 3.1).

C.3.3 Assumptions Made by Research Strategies in Section 5

External validity across units. We face a structurally similar problem in Section 5. But in the single run setting, the population are not the different training runs/recipes $\psi^{(i)}$, but the *components* of the recipe ψ_i, for example the training/validation set assignment of a document i in the dataset. In the context of the within-run approach, the question is whether averages across those within-run units (documents) reflect the effect for a particular within-run unit.

Assumption 6 (External validity of an average effect across within-run units). Let $\tau(a, y; \psi, i)$ denote the unit-specific treatment effect for unit i and $\bar{\tau}(a, y; \psi) := \frac{1}{n}\sum_{i=1}^{n} \tau(a, y; \psi, i)$ the average across units. The average effect $\bar{\tau}(a, y; \psi)$ is externally valid with respect to a target unit i iff

$$\tau(a, y; \psi, i) = \bar{\tau}(a, y; \psi).$$

As with external validity across recipes, the average and the unit-specific effect coincide when effects are homogeneous across units and may diverge under heterogeneity. In Section F.1 we also discuss external validity across models (Assumption 4).

C.4 Construct Validity

In general, the line between external and construct validity can be blurry (Shadish et al., 2002, Ch. 3); Both concern some form of generalization after all. But while external validity concerns generalization across populations, construct validity is about generalization across operationalizations of constructs.

We make a clear distinction: In our discussion *external validity* concerns whether a causal effect transfers across recipes or recipe components *for the same causal question*, while *construct validity* concerns whether the causal question itself is preserved when the treatment or outcome are replaced by a proxy. As follows, we formalize construct validity for proxy treatments and proxy outcomes.

C.4.1 Assumptions Made by Research Strategies in Section 3.2

Assumption 7 (Construct validity of a proxy treatment). Let a be the target treatment and a' a proxy treatment; let y be the outcome and ψ the recipe. The proxy treatment is *construct-valid* with

respect to (a, y, ψ) iff

$$\tau(a, y; \psi) = \tau(a', y; \psi).$$

Assumption 8 (Construct validity of a proxy outcome). Let y be the target outcome and y' a proxy outcome; let a be the treatment and ψ the recipe. The proxy outcome is *construct-valid* with respect to (a, y, ψ) iff

$$\tau(a, y; \psi) = \tau(a, y'; \psi).$$

D The Proxy Approach

This section complements Section 3 of the main paper.

D.1 Proxy Models

Scaling laws in the literature. Hestness et al. (2017) were among the first to recognize that the validation loss falls predictably in the size of the training dataset. Kaplan et al. (2020) and Hoffmann et al. (2022) observed that the validation loss after pretraining follows a power-law relationship in the sizes of model and training set, and that this relationship could be used to predict the optimal training recipe under a fixed compute budget. Subsequently, researchers have proposed scaling laws for many important components of the training recipe, including the batch size, learning rate, and training data composition (Bi et al., 2024; Bergsma et al., 2025; Shukor et al., 2025; Tao et al., 2024).

Success and failure of scaling laws. Many works have demonstrated that proxy models can be used to predict outcomes *across scale*. This is particularly true for predicting the pretraining validation loss as model and dataset size increase (Dey et al., 2023; Hu et al., 2024; Dubey et al., 2024). For example, Bi et al. (2024, Figure 5) accurately predict the performance of their 67B parameter model using proxy models trained with 1000x less compute. However, the literature also contains many examples where predictability breaks down, especially *across different training recipes* (Muennighoff et al., 2023; Kim et al., 2025; Mayilvahanan et al., 2025; Longpre et al., 2026). For example, scaling laws typically require that the composition of the pretraining data remains the same (Bi et al., 2024; Goyal et al., 2024; Ye et al., 2024). Similarly, Wang et al. (2025) show that the validity of proxy models for data selection breaks down once other training parameters change.

Emergent behaviors. An important conceptual challenge for proxy experiments are tasks that exhibit *emergent behaviors* (Wei et al., 2022; Schaeffer et al., 2023). By definition, an emergent behavior cannot be studied with proxies below the threshold at which the behavior starts to emerge. In this respect, it is noteworthy that most scaling laws target the pretraining validation loss, which is known to scale smoothly in model and dataset size. However, a number of works provide scaling laws for other outcomes, in particular benchmark scores, sometimes successfully (Krajewski et al., 2025), and sometimes with mixed results (Lourie et al., 2025).

D.2 Details on the Experiment Depicted in Figure 3

The experiment demonstrates a failure of external validity across scale that cannot be detected from the proxy experiments alone. The target recipe ψ trains a 304M-parameter language model on 30B tokens of unfiltered web data. The treatment a replaces the unfiltered data with quality-filtered data. Because the pool of filtered data is much smaller than the token budget, the treatment implies that the filtered data is repeated, and the number of repetitions grows with the token budget. The outcome y is the validation loss. The proxy recipes ψ' are under-trained versions of the target recipe with up to 3B tokens, and we extrapolate to the target recipe with a scaling law in the number of training tokens.

Model and training. We train 304M parameter models with the OLMo 2 architecture (OLMo et al., 2024). We use a warmup-stable-decay schedule: a single run with constant learning rate proceeds to 27B tokens, and for every horizon D we branch off this run at $0.9D$ tokens and linearly decay the learning rate to zero over the remaining $0.1D$ tokens. Every reported loss is measured at the end of learning rate decay.

Experiment conditions. The unfiltered, fresh arm trains on 30B unique tokens sampled from FineWeb (Penedo et al., 2024). The filtered, repeated arm trains on a pool of 0.75B tokens sampled from FineWeb-Edu (Lozhkov et al., 2024), a subset of FineWeb selected by a classifier for educational

quality. The pool is reshuffled after every pass, so that a horizon of $D = k \cdot 0.75$B tokens corresponds to k passes over the pool. Both arms are evaluated on the same 200M held-out tokens from FineWeb-Edu, which are disjoint from both training pools.

Scaling law estimation. For each arm, we fit $L(D) = E + BD^{-\beta}$ to the proxy models, following the procedure of Hoffmann et al. (2022): We minimize a Huber loss ($\delta = 10^{-3}$) on the log-residuals with a grid of initializations followed by L-BFGS. Similarly to what is done in https://marin.community/, we fix the irreducible loss at $E = 2.81$ for better extrapolation.

D.3 Proxy Treatments

Proxy treatments keep the target recipe ψ but replace the treatment a with a proxy treatment a' that is cheaper to apply. The controlled experiment is then performed with a', that is, we estimate $\tau(a', y; \psi) = \mathbb{E}[y(a') \mid \psi] - \mathbb{E}[y(\emptyset) \mid \psi]$ and use it as an estimate of $\tau(a, y; \psi)$.

Proxy treatments are common for research questions that involve adding or removing parts of the training data, where the ideal experiment would require to re-train the model from scratch. To reduce cost, researchers approximate re-training with a proxy, for example by continuing to train an intermediate model checkpoint (OLMo et al., 2024), or fine-tuning the fully trained model (see (Maini et al., 2024; Pawelczyk et al., 2025) among many examples). In this case, a is *"modify the pretraining data"* and a' is *"modify the continual pretraining data"*. In the second case, a' is *"fine-tune the fully trained model on the modified data"*. In all cases, the treatment is applied at the target scale, and $\tau(a', y; \psi)$ can be estimated well.

When does a proxy treatment work? In its simplest form, the proxy approach assumes that the proxy treatment has the same effect as the original treatment, that is, $\tau(a', y; \psi) = \tau(a, y; \psi)$ (Assumption 7). In practice, proxy treatments are often used to choose among candidate treatments $a_1, \ldots, a_k$, for example, to choose among data sources to add to training (OLMo et al., 2024; Dubey et al., 2024). In this case, we might be satisfied with less than equality: It could be enough for the proxy treatments $a'_1, \ldots, a'_k$ to preserve the ranking of the effects, that is, $\tau(a'_i, y; \psi) > \tau(a'_j, y; \psi)$ whenever $\tau(a_i, y; \psi) > \tau(a_j, y; \psi)$, and for a single treatment, only the sign of the effect. While more plausible, this weaker assumption does not allow to reliably estimate the magnitude of the treatment effect (which means that it could become negligible at scale). At any rate, whether a proxy treatment is valid depends on the mechanism by which it acts on the model, which can differ from that of the original treatment (this is illustrated is Figure 4).

