THE *Magic Correlations*: UNDERSTANDING KNOWLEDGE TRANSFER FROM PRETRAINING TO SUPERVISED FINE-TUNING

Simin Fan, Dimitris Paparas, Natasha Noy
Google Research

Binbin Xiong, Noveen Sachdeva, Berivan Isik
Google Deepmind

ABSTRACT

Understanding how language model capabilities transfer from pretraining to supervised fine-tuning (SFT) is fundamental to efficient model development and data curation. In this work, we investigate four core questions: **RQ1**. To what extent do accuracy and confidence rankings established during pretraining persist after SFT? **RQ2**. Which benchmarks serve as robust cross-stage predictors and which are unreliable? **RQ3**. How do transfer dynamics shift with model scale? **RQ4**. How well does model confidence align with accuracy, as a measure of calibration quality? Does this alignment pattern transfer across training stages? We address these questions through a suite of *correlation protocols* applied to accuracy and confidence metrics across diverse data mixtures and model scales. Our experiments reveal that transfer reliability varies dramatically across capability categories, benchmarks, and scales—with accuracy and confidence exhibiting distinct, sometimes opposing, scaling dynamics. These findings shed light on the complex interplay between pretraining decisions and downstream outcomes, providing actionable guidance for benchmark selection, data curation, and efficient model development.

1 INTRODUCTION

Modern Large Language Model (LLM) training proceeds in stages: pretraining on massive text corpora (Penedo et al., 2023; Weber et al., 2024; Penedo et al., 2024; Li et al., 2025), followed by supervised fine-tuning (SFT) on curated instruction data (Ouyang et al., 2022; Ivison et al., 2023; Chung et al., 2022) and reinforcement learning (Guo et al., 2025; Wang et al., 2025; Cheng et al., 2025). Critical decisions regarding data curation, mixture composition, and resource allocation are often made based on pretraining benchmarks alone, upon small-scale proxy models (Kaplan et al., 2020; Hoffmann et al., 2022; Xie et al., 2023; Fan et al., 2024). This practice relies on the following fundamental assumption: pretraining performance reliably predicts post-SFT performance, and that internal model representations remain stable across training stages. In this paper, we scrutinize this assumption through comprehensive empirical analysis of capability transfer from pretraining to SFT stage. We focus on four core questions:

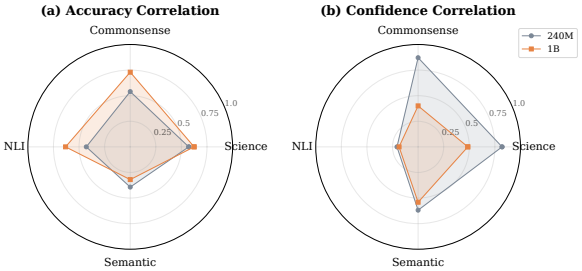


Figure 1: **Cross-stage correlation by capability category.** *(a) Accuracy correlation*: the 1B model generally shows higher transferability; *(b) Confidence correlation*: 240M maintains substantially higher correlation especially in Commonsense (0.87 vs. 0.40) and Science (0.82 vs. 0.49) domains. This transferring pattern indicates that larger models undergo more confidence reorganization during SFT despite better accuracy preservation.

1. **Performance Stability**: To what extent do accuracy and confidence rankings established during pretraining persist after SFT?
2. **Benchmark Reliability**: Which benchmarks serve as robust early-stage predictors, and which are unreliable?
3. **Scale Dynamics**: How do transfer patterns shift with model scale?
4. **Calibration Quality**: How well does model confidence align with accuracy? Does this alignment pattern transfer across training stages?

To address these questions, we propose a suite of *correlation protocols* and conduct systematic experiments on transformer models at two scales (240M and 1B parameters) across 9 diverse pretraining data mixtures. We evaluate on 20 benchmarks spanning four *capability categories*, each corresponding to one particular skill and knowledge domain: ***Commonsense Reasoning*** (Commonsense), ***Scientific Reasoning*** (Science), ***Natural Language Inference*** (NLI), and ***Semantic Understanding*** (Semantic). Our primary findings include:

- **Inverse Scaling of Accuracy and Confidence Transfer**: Larger models exhibit stronger cross-stage *accuracy* correlation but weaker *confidence* correlation, suggesting these metrics capture fundamentally different aspects of capability transfer.
- **Task-Dependent Transfer Reliability**: Transfer dynamics vary dramatically by capability domain. *Commonsense* and *Science* categories show consistently high cross-stage correlation, while *NLI* and *Semantic Understanding* exhibit weaker patterns. We identify specific benchmarks with particularly weak transfer, marking them as unreliable early-stage predictors.
- **Intra-Category Coherence Shifts with Scale**: At smaller scales, benchmarks within the same capacity category often *compete*—data mixtures that improve one benchmark degrade its semantic neighbors. At larger scales, this competition gives way to *synergy*, with positive intra-category correlations emerging.
- **Category-Dependent Calibration Quality**: Performance-confidence alignment—the correlation between model confidence and accuracy—varies dramatically by capability domain, with *Science* showing strong alignment while *Commonsense* and *Semantic* categories exhibit systematic miscalibration that persists through SFT.
- **Data Curation Trade-off between Accuracy and Calibration Quality**: Strict educational-filtered pretraining data (Fineweb-edu) preserves or improve on scientific reasoning accuracy, while exhibits consistent performace-confidence alignment, revealing that strict educational filtering can preserves accuracy while degrades internal calibration patterns.

These findings provide actionable guidance for practitioners: *benchmark selection* should account for category-specific transfer reliability and scale effects; *pretraining data decisions* should consider cross-stage transferability and the divergent impacts on different capability domains.

2 Related Work

Our work intersects several active research areas: benchmark design and valuations, transfer learning dynamics in LLMs, model calibration, and data curation.