Validity Profile. Because proxy treatments operationalize a different construct than the original treatment, we risk measuring the wrong causal effect, $\tau(a', y; \psi) \neq \tau(a, y; \psi)$. This means that proxy treatments threaten *construct* validity (Assumption 7). In contrast, proxy treatments do not threaten *statistical*, *internal*, or *external* validity. This is because the proxy treatment is applied to the target recipe at the target scale, and the experiment is controlled.

> **Proxy Treatments – Validity Threats**
>
> **Construct validity of the proxy treatment:** Does the proxy treatment operationalize the same construct as the target treatment, so that $\tau(a', y; \psi) = \tau(a, y; \psi)$? If not, the measured causal effect may be different from the target effect.

In contrast to the relatively well-recognized external validity challenges of proxy models (Section 3.1), the threat of the construct validity stemming from treatment proxies has received comparatively little attention in the literature. In particular, many papers use fine-tuning proxies without devoting much discussion to their validity.

Experiment In Figure 4 in the main paper, we provide an experiment to demonstrate the lack of construct validity of fine-tuning proxies for privacy leakage. In this experiment, a is single-epoch pretraining on the data in question, a' is multi-epoch fine-tuning on the same data, and y is the success of a privacy attack. From Figure 4, we see that $\tau(a', y; \psi) > \tau(a, y; \psi)$: The proxy treatment overestimates the effect of privacy leakage.

Proxy treatments in the literature. Multi-epoch fine-tuning proxies are widespread in the literature on privacy and unlearning, where the ideal experiment would pretrain the model with and without the data in question (Pawelczyk et al., 2025; Shi et al., 2024; Maini et al., 2024; Dorna et al., 2025).

This is despite the fact that this proxy treatment is known to have different memorization dynamics: Empirical and theoretical results suggest that training for more epochs increases privacy leakage (Song & Mittal, 2021; Dong et al., 2022; Leemann et al., 2023; Tan et al., 2023; Zhang et al., 2024b), whereas examples seen during one-epoch pretraining are often forgotten quickly (Jagielski et al., 2023; Bordt & Pawelczyk, 2026). Therefore, we argue that fine-tuning proxy treatments tend to overestimate the model's susceptibility to privacy attacks (Shokri et al., 2017; Carlini et al., 2021), a point that we illustrate in Figure 4.

Continued training of intermediate checkpoints on a modified data mixture has been used to ablate the composition of mid-training data (OLMo et al., 2024), and to study how models acquire knowledge from documents that are inserted into the training of a checkpoint (Chang et al., 2024).

D.4 Details on the Experiment Depicted in Figure 4

The target recipe ψ pretrains a 546M-parameter language model from scratch on 216B tokens, during which sensitive canary sequences are shown to the model. The proxy treatment a' attempts to simulate this pretraining leakage by fine-tuning an already-pretrained, clean model on a much smaller dataset of 1B tokens containing the same set of canaries. Because fine-tuning updates an already-converged model on a distinct data distribution, it measures a fundamentally distorted construct (memorization due to fine-tuning) rather than the target construct (memorization due to pretraining). The outcome y is privacy leakage, evaluated via a membership inference attack (Shokri et al., 2017; Panda et al., 2025).

Model and training. We train 546M-parameter models using the OLMo 2 architecture. For the target pretraining recipe, the model is trained from scratch on 216B tokens subject to the canary insertions, following the pretraining recipes of Bordt & Pawelczyk (2026). For the proxy fine-tuning recipe, we first pretrain an OLMo 2 546M model from scratch on 216B tokens without any canary insertions. We then fine-tune this clean model for up to 10 epochs on 1B fresh tokens per epoch which include the canary sequences.

Experiment conditions. To measure privacy leakage, we follow the experimental setup of Panda et al. (2025) and place canary sequences into the training data. The canary sequences follow the format `[training text, secret keys]`. For the secret keys, we consider two regimes: plain natural text continuations, and sequences of 8 randomly selected tokens. To evaluate the impact of repetition scale, each specific canary sequence is inserted into the training data either 1, 4, or 16 times.

Leakage evaluation. For both the target and proxy recipes, we evaluate how well a membership inference attack can identify whether the canary sequences were part of the training dataset. We report the privacy leakage in terms of the Area Under the Receiver Operating Characteristic Curve (AUC). An AUC of 0.5 corresponds to random guessing (no measurable privacy leakage), while an AUC of 1.0 indicates perfect memorization by the model.

D.5 Proxy Outcomes

Proxy outcomes keep the target recipe ψ and the treatment a but replace the outcome y with a proxy outcome y' that is cheaper to measure. This is because the outcome that we would like to optimize or measure is either unavailable or expensive to compute. The controlled experiment then estimates $\tau(a, y'; \psi) = \mathbb{E}[y'(a) \mid \psi] - \mathbb{E}[y'(\emptyset) \mid \psi]$, which we use as an estimate of $\tau(a, y; \psi)$. An important caveat in this section is that we do not discuss the construct validity of the outcome itself. This means that we do not ask whether the outcome measures what matters to the phenomenon that it aims to represent (Bean et al., 2025; Freiesleben & Zezulka, 2025; Wallach et al., 2025)—instead, we take the outcome as given, and only ask whether it is accurately represented by the proxy.

When does a proxy outcome work? In its simplest form, the proxy approach assumes that the treatment has the same effect on the proxy outcome as on the original outcome, that is, $\tau(a, y'; \psi) = \tau(a, y; \psi)$ (Assumption 8). Often, the two outcomes are measured on different scales, for example a loss and an accuracy, so that equality of the effects is not meaningful. What is required then is that the proxy preserves the direction of the effect, that is, whether the treatment improves or degrades the outcome, or, for candidate treatments $a_1, \ldots, a_k$, their ranking.

Validity Profile. Because proxy outcomes operationalize a different construct than the original outcome, we risk measuring the wrong causal effect, $\tau(a, y'; \psi) \neq \tau(a, y; \psi)$. Thus, similarly to proxy treatments (Section 3.2), proxy outcomes threaten *construct* validity (Assumption 8). Because the proxy outcome is measured for the target recipe at the target scale, and the experiment remains controlled, proxy outcomes do not threaten *statistical*, *internal*, or *external* validity.

> **Proxy Outcomes – Validity Threats**
>
> **Construct validity of the proxy outcome:** Do changes in the value of the proxy outcome accurately reflect changes in the value of the original outcome, that is $\tau(a, y'; \psi) = \tau(a, y; \psi)$?

We would like to highlight that the validity of the proxy outcome can depend on the treatment (similar to other approaches): For some treatments, causal effects on the original outcome may be accurately reflected in the proxy, whereas for other treatments this may not be the case. In general, this is a complex question of construct validity related to the causal mechanisms by which the treatment influences both the proxy and the original outcome. For example, the treatment may influence the proxy but not the original outcome, or vice-versa.

The pretraining validation loss as a proxy for model performance after post-training. Foundation models are developed in different stages, including pretraining, mid-training, and post-training (Ouyang et al., 2022; Bi et al., 2024; Lambert et al., 2024). The goal of the development process is to obtain a model that performs well on complex and broadly defined downstream tasks after post-training, for example, to obtain a model that has good agentic coding capabilities. However, model architecture and optimization hyperparameters are rarely determined by measuring downstream outcomes. Instead, they are determined based on the cross-entropy loss after pretraining (Yang et al., 2021; Dey et al., 2025; Haas et al., 2025). The implicit assumption behind this approach is that models that have better validation loss after pretraining are also going to perform better on downstream tasks after post-training: In other words, the pretraining validation loss serves as a proxy outcome for downstream capabilities. This approach seems to work reasonably well. However, Han et al. (2026) show that optimizing the weight decay parameter based on pretraining validation loss does not always lead to the best downstream performance after SFT.

Benchmark prediction as a proxy for full benchmark performance. Another example of proxy outcomes in the current literature are benchmark prediction methods. These methods approximate otherwise costly benchmark scores by evaluating the model on a small subset of the benchmark data (Vivek et al., 2024; Li et al., 2024a; Polo et al., 2024; Pacchiardi et al., 2024). While these approximations may work under some conditions, recent work has shown that they fail for the most performant models (Zhang et al., 2025b). Hence, optimizing against the proxy risks overlooking the best of all models, indicating a lack of construct validity of the proxy.

E The Observational Approach

This section complements Section 4 of the main paper.

E.1 The Observational Approach in the Literature

The discussion follows the three steps of Equation 1 that the observational approach threatens: after the available meta-data, it covers statistical estimation (statistical validity), confounding and exchangeability (internal validity), and effect heterogeneity across model families (external validity).

The Fragmented Meta-Data Landscape. Perhaps the most important consideration in the observational approach is the public availability of data about foundation models. For many model families meta-data are publicly available, in particular for open-weight models (Biderman et al., 2023; Groeneveld et al., 2024). However, because there are few large-scale models, the sample size often remains small; The data is scattered across leaderboards and technical reports (Epoch AI, 2026; Fourrier et al., 2024; Chiang et al., 2024; Liu et al., 2025); And, importantly, much of the metadata remains undisclosed. For proprietary models even the parameter counts are routinely kept secret (Li, 2026)— but also for open-weight models details such as the pretraining data mix are commonly missing (Thrush et al., 2025; Liu et al., 2025; Mertens et al., 2026). To work around these limitations, the observational approach is often combined with proxies (Thrush et al., 2025; Liu et al.,

2025; Dominguez-Olmedo et al., 2025), such as per-domain-loss instead of the true mixing weights, inheriting the construct validity problems discussed in Section 3.2.

Small Samples and Statistical Estimation. Since predicting outcomes from the full recipe is infeasible—e.g., the data mixing weights alone can comprise thousands of components, yet the number of fully documented training runs is small (Thrush et al., 2025; Liu et al., 2025)—researchers typically drop "non-treatment features" and impose simple parametric functional forms. For example, Ruan et al. (2024) assume sigmoidal relationships between scale parameters and downstream outcomes, and Thrush et al. (2025) reduce their analysis to pairwise rank correlations. When the treatment is low-dimensional and the parametric assumptions hold, the statistical validity of the observational approach *can* compare favorably, since it can draw on meta-data from multiple families.

Confounding and Exchangeability. But even when the statistical model is accurate, it only describes the *associations* between features and outcomes. Eventually, we would like to arrive at causal conclusions, but in observational data, association and causation do not necessarily coincide. For example, so-called *confounders* may cause both the treatment and the outcome variable, thereby inducing an association between them, even if the treatment has no causal effect. A prominent confounder is *calendar time* (Dominguez-Olmedo et al., 2025; Zhang et al., 2025a): Observational data shows that newer model families require much less compute to achieve a given benchmark performance (Dominguez-Olmedo et al., 2025). This could be attributed to architectural or algorithmic innovations. However, not only the architecture has evolved over time, but also the training data. A range of evidence suggests that the data mix for newer models is increasingly tuned to reflect benchmark tasks (Dekoninck et al., 2024; Dominguez-Olmedo et al., 2025; Zhang et al., 2024a; 2025a). As a result, architectural developments are confounded with downstream outcomes via calendar time and training data composition (Figure 8).

More formally, confounding is a violation of *exchangeability* (Assumption 1), which requires that the potential outcomes are independent of the treatment assignment. Exchangeability is *untestable* based on observational data alone; whether it holds can only be assessed using experiments or prior knowledge about the data-generating process. As examples in the literature illustrate (Dominguez-Olmedo et al., 2025; Singh et al., 2025; Whitfill, 2025), there is substantial reason to believe that this assumption is violated in practice—but nevertheless the issue was largely ignored. In principle, *causal inference* methodology allows to relax the exchangeability assumption: Using prior knowledge about the data-generating process we can systematically remove non-causal associations (Rubin, 1974; Pearl, 2009). But so far, this methodology has found surprisingly little application.

We suspect that in part data limitations are to blame: To adjust for a confounder, the confounder must be observed. To work around these data limitations, a recent approach combines an observational sample with a proxy treatment: To show that "training on the test task" is an important driver of improvements in scaling efficiency, Dominguez-Olmedo et al. (2025) fine-tune the models in their observational sample on task relevant data. While powerful, this strategy requires careful interpretation. The research question shifts from "how does pretraining affect capability" to "how adaptable is the model to the test task", a threat to *construct validity* (Section 3.2).

Effect Heterogeneity across Families. As discussed in Section 3.1, treatment effects can vary substantially across model families: Scaling law coefficients change with the pretraining data composition (Bi et al., 2024, Section 3.3), and data selection conclusions can flip when other hyperparameters are varied (Wang et al., 2025). When the observational analysis pools across families, it averages over this heterogeneity—the same cross-family replication that improves statistical precision can degrade external validity (Assumption 5). This is a fundamental tension for the observational approach: Estimating recipe-specific effects requires parameters that describe the interaction of the treatment with other family-specific parameters, but fitting those parameters requires more data per family.

A second way in which exchangeability fails is *selection bias* (Singh et al., 2025; Whitfill, 2025), which we discuss in Appendix C.2.1.

E.2 Details on the Experiment Depicted in Figure 5

Setup. The experiment studies the effect of confounding on estimates of data quality. The treatment are the data mixing weights $w \in \Delta^{99}$ over 100 web domains (the topic $\times$ format cells of Wettig et al.

(2025)); the outcome is bits-per-byte (BPB) on HellaSwag; the confounder is the model size. As starting point, we use 512 runs with 30M and 1,024 runs with 60M parameters. In every run, the mixing weights were drawn randomly such that they form a controlled experiment (from a Dirichlet distribution centered on the natural corpus distribution, independently of the model size).

Emulating an observational sample. To emulate confounding, we subsample 30M and 60M runs such that for smaller runs high-quality data mixes are more likely. Such a confounding mechanism is plausible in practice: high-quality data is scarce at large token budgets (Muennighoff et al., 2023). More specifically, we assign every data mix in our 30M and 60M pools a quality score u_i: the standardized projection of its centered log-ratio transformed mixing weights onto the log-amplification that the FineWeb-Edu quality filter (Lozhkov et al., 2024) induces on each domain (Wettig et al., 2025). Then 30M runs are retained with probability $\sigma(\lambda u_i)$ and 60M runs with probability $\sigma(-\lambda u_i)$. We calibrate λ such that the correlation between model size and $-u_i$ among the retained runs (about 760 of 1,536) equals a target $\rho \in \{0, 0.15, 0.3, 0.45, 0.6, 0.7\}$, with 20 seeds each.

Estimation and evaluation. All analyses regress the outcome on the 100 mixing weights with the Lasso (regularization $\alpha_{\max}/30$). The *naive* analysis directly fits a model und the subsampled runs, the *size-adjusted* analysis controls for model size, and the *stratified* analysis fits one regression per model size and averages the coefficients. We report the predicted change in BPB for a one-standard-deviation shift of the data mix along the quality direction, and how many of the reference's 20 most beneficial domains the fitted regression recovers (4 expected at random). Table 3 reports means over the 20 seeds.

Table 3: **Results of the experiment depicted in Figure 5.** Mean over 20 seeds. The reference effect of the data quality, estimated from all runs, is -0.0010 BPB. The expected number of recovered domains under random guessing is 4.