Scaling Laws and Transfer Learning. The relationship between pretraining and downstream performance has been extensively studied through scaling laws (Hernandez et al., 2021; Chen et al., 2025). Kaplan et al. (2020) established power-law relationships governing how model performance improves with parameter count, dataset size, and compute budget. Hoffmann et al. (2022) refined it with the Chinchilla scaling laws, demonstrating that prior models were significantly undertrained relative to their parameter counts. More recently, Hernandez et al. (2021) extended scaling analysis to transfer learning settings, showing that SFT efficiency also follows predictable scaling relationships, with (Isik et al., 2025; Lourie et al., 2025; Schaeffer et al., 2025) challenging these findings.

Benchmark Design and Contamination. The validity of evaluation benchmarks has received increasing scrutiny. Sainz et al. (2023) and Deng et al. (2024) documented widespread benchmark contamination in modern LLMs, where test data inadvertently appears in training corpora. Rodriguez et al. (2021) analyzed benchmark difficulty and item response characteristics, while Bowman & Dahl (2021) critiqued construct validity in NLU benchmarks. The HELM framework (Liang et al., 2023) introduced multi-dimensional evaluation spanning accuracy, calibration, robustness, and fairness. Our correlation-based analysis provides a complementary lens: rather than measuring absolute performance, we assess which benchmarks maintain *stable rankings* across training stages—a property essential for reliable early-stage evaluation.

Model Calibration. Well-calibrated confidence estimates are crucial for deploying LLMs in high-stakes applications. Guo et al. (2017) demonstrated that modern neural networks are often poorly calibrated, and proposed temperature scaling as a post-hoc remedy. For language models specifically, Kadavath et al. (2022) showed that larger models exhibit improved calibration on factual questions, while Desai & Durrett (2020) found that pre-trained transformers require careful calibration for downstream tasks. Tian et al. (2023) and Geng et al. (2024) provided comprehensive

surveys of confidence estimation methods for LLMs. Our cross-stage confidence correlation analysis extends this literature by examining whether calibration patterns *persist* through SFT—a question with direct implications for practitioners who must decide when to re-calibrate models.

Data Mixture and Curation Effects. Understanding how pretraining data composition affects downstream capabilities is central to efficient LLM development. Penedo et al. (2023) demonstrated that carefully filtered web data can match curated corpora, while Lozhkov et al. (2024a) showed that educational content filtering improves certain reasoning benchmarks. Li et al. (2025) introduced systematic data curation pipelines with measurable downstream effects. Data attribution methods (Grosse et al., 2023; Park et al., 2023) enable fine-grained analysis of which training examples contribute to specific capabilities. Our work takes a complementary aggregate approach: rather than attributing individual samples, we characterize how broad data mixture choices (web source, code proportion) affect the *transferability* of capabilities across training stages.

3 Experimental Protocols

3.1 Training Pipeline

We investigate the capability transfer dynamics between two critical phases of LLM development: *Pretraining* (PT), where the model acquires foundational knowledge from massive textual corpora, and *Supervised Fine-tuning* (SFT), where the model is adapted via instruction-following data for a more careful calibration.

Pretraining. We train a suite of decoder-only transformer models at two scales (240M and 1B parameters) on 9 distinct data mixtures (Table 3). To systematically disentangle the effects of data source versus data distribution, we construct these mixtures by crossing three *web data sources* and two *code data sources*, with three *mixing proportions*.

Specifically, we utilize: (1) *General Web Data* sourced from RefinedWeb (Penedo et al., 2023), FineWeb-Edu (Lozhkov et al., 2024a), or DCLM (Li et al., 2025); (2) *Code Data* from StarCoder (Li et al., 2023) or The Stack v2 (Lozhkov et al., 2024b); and (3) *Curated Knowledge* from RedPajama-v2 (Weber et al., 2024) (including Wikipedia, ArXiv, Github and StackExchange). The mixing proportions vary the ratio of web data (25%, 45%, 65%) relative to code and curated sources, allowing us to comprehensively study the impact of data mixture composition on downstream knowledge transfer. Complete mixture specifications are provided in Appendix C.3.

Supervised Fine-tuning (SFT). Following the PT stage, we fine-tune the checkpoints pretrained from various pretraining data mixture on `Tulu-v2-mix` (Ivison et al., 2023). Models are pretrained for 12B (240M) and 52B (1B) tokens, followed by 5 epochs of SFT with cosine learning rate scheduling. Hyperparameter configurations are detailed in Appendix C.

3.2 Evaluation Benchmarks

We evaluate the performance of language models across 20 benchmarks organized into four *capability categories*, corresponding to various knowledge domains:

1. ***Commonsense Reasoning*** (Commonsense): CommonsenseQA (Talmor et al., 2019), WinoGrande (Sakaguchi et al., 2019), HellaSwag (Zellers et al., 2019), PIQA (Bisk et al., 2019), SIQA (Sap et al., 2019), COPA (Wang et al., 2020), BoolQ (Wang et al., 2020)—*tasks requiring physical world knowledge and reasoning skill on social relationship*;
2. ***Scienctific Reasoning*** (Science): ARC-Challenge and ARC-Easy (Clark et al., 2018), SciQ (Welbl et al., 2017), OpenBookQA (Mihaylov et al., 2018)—*tasks requiring factual and scientific knowledge*;
3. ***Natural Language Inference*** (NLI): MNLI (Williams et al., 2018), QNLI (Wang et al., 2019), RTE (Wang et al., 2020), CB (Wang et al., 2020)—*tasks requiring inference about entailment relationships*;
4. ***Semantic Understanding*** (Semantic): QQP (Sharma et al., 2019), MRPC (Dolan & Brockett, 2005), WiC (Pilehvar & Camacho-Collados, 2019), WSC (Levesque et al., 2012), MultiRC (Khashabi et al., 2018)—*tasks requiring understanding of semantic equivalence, coreference, and textual relationships*.

We provide full benchmark details in Appendix C.4. For each benchmark, we report two primary metrics: ***accuracy***, defined as the proportion of correct predictions; and ***confidence***, defined as the probability assigned to the selected answer averaged across the evaluation set.