	Effect of data quality (BPB)		**Top-20 domains recovered**	
Confounding strength ρ	naive	size-adjusted	naive	size-adjusted
0.00	-0.0020	-0.0010	6.95	16.10
0.15	$+0.0056$	-0.0009	6.45	16.30
0.30	$+0.0121$	-0.0010	5.30	16.10
0.45	$+0.0203$	-0.0009	3.75	15.90
0.60	$+0.0282$	-0.0010	2.95	16.10
0.70	$+0.0334$	-0.0009	2.60	16.00

Results. Without confounding, the naive estimate (-0.0020 ± 0.0027) agrees with the reference (-0.0010). With confounding, estimates become biased: the estimated effect of data quality changes sign, and for $\rho \geq 0.45$ the naive analysis recovers fewer beneficial domains than random guessing. The size-adjusted and the stratified analysis are unaffected by ρ. Even without confounding, the naive analysis recovers only 7 of the 20 domains, against 16 for the size-adjusted analysis.

F The Single-run Approach

F.1 Theoretical Control Outcomes

To gain more fine-grained insight about the causal effects for individual units, *theoretical controls* can be employed. We discuss two settings: One without experimental control over the model run, and one where we can inject data with known properties into training to obtain mathematical guarantees.

Observing vs. controlling one training run. In the first setting, theoretical controls are derived without any control over the training run. One instance of this are Influence Functions, which are used to approximate the contribution of individual training examples to a particular prediction (Koh & Liang, 2017; Grosse et al., 2023; Bae et al., 2024; Mlodozeniec et al., 2025). To do so, they replace the effect of "removing one data point" with "reducing the weight of one data point", and employ strong statistical assumptions for its approximation. While Influence Functions accurately capture the LOO retraining effect for simple models like linear or logistic regression (Hampel, 1974; Giordano et al., 2019; Koh & Liang, 2017), the assumptions underlying the approximation fail for frontier

models with highly non-convex loss landscapes and complex stochastic training dynamics (Schioppa et al., 2023; Bae et al., 2022; Li et al., 2024b).

In the second setting, we control the training run, allowing us to *inject* signals that can be used to derive a mathematically guaranteed control (Andrew et al., 2024; Pawelczyk et al., 2025; Bordt & Pawelczyk, 2026; Sablayrolles et al., 2020; Christ et al., 2024; Bouaziz et al., 2025). For example, in privacy auditing, Steinke et al. (2023) randomly assign "canary" data to the training or validation set. The analytical control is derived from the binomial distribution governing these independent assignments. Applying a similar principle, Pawelczyk et al. (2025) inject independent noise canaries (e.g., Gaussian canaries) into gradient updates, training data, or input embeddings (Pawelczyk et al., 2025; Bordt & Pawelczyk, 2026). Because the expected dot product between a model's output and a Gaussian canary from the control group is analytically zero, the control baseline is mathematically guaranteed. However, canary insertion may degrade model performance depending on signal strength—ranging from up to 5% accuracy drop (Steinke et al., 2023) to negligible impact when using Gaussian canaries (Pawelczyk et al., 2025).

Validity threats. Interference (Assumption 3) remains a central threat for *internal* validity: All units share the same model parameters, so training on canary i may affect the outcome for canary j (Keinan et al., 2025). Depending on the particular approach, *statistical*, *external*, and *construct* validity may all be at risk: As the case with influence functions, analytical approximations may rely on proxy treatments (*construct* validity) and on strong statistical assumptions (*statistical* validity). Injecting more "canary" data improves *statistical* validity but modifies the training recipe and may reduce *external* validity *across models*. However, analytical controls can be used to estimate effects for individual units, thereby maintaining *external* validity *across units*.

F.2 Details on the Experiment Depicted in Figure 7

Here we specify the experimental details supporting the validity claims of Section 5 where we study the single-run memorisation estimator of Lesci et al. (2024). We keep their design unchanged (one training run; a panel of sampled instances evaluated across checkpoints; the difference and difference-in-differences estimators given in Equations 5 and 6 below) and add two aspects: (i) a ground-truth counterfactual obtained from *paired* training runs that hold all sources of training variance fixed, and (ii) validation sets constructed to violate exchangeability in controlled ways.

Notation and Estimators. Units are packed training sequences x of L tokens. The treatment $G(x) = g$ is the macro-batch in which x is trained; $G(x) = \infty$ denotes a validation instance that is never trained on. The outcome at checkpoint c is the sequence log-likelihood $Y_c(x) = \log p_{\theta_c}(x)$, summed over the $L - 1$ predicted tokens. The estimand is the expected counterfactual memorisation of cohort g at checkpoint c,

$$\tau_{g,c} = \mathbb{E}\big[Y_c(x; g) - Y_c(x; \infty) \mid G(x) = g\big], \tag{4}$$

which is an average treatment effect on the treated for the specific training run. The two estimators we consider are

$$\hat{\tau}^{\text{diff}}_{g,c} = \overline{Y}_c(g) - \overline{Y}_c(\infty), \qquad \text{(Difference Estimator)} \tag{5}$$

$$\hat{\tau}^{\text{did}}_{g,c} = \big(\overline{Y}_c(g) - \overline{Y}_{g-1}(g)\big) - \big(\overline{Y}_c(\infty) - \overline{Y}_{g-1}(\infty)\big) \qquad \text{(Diff-in-Diff Estimator)}, \tag{6}$$

where $\overline{Y}_c(\cdot)$ denotes the sample mean over the panel instances of the respective group and $g - 1$ denotes the last checkpoint before cohort g is trained (step $g-50$ in our setup). Equation 5 is unbiased under i.i.d. sampling and random assignment (exchangeability); Equation 6 is unbiased under parallel trends and no anticipation.

Model and data. For this experiment, we train a 179M parameter OLMo 2 model (OLMo et al., 2024), initialized from scratch and trained for a single epoch on 3.15B tokens of web text of the OLMo-2 pretraining data.

Checkpoints and macro-batches. Checkpoints are saved at $c \in \{0, 50, 100, \ldots, 1500\}$. All instances trained between checkpoints $c - 50$ and c form macro-batch $\mathcal{G}_c$ ($50 \times 512 = 25{,}600$ sequences); we index cohorts by $g \in \{50, \ldots, 1500\}$. The pre-treatment checkpoint for cohort g is $g - 50$ (denoted $g - 1$ in Equation 6).

Determinism. To make sure the paired runs only differ in 100 sequences, all runs use identical seeds, data order and hyper-parameters, and are run on the same H100 GPU. Each paired run is *branched*

from Run T's step-$(g^\star - 50)$ checkpoint with restored data-loader and RNG state, and the swapped sequences are the only rows of macro-batch $\mathcal{G}_{g^\star}$ that differ between the two runs. All 100 swapped sequences sit in a single batch, so the two runs are identical up to that batch.

Reserved held-out data. Before training, the data stream is split by source and stream order into (i) the training stream (768,000 DCLM sequences), (ii) a reserve pool $\mathcal{R}$ of 20,000 DCLM sequences from the same stream (i.i.d. validation set, replacement instances, hard validation set, matched probes), (iii) a held-out DCLM candidate pool of 1.4M sequences for near-duplicate selection, and (iv) a pool of 10,000 StarCoder sequences from the same stream for the domain-shifted validation set. The models never trained on information from (ii)–(iv).

Ground Truth from Paired Runs. We construct the counterfactual for the *specific* run by holding all training variance fixed.

For three cohorts $g^\star \in \{150, 750, 1350\}$ (end of warm-up, mid-training, late training), we train Run $\mathrm{C}_{g^\star}$: identical to Run T in seed, data order and hyper-parameters, except that the $n_{g^\star} = 100$ panel instances of macro-batch $\mathcal{G}_{g^\star}$ are replaced *in place* (same step, same position within the batch) by 100 fresh sequences from $\mathcal{R}$ that have no near-duplicate in the training data stream. The cohorts serve to compare early, mid, and late training and are not pooled: Figure 7 uses $g^\star = 750$ only. For this cohort we trained a second paired run C'_{750}, and panel (a) averages the estimates and the ground truth over the two paired runs.