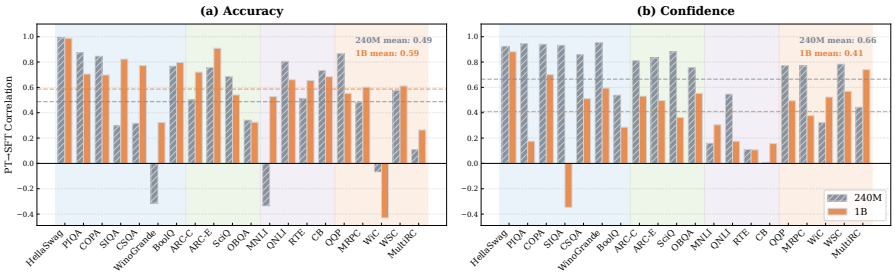


Figure 2: **Cross-stage Correlation across various benchmarks.** Each bar shows the Pearson correlation between PT and SFT performance on the certain benchmark across data mixtures. (a) *Accuracy Correlation*: the 1B model achieves higher transferrability than 240M (in average $\bar{r}$=0.59 v.s. 0.49). (b) *Confidence Correlation*: the pattern *reverses*—240M achieves substantially stronger transfer than 1B model ($\bar{r}$=0.41 v.s. 0.66). Background colors indicate capability categories (Commonsense, Science, NLI, Semantic).

3.3 The Lens of Correlations

To quantify the preservation and reorganization of capabilities across training stages, we analyze model behavior through five complementary correlation protocols, along three dimensions.

3.3.1 Cross-Stage Correlation

We propose the following correlation protocols to measure *how the accuracy and confidence calibration patterns established during pretraining persist through SFT*.

Cross-Stage Accuracy Correlation ($r_{\text{acc}}^{\text{stage}}$). For a given benchmark, we compute the Pearson correlation between pretraining and SFT accuracy across data mixtures:

$$r_{\text{acc}}^{\text{stage}} = \text{corr}\left(\mathbf{a}^{\text{PT}}, \mathbf{a}^{\text{SFT}}\right) \tag{1}$$

where $\mathbf{a}^{\text{PT}}, \mathbf{a}^{\text{SFT}} \in \mathbb{R}^M$ are pretraining and SFT accuracy scores collected across $M = 9$ mixtures. High $r_{\text{acc}}^{\text{stage}}$ indicates the benchmark is a *reliable early-stage proxy* from pretraining stage for downstream capability.

Cross-Stage Confidence Correlation ($r_{\text{conf}}^{\text{stage}}$). Analogously, for the confidence scores across various pretraining data mixture, we compute the cross-stage confidence correlation:

$$r_{\text{conf}}^{\text{stage}} = \text{corr}\left(\mathbf{c}^{\text{PT}}, \mathbf{c}^{\text{SFT}}\right) \tag{2}$$

where $\mathbf{c}^{\text{PT}}$ and $\mathbf{c}^{\text{SFT}}$ represent the confidence scores across various data mixtures. A high $r_{\text{conf}}^{\text{stage}}$ suggests that the model's calibration "fingerprint"—its level of uncertainty about specific inputs—derived from pretraining that persists despite the perturbations of SFT.

3.3.2 Intra-Category Correlation

Benchmarks are typically organized into semantic categories (e.g., commonsense reasoning, natural language inference) based on the assumption that tasks within a category tap similar underlying capabilities. But do benchmarks within the same category actually behave coherently when data mixtures change? We investigate this question through three complementary protocols that capture different aspects of intra-category dynamics.

Pretraining Coherence ($r_{\text{intra}}^{\text{PT}}$). For benchmarks i, j in category C:

$$r^{\text{PT}}(i,j) = \text{corr}\left(\mathbf{a}_i^{\text{PT}}, \mathbf{a}_j^{\text{PT}}\right), \quad i < j \tag{3}$$

This metric measures how data mixtures that improve benchmark i also improve benchmark j *during pretraining*. High positive values indicate *pretraining-stage synergy*; negative values indicate *competition* where optimizing for one benchmark degrades another.

SFT Coherence ($r_{\text{intra}}^{\text{SFT}}$). Analogously for post-SFT:

$$r_{\text{intra}}^{\text{SFT}}(i,j) = \text{corr}\left(\mathbf{a}_i^{\text{SFT}}, \mathbf{a}_j^{\text{SFT}}\right), \quad i < j \tag{4}$$

Similarly, the within-SFT coherence assesses data mixtures that improve benchmark i also improve benchmark j *after SFT*. Comparing $r_{\text{intra}}^{\text{SFT}}$ with $r_{\text{intra}}^{\text{PT}}$ reveals whether SFT *introduces* or *resolves* intra-category competition.

Cross-Stage Intra-Category Coherence ($r_{\text{intra}}^{\text{cross}}$). For transfer between *different* benchmarks:

$$r^{\text{cross}_{\text{intra}}}(i,j) = \text{corr}\left(\mathbf{a}_i^{\text{PT}}, \mathbf{a}_j^{\text{SFT}}\right), \quad i \neq j \tag{5}$$

The cross-stage transfer coherence answers whether the strong pretraining performance on benchmark i can predict strong post-SFT performance on a *different* benchmark j within the same capability category. This measures *cross-stage cross-benchmark transfer reliability*—whether capability improvements during pretraining generalize to related tasks after SFT.

For each protocol, we report the category-level average by averaging over all valid pairs within the same capability category:

$$\bar{r}_{\text{intra}}(C) = \frac{2}{|C|(|C|-1)} \sum_{i<j,\ i,j \in C} r_{\text{intra}}(i,j) \tag{6}$$

3.3.3 Performance-confidence Alignment

Beyond transfer dynamics, we examine whether models develop well-calibrated representations through ***performance-confidence alignment***—the degree to which a model's confidence reflects its actual accuracy. A model exhibits strong alignment when it is confident on correct predictions and uncertain on incorrect ones; poor alignment manifests as over-confidence on errors or under-confidence on correct answers.

Accuracy-Confidence Correlation (r_{align}). To quantify alignment quality, we correlate accuracy and confidence across benchmarks within a single model configuration:

$$r_{\text{align}} = \text{corr}\,(\mathbf{a}, \mathbf{c}) \tag{7}$$

where $\mathbf{a}$ and $\mathbf{c}$ denote accuracy and confidence vectors across the benchmark suite. This metric serves as a *proxy for global calibration quality*: High r_{align} indicates the model is confident when correct and uncertain when incorrect, indicating a well-calibrated model. We use this metric to identify which pretraining data sources yield representations best aligned with target task demands.

4 Results & Findings

We present empirical observations centered around the four core research questions: ***Cross-stage transfer*** (§4.1), examining whether pretraining metrics persist after SFT; ***Benchmark reliability*** (§4.2), identifying which benchmarks serve as robust predictors across various capacity categories; ***Scaling dynamics*** (§4.3), analyzing how transfer patterns shift with model scale; and ***Performance-confidence alignment*** (§4.4), assessing calibration quality through accuracy-confidence correlation and its cross-stage transferability.

4.1 Cross-Stage Transfer (RQ1)

We first examine to what extent the task performance (*accuracy*) and calibration pattern (*confidence*) established during pretraining persist after SFT. We present the full results on cross-stage statistics and analysis in Appendix D.1.

Finding 1: Transfer reliability is strongly category-dependent. Figure 1(a) reveals that capability categories exhibit markedly different transfer profiles. Specifically, *Science* and *Commonsense* benchmarks show consistently high cross-stage accuracy correlation across both model scales ($\bar{r} > 0.5$), indicating that pretraining performance on these domains reliably predicts post-SFT outcomes. In contrast, *Semantic* tasks demonstrate weaker transferability, suggesting that SFT fundamentally reorganizes how models approach these linguistically nuanced tasks rather than building upon pretraining representations.

Finding 2: Confidence patterns persist more strongly than accuracy for reasoning tasks. Beyond the commonly-adopted accuracy metric, we examine whether confidence calibration established during pretraining persists after SFT. Analyzing the diagonal of the PT→SFT confidence correlation matrix (Figure 3), we find remarkably strong persistence, especially at 240M model scale: mean benchmark-wise confidence correlation is $\bar{r} = 0.68$, with all evaluated benchmarks showing positive cross-stage correlation and 79% exceeding $r > 0.5$.

This persistence is strongly category-dependent. *Commonsense* and *Science* benchmarks exhibit strong cross-stage confidence transfer ($\bar{r}$ = 0.87, 0.82), indicating that a model's confidence profile on physical, social and scientific reasoning is largely determined during the pretraining stage. In contrast, *NLI* tasks show weak confidence persistence ($\bar{r} = 0.21$), suggesting that SFT substantially re-calibrates uncertainty on natural language inference tasks.

The practical implication is significant: for Commonsense and Science benchmarks, *confidence-based model selection during pretraining remains a valid proxy for post-SFT calibration quality*, enabling early identification of well-calibrated models without expensive SFT iterations.

Finding 3: Cross-benchmark confidence structure persists across training stages. A stronger form of persistence emerges when comparing the *structure* of confidence correlations across stages. The PT-PT heatmap (Figure 4a) and SFT-SFT heatmap (Figure 4b) demonstrate the cross-benchmark predictability of confidence scores within the same training stage. According to Figure 4a and 4b, at 240M exhibit striking structural similarity: these two within-stage correlation matrices are highly correlated with $r = 0.73, p < 10^{-32}$, indicating that *benchmark pairs that co-vary in confidence during pretraining continue to co-vary similarly after SFT*.

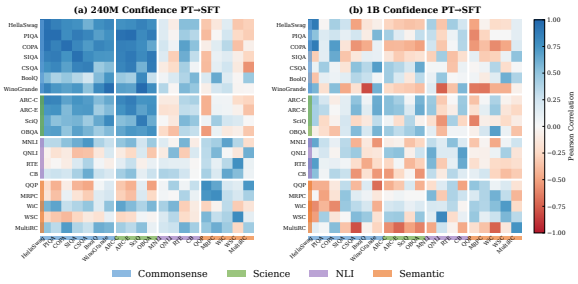


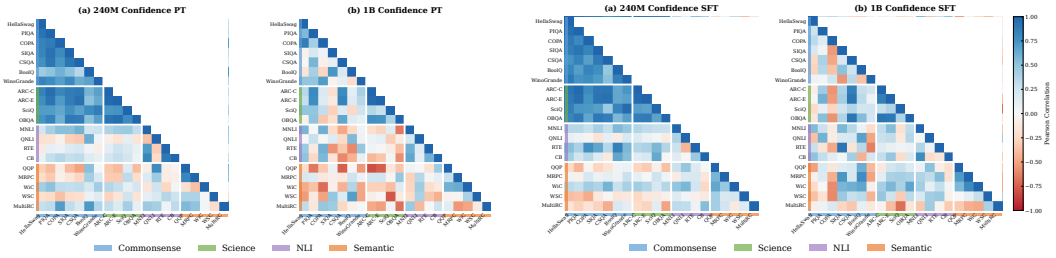
Figure 3: **Cross-stage confidence correlation (PT→SFT).** Each cell shows the Pearson correlation between benchmark i's PT confidence and benchmark j's SFT confidence across data mixtures; the diagonal represents the benchmark-wise transfer pattern. **Left**: At 240M, the *Commonsense*–*Science* block shows high ***positive*** correlations; **Right**: At 1B, greater heterogeneity emerges with ***negative*** correlations.

This cross-benchmark cross-stage consistency is particularly strong for *Commonsense* and *Science* categories. The *Commonsense–Science* "confidence block" not only persists but slightly *strengthens* through SFT, suggesting that SFT reinforces rather than disrupts the shared calibration structure between these reasoning categories.

4.2 Benchmark Reliability (RQ2)

We further conduct benchmark-wise analysis to investigate which benchmarks serve as reliable predictors of post-SFT performance and which should be treated with caution during early-stage evaluation. Detailed statistics are provided in Appendix D.2.