Batch composition, token count and schedule are thus unchanged. For $c \geq g^\star$, Run $\mathrm{C}_{g^\star}$ observes $Y_c(x; \infty)$ for the swapped instances, and the ground truth is

$$\tau^{\text{truth}}_{g^\star,c} = \frac{1}{n_{g^\star}} \sum_{x \in \mathcal{S}_{g^\star}} \left[Y_c^{\mathrm{T}}(x) - Y_c^{\mathrm{C}_{g^\star}}(x) \right], \tag{7}$$

where $\mathcal{S}_{g^\star}$ is the swapped set (the set S of Section 5 for $g^\star = 750$). The same paired run gives $\mu_c^\infty(g^\star)$ directly (the mean of $Y_c^{\mathrm{C}_{g^\star}}$ over $\mathcal{S}_{g^\star}$).

Validation Sets. Each validation set has 2,000 sequences, is never trained on, and is evaluated at all checkpoints of Run T. Only the control group changes; the treated data and Run T are shared.

(V1) **i.i.d. split** ($\mathcal{V}_{\text{iid}}$) is sampled from the reserved held-out pool $\mathcal{R}$. Here, exchangeability holds by construction and we expect both estimators to be unbiased. This serves as a baseline.

(V2) **Deduplication mismatch** ($\mathcal{V}_{\text{dedup}}$) contains held-out sequences from the same stream that have a MinHash near-duplicate (Jaccard ≥ 0.7 over token 5-grams, 128 permutations) in the training stream. Because the DCLM stream is already fuzzily deduplicated, the natural near-duplicate rate is low (0.18% of held-out candidates at Jaccard ≥ 0.7); we therefore selected a 1.4M-sequence held-out candidate pool, obtained 2,525 candidates and sampled 2,000 (mean maximal Jaccard to the training stream 0.82). This is the scenario Lesci et al. (2024) flag (deduplicated training set, non-deduplicated validation). For each sequence we record the step at which its training near-duplicate is first seen. We expect a level gap and, because the near-duplicates are trained on throughout the run, a trend gap that grows with c.

(V3) **Domain shift** ($\mathcal{V}_{\text{domain}}$) is made of StarCoder sequences from the same OLMo-2 stream (the training stream is DCLM web text only), so that the domain shift is clean. Here we expect level and trend gaps as well.

(V4) **Difficulty shift** ($\mathcal{V}_{\text{hard}}$) sampled from $\mathcal{R}$ but restricted to the top quartile of sequence loss under an independent 179M reference model (a stage-1 checkpoint trained on 210B tokens of a disjoint region of the same stream; 4,351 eligible sequences, 2,000 sampled). Here we expect at least a level gap by construction.

G Case Studies

Here we provide case studies where we apply the checklist to representative papers (Khaddaj et al., 2025; Muennighoff et al., 2023; Dong et al., 2024; Chang et al., 2024; Ruan et al., 2024; Dominguez-Olmedo et al., 2025; Lesci et al., 2024). The papers that we discuss often contain a large number of experiments and different settings. To simplify the discussion, we complete the checklist only for the main experimental result of the paper. To present the results of the case studies in a compact way, we summarize the answers to the checklist in 1-page figures.

Below we discuss the case study results for one representative checklist for each research strategy.

G.1 Case Study for the Proxy Approach (Section 3)

Khaddaj et al. (2025) studies proxy models. More specifically, the paper studies the validity of proxy models itself. The authors train a 760M parameter reference model, then compares the cross-entropy loss of this model after pretraining to smaller proxy models (40M–760M parameters). Khaddaj et al. (2025) find that test losses are highly correlated even at $175\times$ less compute, with exceptions for the smallest proxies and for some test sets. Because the paper primarily measures the cross-entropy loss after pretraining, but the claims in the paper are more generally about "model behavior", this paper also employs a proxy outcome.

Figure 9 provides a summary of the most important checklist answers for Khaddaj et al. (2025). While Khaddaj et al. (2025) discuss the validity of the smaller proxy models in great details, a critical dimension of external validity is not being discussed: Whether the 760M parameter reference model is itself large enough to stand in as a reliable proxy for the "large-scale models" that the results in the paper are supposedly about.

G.2 Case Study for the Observational Approach (Section 4)

Ruan et al. (2024) fit scaling laws to publicly released models instead of training their own: a low-dimensional capability score, extracted from the standard benchmark scores of 77 models, predicts downstream outcomes such as emergent-task performance, agent benchmarks, and the gains from chain-of-thought. All three are found to change smoothly and predictably with the capability score, verified on held-out and later models. The paper also employs a proxy treatment, since the capability score stands in for scale (training compute).

Figure 13 summarizes the checklist. It identifies internal validity as a central threat to the validity of the conclusions that receives insufficient discussion in the paper. Moreover, it identifies the construct validity threat stemming from the proxy treatment.

G.3 Case Study for the Single-Run Approach (Section 5)

Lesci et al. (2024) employ a single-run design to estimate the effect of training on a sequence at step t on its memorisation at a later step t'. To address internal validity, they employ a difference-in-differences design. They use the estimator to compare memorisation across models of different sizes, training steps, and learning-rate schedules, finding that larger models memorise more and more persistently.

The checklist (Figure 15) identifies a key threat to internal validity that is left undiscussed: interference between units. That is, training on one document may have a causal effect on the outcome of another document, thereby biasing the analysis. The remaining internal validity threats receive proper discussion, as do statistical and construct validity; external validity is partly addressed, since the claims are limited to the Pythia suite and to cohort averages, but the predictability of larger models from smaller ones is not argued beyond the studied range.

Checklist Summary from the Perspective of a Reviewer: Khaddaj et al. Small-to-Large Generalization: Data Influences Models Consistently Across Scale. ICLR, 2025.

Research question Do changes in pretraining data influence smaller and larger models identically? The paper trains proxies (40M–760M parameters) and a 760M reference model on ten data distributions and correlates their test losses across distributions.

Conclusions The choice of training data distribution generally affects models across scale similarly, even when the difference in compute is large. Specifically, test losses are highly correlated even at $175\times$ less compute, with exceptions for the smallest proxies and for some test sets. Small proxy models can be a powerful tool for practically and effectively answering how changes in training setup affect the behavior of large scale models.

Part A — Approximation Strategies

A2	Proxy models	[Yes]	The 760M reference model stands in for the large-scale models the conclusions are about. The other proxies (40M–760M) are themselves the object of study.
A3	Proxy treatments	[No]	Separate models are trained from scratch on each of the ten data distributions.
A4	Proxy outcomes	[Yes]	The paper mostly measures the cross-validation loss after pretraining whereas the title and abstract speak of predictions and behavior.
A5	Observational data	[No]	All training runs are the authors' own.
A6	Single-run units	[No]	Units are separate training runs.
A7	Theoretical controls	[No]	Not applicable.
A8	Other approximations	[No]	No approximation beyond A2 and A4.

Part B — Validity Threats

SV	Statistical validity	[Partly]	The measured effects are reasonably large and well-separated, however, there is one run per configuration, ten points per correlation, and no error bars.
IV	Internal validity	[Yes]	The authors assign the treatment to separate, otherwise identical training runs.
EV	External validity	[No]	The paper discusses the problem of the validity of proxy models, and the validity of the smaller proxy models is studied carefully. However, whether the setup itself generalizes beyond the 760M reference model (which is itself a proxy model for the large-scale models mentioned in the conclusion) is *not* discussed in the paper.
CV	Construct validity	[Partly]	Loss versus accuracy is examined (Fig. 3), but whether loss consistency carries over to the accuracy-type outcomes the framing names is not discussed.

Figure 9: **Case study for Khaddaj et al. (2025).** This figure depicts an overview of the most important answers to the checklist (Appendix H) for Khaddaj et al. (2025), a paper which studies the validity of small-scale proxies for training data selection, among others.

Checklist Summary from the Perspective of a Reviewer: Muennighoff et al. Scaling Data-Constrained Language Models. NeurIPS, 2023.

Research question The paper studies scaling laws for repeated data, and how the optimal compute budget should be split between model parameters and epochs when unique training data is limited. The paper trains over 400 models of 7M–8.7B parameters for up to 1500 epochs, fits scaling laws, and verifies on held-out loss and downstream tasks.