Finding 4: Certain benchmarks are unreliable cross-stage predictors. Figure 2(a) reveals that individual benchmarks vary dramatically in their predictive reliability. WiC and MultiRC exhibit particularly weak or negative accuracy correlations ($r_{\text{acc}}^{\text{stage}} < 0.3$ across scales), indicating that strong pretraining performance on these benchmarks does not guarantee post-SFT success. Similarly, WinoGrande and MNLI show negative transfer at 240M scale ($r = -0.31$ and $r = -0.34$ respectively). These unreliable benchmarks share a common characteristic: they require nuanced linguistic understanding that appears to be substantially reorganized during instruction tuning. *Practitioners should exercise caution when using these benchmarks for early-stage model selection or data curation decisions.*



(a) PT confidence correlation

(b) SFT confidence correlation

Figure 4: **Within-stage confidence correlation comparison.** (a) PT-PT and (b) SFT-SFT cross-benchmark confidence correlations at 240M and 1B scale. The *Commonsense*–*Science* block is nearly identical across stages, demonstrating that confidence correlation patterns established during pretraining persist through SFT.

Finding 5: Intra-category competition undermines single-benchmark evaluation. Beyond individual benchmark reliability, we examine whether benchmarks within the same capacity category respond coherently to data mixture changes. As shown in Figure 5, at both model scales, *Commonsense* and *Semantic* categories consistently exhibit negative intra-category coherence on accuracy.

This "*intra-category competition*" implies that data mixtures optimizing one benchmark often degrade its semantic neighbors—a critical warning for practitioners who assume that improving HellaSwag indicates general improvements in Commonsense Reasoning. Notably, *Science* demonstrate positive intra-category coherence on large-scale model (Figure 5 (b)), suggesting more unified underlying representations for scientific knowledge.

Finding 6: Certain benchmarks behave incoherently within their semantic categories. Beyond cross-stage reliability, we identify benchmarks that fail to co-vary with other benchmarks in the same capability category—acting as "black sheep" that respond differently to data mixture changes than semantically related tasks. Within *Commonsense* category, WinoGrande exhibits consistently negative correlations with other category members (mean pairwise $r = -0.18$ at 240M); within *Semantic* group, WSC shows weak coherence with $r < 0.15$, indicating that coreference resolution may not share underlying representations with other linguistic understanding benchmarks.

These incoherent benchmarks pose a subtle risk: *even when a benchmark shows reasonable cross-stage transfer individually, the low category coherence means that the task-specific improvements cannot generalize to semantically related tasks.*

4.3 Scaling Dynamics (RQ3)

We now examine how transfer patterns shift as models scale from 240M to 1B, revealing systematic differences in accuracy and confidence patterns across scales.

Finding 7: Accuracy and confidence transfer exhibit inverse scaling dynamics. Figure 2 reveals a striking dissociation between accuracy and confidence as models scale. For *accuracy*, the 1B model achieves stronger cross-stage transfer than 240M ($\bar{r}_{\text{acc}} = 0.60$ vs. 0.51), consistent with the intuition that larger models learn more transferable representations. However, for *confidence*, the pattern *reverses*: 240M maintains substantially higher cross-stage correlation ($\bar{r}_{\text{conf}} = 0.68$ vs. 0.39). This inverse scaling dynamic suggests that larger models undergo more extensive confidence reorganization during SFT, developing task-specific calibration profiles rather than retaining the monolithic uncertainty patterns observed on small-scale models. The implication for practitioners is significant: *as models scale, confidence calibration from pretraining becomes less predictive of post-SFT behavior, warranting explicit calibration techniques for SFT.*

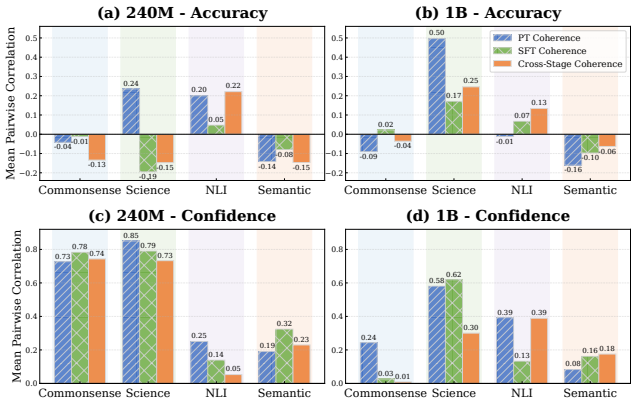


Figure 5: **Intra-category coherence across three correlation protocols. Top**: the coherence scores on *accuracy*, where Science shows PT→SFT degradation; NLI preserves cross-stage coherence despite a drop in SFT coherence. **Bottom**: the coherence scores on *confidence*, where 240M model maintains high coherence while 1B shows substantial degradation.

Finding 8: Scaling induces a transition from intra-category competition to synergy in accuracy. Figure 5 reveals that model scale fundamentally alters intra-category dynamics for accuracy. For both *SFT coherence* and *cross-stage coherence*, most capability categories shift toward positive coherence at 1B scale. The *Science* category exhibits particularly significant improvements across all three protocols with 240M→1B scaling (PT: $0.24 \rightarrow 0.50$; SFT: $-0.19 \rightarrow 0.17$; Cross-stage: $-0.15 \rightarrow 0.25$). This suggests that larger models develop more unified reasoning representations where knowledge transfers constructively across related benchmarks. It implies that *at smaller scales, optimizing for a single benchmark per category may harm related tasks; while at larger scales, the generalization within the same capacity category becomes more viable.*

Finding 9: Confidence coherence degrades with scale while remaining positive. For confidence calibration, we observe the opposite scaling trend (Figure 5). At 240M, *Commonsense* and *Science* exhibit strong positive coherence ($\bar{r} = 0.73$ and 0.85 respectively); while at 1B, this coherence diminishes substantially ($\bar{r} = 0.24$ and 0.58). Rather than indicating performance degradation, this drop reflects the development of more *task-specific uncertainty estimates* on larger language models—when scaling-up, the model learns to calibrate confidence differently for distinct tasks or benchmarks, instead of adopting a universal uniform calibration pattern of smaller models.

Finding 10: NLI exhibits anomalous scaling patterns. The NLI category presents a distinct scaling patterns varying from other benchmarks described above. For confidence coherence, NLI is the *only* category where 1B shows higher coherence than 240M —the inverse of the trend where scale typically reduces confidence coherence. This twofold anomaly suggests that NLI tasks engage qualitatively different representational mechanisms. One hypothesis is that larger models develop

more consistent inferential strategies for NLI specifically, improving calibration coherence even as other categories become more task-specifically calibrated.

4.4 Calibration Quality (RQ4)

Finally, we examine *performance-confidence alignment*—the correlation between model confidence and accuracy scores (r_{align})—as a measure of *calibration quality*. A well-calibrated model should be confident precisely when it is correct and uncertain when it is incorrect, yielding high r_{align}. We investigate how this alignment varies across capability categories, data mixtures, and training stages.

Finding 11: Performance-confidence alignment varies significantly across capability categories. Figure 6 reveals striking differences in how well models' confidence aligns with actual performance across capability domains. *Science* tasks exhibit remarkably high alignment ($r_{\text{align}} \approx 0.8$), indicating that models are confident precisely when they perform well—the prior knowledge from pretraining aligns well with downstream task requirements. In stark contrast, *Commonsense* shows consistently weak or negative alignment ($r_{\text{align}} \approx -0.1$), meaning models tend to be *more* confident on incorrect predictions. *Semantic* tasks exhibit similar miscalibration ($r_{\text{align}} \approx -0.2$). This systematic miscalibration suggests a fundamental mismatch between pretraining knowledge and the implicit reasoning required for commonsense understanding—the model's confidence reflects surface-level pattern matching rather than genuine comprehension.

Finding 12: Alignment patterns persist through SFT. Comparing the left and right panels of Figure 6, the category-level alignment structure remains largely stable across training stages. *Science* tasks maintain high r_{align} values ($\sim$0.8) in both pretraining and post-SFT evaluation, while the miscalibration for *Commonsense* and *Semantic* also persists. This suggests that SFT does not fundamentally reorganize the model's confidence structure—the alignment "fingerprint" established during pretraining carries through to downstream performance. Combined with our cross-stage analysis (§4.1), this finding reinforces that *pretraining data composition has lasting effects on model behavior that cannot be easily overwritten by instruction tuning*. Practitioners seeking well-calibrated models should prioritize appropriate pretraining data curation over post-hoc calibration techniques.

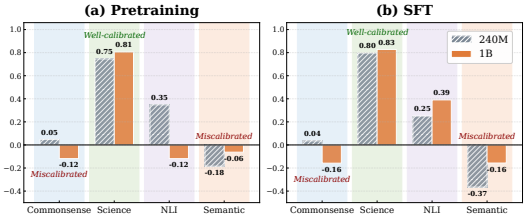


Figure 6: **Performance-confidence alignment (r_{align}) varies by category.** Science tasks show positive alignment, indicating well-calibrated confidence. Commonsense and Semantic show negative alignment, indicating systematic overconfidence on incorrect predictions.

Finding 13: Educational content filtering exhibits scale-dependent accuracy-calibration trade-offs. Figure 7 reveals that the effects of educational content filtering (FineWeb-Edu) vary dramatically with model scale. At 240M, FineWeb-Edu *improves* NLI accuracy by +5.0pp compared to RefinedWeb, yet simultaneously *degrades* alignment from $r_{\text{align}} = 0.68$ to $r_{\text{align}} = -0.12$ ($\Delta = -0.80$). Remarkably, at 1B scale, this pattern *reverses*: FineWeb-Edu *degrades* NLI accuracy by -4.4pp while slightly *improving* alignment ($\Delta = +0.16$). On *Scientific Reasoning* tasks, exhibits consistent benefits across scales: FineWeb-Edu improves accuracy at both 240M (+2.1pp) and 1B (+2.5pp), with slight alignment degradation. It indicates that: *data curation decisions optimized at small proxy scales may not transfer to larger models*; and *strict educational filtering could preserves or improve on accuracy scores while undermining internal calibration pattern.*

5 Conclusion

We systematically analyze capability transfer from pretraining to SFT using correlation-based protocols across diverse data mixtures, model scales, and a large variety of evaluation benchmarks. Our findings provide practical guidance for early-stage benchmark selection and data mixture design, while highlighting the lasting impact of pretraining decisions on model behavior. We provide detailed practical implementation in Appendix A.

Limitations and Future Direction. Our study has several limitations that suggest directions for future work. Because of resource constraints, our experiments are conducted at 240M and 1B parameter scales using a single SFT dataset (`Tulu-v2-mix`), which may not fully capture phenomena at larger scales or under different post-training regimes. Additionally, while we evaluate 20 benchmarks across diverse capability categories, important domains such as long-context reasoning, code generation, and safety remain unexplored. We provide a detailed discussion of limitations and future directions in Appendix B.

REFERENCES

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language, 2019. URL https://arxiv.org/abs/1911.11641.

Samuel R. Bowman and George Dahl. What will it take to fix benchmarking in natural language understanding? In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4843–4855, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.385. URL https://aclanthology.org/2021.naacl-main.385/.

Hao Chen, Abdul Waheed, Xiang Li, Yidong Wang, Jindong Wang, Bhiksha Raj, and Marah I. Abdin. On the diversity of synthetic data and its impact on training large language models, 2024. URL https://arxiv.org/abs/2410.15226.

Yangyi Chen, Binxuan Huang, Yifan Gao, Zhengyang Wang, Jingfeng Yang, and Heng Ji. Scaling laws for predicting downstream performance in llms, 2025. URL https://arxiv.org/abs/2410.08527.

Zhoujun Cheng, Shibo Hao, Tianyang Liu, Fan Zhou, Yutao Xie, Feng Yao, Yuexin Bian, Yonghao Zhuang, Nilabjo Dey, Yuheng Zha, Yi Gu, Kun Zhou, Yuqi Wang, Yuan Li, Richard Fan, Jianshu She, Chengqian Gao, Abulhair Saparov, Haonan Li, Taylor W. Killian, Mikhail Yurochkin, Zhengzhong Liu, Eric P. Xing, and Zhiting Hu. Revisiting reinforcement learning for llm reasoning from a cross-domain perspective, 2025. URL https://arxiv.org/abs/2506.14965.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models, 2022. URL https://arxiv.org/abs/2210.11416.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018. URL https://arxiv.org/abs/1803.05457.