Conclusions Up to four epochs of repeated data cost almost nothing in loss. After that, returns diminish sharply beyond about 16 epochs.

Part A — Approximation Strategies

A2	Proxy models	[Yes]	Models of 7M–8.7B parameters stand in for the large language models the scaling law is meant for, including the 120B Galactica model.
A3	Proxy treatments	[No]	Not applicable.
A4	Proxy outcomes	[Yes]	Held-out loss is the outcome of the scaling law. Claims concern model performance, which is measured on 19 downstream tasks.
A5	Observational data	[No]	Not applicable.
A6	Single-run units	[No]	Not applicable.
A7	Theoretical controls	[No]	Not applicable.
A8	Other approximations	[No]	Not applicable.

Part B — Validity Threats

SV	Statistical validity	[Yes]	Five seeds with standard errors.
IV	Internal validity	[Yes]	The authors train every model from scratch on nested data subsets with fixed hyperparameters.
EV	External validity	[Yes]	Experimentally validated up to 8.7B parameters, then extrapolated to 120B.
CV	Construct validity	[Yes]	Conclusions based on the loss are confirmed on downstream tasks.

Figure 10: **Case study with Muennighoff et al. (2023).** Overview of the answers to the checklist (Appendix H) for Muennighoff et al. (2023), a paper which fits a scaling law for repeated data to models of up to 8.7B parameters and applies it to larger ones.

Checklist Summary from the Perspective of a Reviewer: Dong et al. Generalization or Memorization: Data Contamination and Trustworthy Evaluation for Large Language Models. ACL Findings, 2024.

Research question Can data contamination of an LLM, that is, benchmark test data in the pretraining training data, be detected from sampled outputs? The paper simulates contamination by LoRA-fine-tuning four 7B models on HumanEval and GSM8K. It evaluates a detector (CDD) and corrected metric (TED) on the resulting models.

Conclusions Contamination makes the output distribution peaked; CDD detects it better than n-gram, perplexity, and embedding baselines. TED removes most of the performance gain from contamination.

Part A — Approximation Strategies

A2	Proxy models	[Yes]	Four open 7B models fine-tuned with LoRA stand in for the proprietary LLMs that the claims are about.
A3	Proxy treatments	[Yes]	Contamination is implemented as LoRA fine-tuning on benchmark data as a proxy for modifying the pretraining data.
A4	Proxy outcomes	[No]	Not applicable.
A5	Observational data	[No]	Not applicable.
A6	Single-run units	[No]	Not applicable.
A7	Theoretical controls	[No]	Not applicable.
A8	Other approximations	[No]	Not applicable.

Part B — Validity Threats

SV	Statistical validity	[Partly]	The measured effects are reasonably large and well-separated, however, there is a lack of error bars and statistical tests.
IV	Internal validity	[Yes]	The treatment is assigned in controlled experiments. That the base models used as controls may themselves be contaminated is discussed.
EV	External validity	[Partly]	Heterogeneity across settings and domains is examined. Whether a detector calibrated on 7B fine-tuned models transfers to large chat models is not discussed.
CV	Construct validity	[No]	While the paper discusses LoRA versus full fine-tuning, there is no discussion of fine-tuning versus pretraining contamination, and fine-tuning itself as a possible cause of peakedness is not discussed.

Figure 11: **Case study with Dong et al. (2024).** This figure depicts an overview of the most important answers to the checklist (Appendix H) for Dong et al. (2024), a paper which proposes methods to detect data contamination from an LLM's sampled outputs.

Checklist Summary from the Perspective of a Reviewer: Chang et al. How Do Large Language Models Acquire Factual Knowledge During Pretraining? NeurIPS, 2024.

Research question How do LLMs acquire and forget factual knowledge during pretraining, and how does this depend on the data and the training conditions? The paper performs continual pretraining of OLMo-1B and 7B checkpoints, injecting 120 fictional passages in different conditions. It tracks the log-probability of probes about each passage at every step.

Conclusions Knowledge is acquired by accumulating small increases of probability at each encounter.

Part A — Approximation Strategies

A2	Proxy models	[Yes]	OLMo models of 1B and 7B parameters stand in for LLMs in general.
A3	Proxy treatments	[Yes]	Fictional passages written by GPT-4 stand in for facts that occur naturally in the pretraining corpus.
A4	Proxy outcomes	[Yes]	The log-probability of probe target spans stands in for acquired knowledge.
A5	Observational data	[No]	Not applicable.
A6	Single-run units	[Yes]	The 120 injected passages are units within one continued-pretraining run.
A7	Theoretical controls	[No]	Not applicable.
A8	Other approximations	[No]	Not applicable.

Part B — Validity Threats

SV	Statistical validity	[Yes]	Forgetting fits come with R^2 and slope uncertainties.
IV	Internal validity	[Partly]	Interference between the 120 passages is not discussed but unlikely to affect the conclusion.
EV	External validity	[Partly]	Transfer beyond OLMo-7B and the extrapolation of results beyond 2500 observed training steps are not discussed.
CV	Construct validity	[Partly]	The gap between probe probability and generated output is mentioned. Whether injected synthetic passages behave like naturally occurring facts is touched on.

Figure 12: **Case study with Chang et al. (2024).** Overview of the answers to the checklist (Appendix H) for Chang et al. (2024), a paper which tracks how OLMo checkpoints acquire and forget fictional passages injected during continued pretraining.

Checklist Summary from the Perspective of a Reviewer: Ruan et al. Observational Scaling Laws and the Predictability of Language Model Performance. NeurIPS, 2024.

Research question Can we build a single scaling law that generalizes across families and allows us to predict downstream performance for larger models based on smaller models? Can we predict the effects of post-training interventions like CoT?

Conclusions We can predict downstream performance for larger models from smaller models for several relevant outcomes. The relevant outcomes are: so-called "emergent" capabilities, agentic capabilities, and performance after post-training interventions.

Part A — Approximation Strategies

A2	Proxy models	[Yes]	Laws are fitted on weaker models (below Llama-2-7B compute) standing in for the stronger held-out and future models the claims are about.
A3	Proxy treatments	[Yes]	Predictions are not directly based on scale in the sense of size, but based on *capability*, which is extracted from model benchmark scores.
A4	Proxy outcomes	[No]	Outcomes are measured directly: agent benchmark scores, the scores on the emergent tasks, and the gains from post-training interventions.
A5	Observational data	[Yes]	All 77 models were trained by third parties; scores and metadata come from leaderboards and papers.
A6	Single-run units	[No]	Not applicable.
A7	Theoretical controls	[No]	Not applicable.
A8	Other approximations	[No]	Not applicable.

Part B — Validity Threats

SV	Statistical validity	[Yes]	The functional form assumption is tested by holdouts; the danger of overfitting in light of the small sample sizes is discussed; robustness checks and preregistered experiments are performed.
IV	Internal validity	[No]	Despite the predictive framework, the paper suggests a causal interpretation, that is, the laws allow us to predict the effect of scale (and post-training interventions) on outcomes. However, the paper does not discuss any causal assumptions; for example, potential threats imposed by confounders or selection bias are not discussed in the paper and could potentially bias the result.
EV	External validity	[Yes]	Transfer across scale and families is the central topic of the paper and discussed in various places in the paper. The claims are substantiated, for example, with preregistered experiments and holdout predictions.
CV	Construct validity	[Partly]	The paper shows for models with known scale that capability is log-linearly correlated with scale; but all predictions are still based on capability instead of scale.

Figure 13: **Case study with Ruan et al. (2024).** Overview of the answers to the checklist (Appendix H) for Ruan et al. (2024), a paper which fits scaling laws to publicly released models instead of training its own.

Checklist Summary from the Perspective of a Reviewer: Dominguez-Olmedo et al. Training on the Test Task Confounds Evaluation and Emergence. ICLR, 2025.

Research question Do newer language models outperform older ones at equal pretraining compute because they train more on the test task, and does fine-tuning every model on the same task data before evaluation adjust for it? The paper compares 56 released base models before and after such fine-tuning.