DatologyAI, :, Pratyush Maini, Vineeth Dorna, Parth Doshi, Aldo Carranza, Fan Pan, Jack Urbanek, Paul Burstein, Alex Fang, Alvin Deng, Amro Abbas, Brett Larsen, Cody Blakeney, Charvi Bannur, Christina Baek, Darren Teh, David Schwab, Haakon Mongstad, Haoli Yin, Josh Wills, Kaleigh Mentzer, Luke Merrick, Ricardo Monti, Rishabh Adiga, Siddharth Joshi, Spandan Das, Zhengping Wang, Bogdan Gaza, Ari Morcos, and Matthew Leavitt. Beyondweb: Lessons from scaling synthetic data for trillion-scale pretraining, 2025. URL https://arxiv.org/abs/2508.10975.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. Investigating data contamination in modern benchmarks for large language models, 2024. URL https://arxiv.org/abs/2311.09783.

Shrey Desai and Greg Durrett. Calibration of pre-trained transformers. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 295–302, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.21. URL https://aclanthology.org/2020.emnlp-main.21/.

William B. Dolan and Chris Brockett. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, 2005. URL https://aclanthology.org/I05-5002/.

Simin Fan, Matteo Pagliardini, and Martin Jaggi. Doge: Domain reweighting with generalization estimation, 2024. URL https://arxiv.org/abs/2310.15393.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. A survey of confidence estimation and calibration in large language models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6577–6595, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.366. URL https://aclanthology.org/2024.naacl-long.366/.

Roger Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, Evan Hubinger, Kamilė Lukošiūtė, Karina Nguyen, Nicholas Joseph, Sam McCandlish, Jared Kaplan, and Samuel R. Bowman. Studying large language model generalization with influence functions, 2023. URL https://arxiv.org/abs/2308.03296.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. Textbooks are all you need, 2023. URL https://arxiv.org/abs/2306.11644.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks, 2017. URL https://arxiv.org/abs/1706.04599.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645 (8081):633–638, September 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL http://dx.doi.org/10.1038/s41586-025-09422-z.

Danny Hernandez, Jared Kaplan, Tom Henighan, and Sam McCandlish. Scaling laws for transfer, 2021. URL https://arxiv.org/abs/2102.01293.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models, 2022. URL https://arxiv.org/abs/2203.15556.

Berivan Isik, Natalia Ponomareva, Hussein Hazimeh, Dimitris Paparas, Sergei Vassilvitskii, and Sanmi Koyejo. Scaling laws for downstream task performance in machine translation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=vPOMTkmSiu.

Hamish Ivison, Yizhong Wang, Valentina Pyatkin, Nathan Lambert, Matthew Peters, Pradeep Dasigi, Joel Jang, David Wadden, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. Camels in a changing climate: Enhancing lm adaptation with tulu 2, 2023. URL https://arxiv.org/abs/2311.10702.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know, 2022. URL https://arxiv.org/abs/2207.05221.

Feiyang Kang, Newsha Ardalani, Michael Kuchnik, Youssef Emad, Mostafa Elhoushi, Shubhabrata Sengupta, Shang-Wen Li, Ramya Raghavendra, Ruoxi Jia, and Carole-Jean Wu. Demystifying synthetic data in llm pre-training: A systematic study of scaling laws, benefits, and pitfalls, 2025. URL https://arxiv.org/abs/2510.01631.

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, 2020. URL https://arxiv.org/abs/2001.08361.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 252–262, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1023. URL https://aclanthology.org/N18-1023/.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. The winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, pp. 552–561. AAAI Press, 2012. ISBN 9781577355601.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel, Jean Mercat, Mayee Chen, Suchin Gururangan, Mitchell Wortsman, Alon Albalak, Yonatan Bitton, Marianna Nezhurina, Amro Abbas, Cheng-Yu Hsieh, Dhruba Ghosh, Josh Gardner, Maciej Kilian, Hanlin Zhang, Rulin Shao, Sarah Pratt, Sunny Sanyal, Gabriel Ilharco, Giannis Daras, Kalyani Marathe, Aaron Gokaslan, Jieyu Zhang, Khyathi Chandu, Thao Nguyen, Igor Vasiljevic, Sham Kakade, Shuran Song, Sujay Sanghavi, Fartash Faghri, Sewoong Oh, Luke Zettlemoyer, Kyle Lo, Alaaeldin El-Nouby, Hadi Pouransari, Alexander Toshev, Stephanie Wang, Dirk Groeneveld, Luca Soldaini, Pang Wei Koh, Jenia Jitsev, Thomas Kollar, Alexandros G. Dimakis, Yair Carmon, Achal Dave, Ludwig Schmidt, and Vaishaal Shankar. Datacomp-lm: In search of the next generation of training sets for language models, 2025. URL https://arxiv.org/abs/2406.11794.

Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, Qian Liu, Evgenii Zheltonozhskii,

Terry Yue Zhuo, Thomas Wang, Olivier Dehaene, Mishig Davaadorj, Joel Lamy-Poirier, João Monteiro, Oleh Shliazhko, Nicolas Gontier, Nicholas Meade, Armel Zebaze, Ming-Ho Yee, Logesh Kumar Umapathi, Jian Zhu, Benjamin Lipkin, Muhtasham Oblokulov, Zhiruo Wang, Rudra Murthy, Jason Stillerman, Siva Sankalp Patel, Dmitry Abulkhanov, Marco Zocca, Manan Dey, Zhihan Zhang, Nour Fahmy, Urvashi Bhattacharyya, Wenhao Yu, Swayam Singh, Sasha Luccioni, Paulo Villegas, Maxim Kunakov, Fedor Zhdanov, Manuel Romero, Tony Lee, Nadav Timor, Jennifer Ding, Claire Schlesinger, Hailey Schoelkopf, Jan Ebert, Tri Dao, Mayank Mishra, Alex Gu, Jennifer Robinson, Carolyn Jane Anderson, Brendan Dolan-Gavitt, Danish Contractor, Siva Reddy, Daniel Fried, Dzmitry Bahdanau, Yacine Jernite, Carlos Muñoz Ferrandis, Sean Hughes, Thomas Wolf, Arjun Guha, Leandro von Werra, and Harm de Vries. Starcoder: may the source be with you!, 2023. URL https://arxiv.org/abs/2305.06161.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models, 2023. URL https://arxiv.org/abs/2211.09110.