Conclusions Newer models outperform older ones at equal compute. But after fine-tuning all models on the same task-relevant data the difference vanishes, so the improvement is attributed to training on the test task.

Part A — Approximation Strategies

A2	Proxy models	[No]	Not applicable.
A3	Proxy treatments	[Yes]	Fine-tuning after pretraining stands in for training on the test task during pretraining, and model recency (trained after November 2023) for the unobserved amount of such training.
A4	Proxy outcomes	[No]	Not applicable.
A5	Observational data	[Yes]	All models were trained and released by third parties.
A6	Single-run units	[No]	Not applicable.
A7	Theoretical controls	[No]	Not applicable.
A8	Other approximations	[No]	Not applicable.

Part B — Validity Threats

SV	Statistical validity	[Yes]	Family-clustered standard errors, large R^2 for every fit, and robustness to the cutoff, to an alternative split, and to the fine-tuning learning rate.
IV	Internal validity	[Yes]	The within-model effects of fine-tuning are assigned by design, and the cross-time attribution states its causal assumptions and estimates a null direct effect of recency (App. B). Minor: Selection into the sample of released, documented models is not discussed.
EV	External validity	[Yes]	The claims concern released models in general, estimated on 56 open base models with disclosed compute. Transfer to instruction and chat models is tested (36 models, App. F) and heterogeneity across families is examined (Sec. 4.1). Minor: transfer beyond the sampled range, to the larger and proprietary models that motivate the paper, is neither argued nor limited.
CV	Construct validity	[Partly]	The adjusted score is framed as performance after task adaptation (Sec. 4.1, Sec. 6), and the gap between post-hoc fine-tuning and training on the test task during pretraining is stated as the key assumption (App. B) and tested in outcome space only (App. C.3); several conclusions nevertheless speak of capabilities and compute efficiency (Sec. 1, Sec. 4.1).

Figure 14: **Case study with Dominguez-Olmedo et al. (2025).** Overview of the answers to the checklist (Appendix H) for Dominguez-Olmedo et al. (2025), a paper which compares publicly released models across time and fine-tunes each of them on the same task data to adjust for training on the test task.

Checklist Summary from the Perspective of a Reviewer: Lesci et al. Causal Estimation of Memorisation Profiles. ACL, 2024.

Research question How much does training on a sequence at step g change the log-likelihood a model assigns to it at a later checkpoint, and how does this memorisation profile vary with training step, checkpoint, and model size? The paper estimates this within single Pythia runs (70M–12B) by comparing sequences trained at step g against never-trained validation sequences with a difference-in-differences estimator.

Conclusions Memorisation is stronger and more persistent in larger models, is shaped by data order and the learning-rate schedule, and has a profile that is stable across model sizes, so that memorisation in larger models is predictable from smaller ones.

Part A — Approximation Strategies

A2	Proxy models	[No]	Not applicable.
A3	Proxy treatments	[No]	Not applicable.
A4	Proxy outcomes	[Yes]	Memorisation is operationalised as the effect on sequence log-likelihood; other metrics give different estimates (App. D).
A5	Observational data	[Yes]	Pythia was trained by EleutherAI; the authors did not decide which sequence is trained when.
A6	Single-run units	[Yes]	Units are sequences within one run: the training sequences of a macro-batch form the treated cohort, validation sequences the controls.
A7	Theoretical controls	[No]	Not applicable.
A8	Other approximations	[No]	Not applicable.

Part B — Validity Threats

SV	Statistical validity	[Yes]	Sample sizes are large; simultaneous bootstrap confidence intervals. Minor: cross-size comparisons rest on one run per size, corresponding uncertainties not discussed.
IV	Internal validity	[No]	Exchangeability, positivity, consistency, and the difference-in-differences assumptions are stated and argued, but the no-interference assumption, e.g., regarding near-duplicates or spillover onto the validation controls, is never discussed.
EV	External validity	[Partly]	Claims are limited to the Pythia suite and to cohort averages; the predictability of larger models from smaller ones is neither restricted to nor argued beyond the 70M–12B range.
CV	Construct validity	[Yes]	The gap between log-likelihood and memorisation is named and tested with alternative metrics (App. D).

Figure 15: **Case study with Lesci et al. (2024).** Overview of the answers to the checklist (Appendix H) for Lesci et al. (2024), a paper which estimates memorisation from the checkpoints of single Pythia training runs.

H CHECKLIST

Here we present the full checklist for authors and reviewers.

How to use this checklist.

- **Part A** states the research question and identifies the approximation strategies of the experiment. A [Yes] in Part A makes specific Part B threats applicable, so [NA] is not allowed for those questions; the last guideline of each Part A question says which.
- **Part B** asks, for each of the four types of validity, whether the paper *properly discusses* the threats to it: names them and argues them away, tests them, bounds them, or limits its claims accordingly. It does not ask whether the assumptions hold. A [No] is not disqualifying; it flags threats the paper should discuss.
- **Answers** are [Yes] / [No], and [NA] where the guidelines allow it. [NA] means the threat cannot arise in this design. If it is applicable but undiscussed, answer [No].
- **Justification:** one or two sentences pointing to the supporting section, quote, or figure.

H.1 PART A — RESEARCH QUESTION AND APPROXIMATION STRATEGIES

A1. What is the research question? Research question: [one sentence naming treatment, outcome, and target models]
Conclusions: [free text]
Guidelines:

- The **treatment** is the change to the training process or data whose effect the paper wants to know (e.g., adding math data to the pre-training mix, lowering the learning rate, including a document in the training set). The **control** is the same setup without that change.
- The **outcome** is the property measured to see the effect (e.g., a benchmark score, the training loss, how well a document is memorized).
- The **target models** are the models the paper's *conclusions* are about. They may be larger, or trained differently, than the models the paper actually trains.
- A **unit** is the thing that does or does not receive the treatment and on which the outcome is measured. Usually a unit is one training run; in some designs it is a document, token, or subword within one run. Units with the treatment form the **treatment group**, units without it the **control group**.
- Example: How does the MMLU benchmark score (outcome) of Olmo 3 32B (target model) change when we remove all math competition data from its pretraining data (treatment)?
- List the main conclusions the paper draws from this experiment; Part B refers back to them.

A2. Do the experiments involve proxy models? Answer: [Yes] / [No]
Justification:
Guidelines:

- A proxy model is a model that is smaller or cheaper to train than the target models the claims are about, and that stands in for them in the experiment.
- Example: 1B models standing in for 32B models; a series of small models fitted with a scaling law to predict behavior at a larger size.
- Note: If the answer is [Yes], the threat to external validity is applicable and thus [NA] is not allowed for question EV.

A3. Do the experiments involve proxy treatments? Answer: [Yes] / [No]
Justification:
Guidelines:

- A proxy treatment replaces the treatment the paper cares about with a cheaper stand-in, or only observes a proxy of it instead of implementing it.
- Example: fine-tuning on private data instead of pre-training on it; per-domain loss as a stand-in for the true data-mixture weights.

- Note: If the answer is [Yes], the threat to construct validity (of the treatment) is applicable and thus [NA] is not allowed for question CV.

A4. Do the experiments involve proxy outcomes? Answer: [Yes] / [No]
Justification:
Guidelines:

- A proxy outcome replaces the outcome the paper cares about with a cheaper stand-in.
- Example: "training loss" as stand-in for "benchmark score"; a small subset of benchmark questions as stand-in for the full benchmark.
- Note: If the answer is [Yes], the threat to construct validity (of the outcome) is applicable and thus [NA] is not allowed for question CV.

A5. Do the experiments involve observational data? Answer: [Yes] / [No]
Justification:
Guidelines:

- Data is *experimental* if the researchers decided which units receive the treatment (they ran the training runs, ablations, or sweeps themselves). It is *observational* if the runs already existed and someone else made those decisions, so the researchers did not control and do not know how the treatment was assigned.
- Example: relating recipe choices to benchmark scores across publicly released models; analyzing released checkpoints of an existing model suite; using leaderboards or model cards.
- Note: If the answer is [Yes], the threat to internal validity is applicable and thus [NA] is not allowed for question IV.

A6. Do the experiments treat elements of a single training run as independent units? Answer: [Yes] / [No]
Justification:
Guidelines:

- This strategy emulates two or more training runs from a single run: elements of the run (documents, tokens, subwords, . . .) are treated as independently treatable units, each with its own treatment status and outcome.
- Example: documents in the training set form the treatment group, held-out documents the control group, and the model's loss on each document is its outcome.
- Note: If the answer is [Yes], the threat to internal validity is applicable and thus [NA] is not allowed for question IV.

A7. Do the experiments use theoretical controls? Answer: [Yes] / [No]
Justification:
Guidelines:

- A theoretical control is an outcome that was not measured but computed from theory: the paper analytically approximates what the outcome *would have been* under the treatment or under the control, and compares against that.
- Example: influence functions analytically approximating "what if this data point were not in the training data".
- Note: If the answer is [Yes], the threats to internal validity and to construct validity (of the treatment) are applicable and thus [NA] is not allowed for questions IV and CV.

A8. Do the experiments involve other approximations? Answer: [Yes + description] / [No]
Justification:
Guidelines:

- Any way in which the design deviates from the ideal experiment that is not covered by A2 to A7.
- Example: measuring the outcome only at coarse checkpoint intervals; analyzing a subsample of the units.

- Note: If the answer is [Yes], name the validity types that the approximation threatens; [NA] is not allowed for those questions.

H.2 Part B — Validity Threats

Each question asks whether the paper properly discusses the threats to one type of validity. The guidelines explain the type and list the assumptions that have to be established for it.

SV. Statistical validity. **Question:** Does the paper discuss and establish the statistical validity of its main estimate?
Answer: [Yes] / [No]
Justification:
Guidelines:

- Always applicable; [NA] is not an option.
- Statistical validity asks whether the number the paper estimates from a finite, noisy sample reliably reflects the true underlying quantity, that is, whether it would come out about the same if the experiment were repeated. Establishing it means addressing, where applicable:
  - *Uncertainty quantification:* outcomes vary with the random seed, data ordering, sampled evaluation subset, and so on. The paper should report error bars and say which randomness they capture and which they cannot (one run per configuration cannot show seed-to-seed variation). Example: the difference between two recipes can be smaller than the seed-to-seed variation of one recipe.
  - *Functional-form assumptions:* when a formula is fitted (a power law, a linear regression, a sigmoid), the estimate is only as good as the formula, especially when extrapolating beyond the measured range. The paper should state and justify or test the assumed form. Example: scaling-law fits assume a power law over the fitted range.
  - *Sample size and dependence:* few units, or units that are not independent, make the estimate less precise than the raw count suggests. Example: fine-tunes of one base model are not independent units.
  - *Multiple comparisons:* the best of many noisy comparisons (arms, benchmarks, checkpoints) looks strong by chance.
- Answer [No] if any applicable aspect is undiscussed, and name it.

IV. Internal validity. **Question:** Does the paper discuss and establish that its estimate reflects a genuine causal effect of the treatment?
Answer: [Yes] / [No] / [NA]
Justification:
Guidelines:

- [NA] is not allowed if A5, A6, or A7 = Yes. [NA] means the experimenter assigned the treatment to separate, independent training runs and the groups differ in nothing but the treatment; the assumptions below then hold by design.
- Internal validity asks whether a reliable estimate reflects a genuine *causal* effect of the treatment, rather than some other difference between the compared groups or an influence of the units on each other. An association between treatment and outcome is causal only under *identifying assumptions*, which the paper has to establish:
  - *Exchangeability:* treatment and control groups do not differ systematically in anything other than the treatment. It fails through *confounders* (third factors that influence both which units get the treatment and their outcome) and *selection bias* (units enter the sample in a way that depends on treatment and outcome, e.g., only the best-performing models are published). Adjusting for measured confounders relaxes this to *conditional* exchangeability: the groups are comparable once those confounders are held fixed. Example: architecture and training data composition are confounded by calendar time; a correlation between architecture and benchmark score may actually be caused by data composition.
  - *Positivity:* every treatment value the paper draws conclusions about occurs in the sample, among otherwise comparable units; when other variables are adjusted for, within every combination of those variables. Effects of treatment values that no unit received can only be extrapolated, not

estimated. Example: if a document is not included in any training run in the sample, we cannot estimate its effect.
  - *SUTVA* (stable unit treatment value assumption), which has two parts. *No interference:* each unit's outcome depends only on its own treatment, not on the treatments other units received. It is strained whenever units share something through which a treatment can spread (training data or evaluation pools across runs, model parameters within a run). Example: many released models are fine-tuned on synthetic data generated by one model, so a change to that model's recipe also changes the outcomes of the models trained on its outputs. *Consistency:* the treatment is well defined, that is, all treated units received the same version of it, so the observed outcome is the outcome of the treatment the paper names. Example: "document is in the training set" hides deduplication, number of epochs, and ordering.
  - *Other identification strategies:* if the paper does not compare treatment and control groups directly but uses difference-in-differences (comparing the change over time in treated units to that in untreated units), instrumental variables, or a regression discontinuity design, the assumptions of that strategy have to be discussed. Example: difference-in-differences assumes parallel trends, that is, that treated and untreated units would have developed alike without the treatment.
- A [Yes] requires the paper to name each applicable assumption and argue for it, test it, bound its violation, or limit its claims. Answer [No] if any applicable assumption is undiscussed, and name it.

EV. External validity. Question: If the paper's claims go beyond the setting it studied, does it discuss whether the effect transfers to the setting the claims are about?
Answer: [Yes] / [No] / [NA]
Justification:
Guidelines:

- [NA] is not allowed if A2 = Yes. [NA] means the experiments run at the scale, with the training recipes, model families, and units, and at the aggregation level that the claims are about.
- External validity asks whether a correctly measured causal effect carries over from the setting in which it was measured to the setting the claims are about. Three transfers are typical:
  - *Across scale:* effects measured on smaller models need not hold for larger ones; they can change in size, or even in sign, with model size or amount of training. Example: the optimal learning rate decreases with model size; the best value at 1B can be wrong at 70B.
  - *Across training recipes, model families, and units:* a training recipe is everything that determines a training run (architecture, data, optimizer, hyperparameters, . . .); models sharing central parts of their recipe form a family. An effect can depend on everything that was held fixed in the experiment. Example: data-selection conclusions can flip when other hyperparameters change.
  - *Across aggregation levels:* an average across many units can hide that individual units or subgroups react very differently (*effect heterogeneity*). Example: an average memorization effect near zero is compatible with strong memorization of particular documents.
- A [Yes] requires verification in the target setting, an argument from established behavior (e.g., known scaling behavior of the quantity), or claims explicitly limited to what was studied. Answer [No] if any applicable transfer is undiscussed, and name it.

CV. Construct validity. Question: If the implemented treatment or the measured outcome differs from the treatment or outcome named in the conclusions, does the paper discuss whether findings about the one carry over to the other?
Answer: [Yes] / [No] / [NA]
Justification:
Guidelines:

- [NA] is not allowed if A3, A4, or A7 = Yes. [NA] means treatment and outcome are implemented exactly as stated in the research question.
- Construct validity asks whether the treatment as implemented and the outcome as measured really capture the concepts named in the research question. The concrete implementation of a concept (fine-tuning on private data for "pre-training on private data"; one benchmark score for "reasoning ability") is its *operationalization*. Two gaps can open:

  - *Treatment:* two treatments that look alike can act on the outcome in different ways, so the measured effect may be a different effect than the one claimed. Example: multi-epoch fine-tuning memorizes far more than single-epoch pre-training, so it overestimates privacy leakage.
  - *Outcome:* the treatment may affect the measured outcome but not the outcome of interest, or vice versa. Example: improvement in "validation loss" does not necessarily reflect improvement in "reasoning capabilities".
- A [Yes] requires an argument that the effects coincide (treatment) or that the two outcomes move together (outcome), or claims aligned with what was implemented and measured. Answer [No] if either applicable gap is undiscussed, and name it.