Nicholas Lourie, Michael Y. Hu, and Kyunghyun Cho. Scaling laws are unreliable for downstream tasks: A reality check, 2025. URL https://arxiv.org/abs/2507.00885.

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

Anton Lozhkov, Raymond Li, Loubna Ben Allal, Federico Cassano, Joel Lamy-Poirier, Nouamane Tazi, Ao Tang, Dmytro Pykhtar, Jiawei Liu, Yuxiang Wei, Tianyang Liu, Max Tian, Denis Kocetkov, Arthur Zucker, Younes Belkada, Zijian Wang, Qian Liu, Dmitry Abulkhanov, Indraneil Paul, Zhuang Li, Wen-Ding Li, Megan Risdal, Jia Li, Jian Zhu, Terry Yue Zhuo, Evgenii Zheltonozhskii, Nii Osae Osae Dade, Wenhao Yu, Lucas Krauß, Naman Jain, Yixuan Su, Xuanli He, Manan Dey, Edoardo Abati, Yekun Chai, Niklas Muennighoff, Xiangru Tang, Muhtasham Oblokulov, Christopher Akiki, Marc Marone, Chenghao Mou, Mayank Mishra, Alex Gu, Binyuan Hui, Tri Dao, Armel Zebaze, Olivier Dehaene, Nicolas Patry, Canwen Xu, Julian McAuley, Han Hu, Torsten Scholak, Sebastien Paquet, Jennifer Robinson, Carolyn Jane Anderson, Nicolas Chapados, Mostofa Patwary, Nima Tajbakhsh, Yacine Jernite, Carlos Muñoz Ferrandis, Lingming Zhang, Sean Hughes, Thomas Wolf, Arjun Guha, Leandro von Werra, and Harm de Vries. Starcoder 2 and the stack v2: The next generation, 2024b. URL https://arxiv.org/abs/2402.19173.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering, 2018. URL https://arxiv.org/abs/1809.02789.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. 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 Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. Trak: Attributing model behavior at scale, 2023. URL https://arxiv.org/abs/2303.14186.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data, and web data only, 2023. URL https://arxiv.org/abs/2306.01116.

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, 2024. URL https://arxiv.org/abs/2406.17557.

Mohammad Taher Pilehvar and Jose Camacho-Collados. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 1267–1273, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1128. URL https://aclanthology.org/N19-1128/.

Pedro Rodriguez, Joe Barrow, Alexander Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. Evaluation examples are not equally informative: How should that change NLP leaderboards? In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4486–4503, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.346. URL https://aclanthology.org/2021.acl-long.346/.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10776–10787, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.722. URL https://aclanthology.org/2023.findings-emnlp.722/.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale, 2019. URL https://arxiv.org/abs/1907.10641.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. Socialiqa: Commonsense reasoning about social interactions, 2019. URL https://arxiv.org/abs/1904.09728.

Rylan Schaeffer, Hailey Schoelkopf, Brando Miranda, Gabriel Mukobi, Varun Madan, Adam Ibrahim, Herbie Bradley, Stella Biderman, and Sanmi Koyejo. Why has predicting downstream capabilities of frontier AI models with scale remained elusive? In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=I1NtlLvJal.

Lakshay Sharma, Laura Graesser, Nikita Nangia, and Utku Evci. Natural language understanding with the quora question pairs dataset, 2019. URL https://arxiv.org/abs/1907.01041.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge, 2019. URL https://arxiv.org/abs/1811.00937.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback, 2023. URL https://arxiv.org/abs/2305.14975.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding, 2019. URL https://arxiv.org/abs/1804.07461.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems, 2020. URL https://arxiv.org/abs/1905.00537.

Yiping Wang, Qing Yang, Zhiyuan Zeng, Liliang Ren, Liyuan Liu, Baolin Peng, Hao Cheng, Xuehai He, Kuan Wang, Jianfeng Gao, Weizhu Chen, Shuohang Wang, Simon Shaolei Du, and Yelong Shen. Reinforcement learning for reasoning in large language models with one training example, 2025. URL https://arxiv.org/abs/2504.20571.

Maurice Weber, Daniel Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, Ben Athiwaratkun, Rahul Chalamala, Kezhen Chen, Max Ryabinin, Tri Dao, Percy Liang, Christopher Ré, Irina Rish, and Ce Zhang. Redpajama: an open dataset for training large language models, 2024. URL https://arxiv.org/abs/2411.12372.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions, 2017. URL https://arxiv.org/abs/1707.06209.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference, 2018. URL https://arxiv.org/abs/1704.05426.

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy Liang, Quoc V. Le, Tengyu Ma, and Adams Wei Yu. Doremi: Optimizing data mixtures speeds up language model pretraining, 2023. URL https://arxiv.org/abs/2305.10429.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence?, 2019. URL https://arxiv.org/abs/1905.07830.

A Practical Implications

Our findings offer several actionable insights for practitioners developing and evaluating language models.

Benchmark selection for early-stage evaluation. Not all benchmarks are equally informative during pretraining evaluation. According to the clear dichotomy we have observed: *Commonsense Reasoning* benchmarks (e.g. HellaSwag, PIQA, COPA) and *Scientific Reasoning* benchmarks exhibit high cross-stage correlation, making them reliable proxies for post-SFT performance; while *Semantic Understanding* tasks (WiC, WSC, MultiRC) and *NLI* benchmarks (MNLI) show weak or negative transfer, suggesting their pretraining performance provides limited signal about final model capabilities.

Implication 1: *Practitioners should prioritize high-transfer benchmarks while avoiding unreliable ones when making early-stage decisions about data curation and resource allocation.*

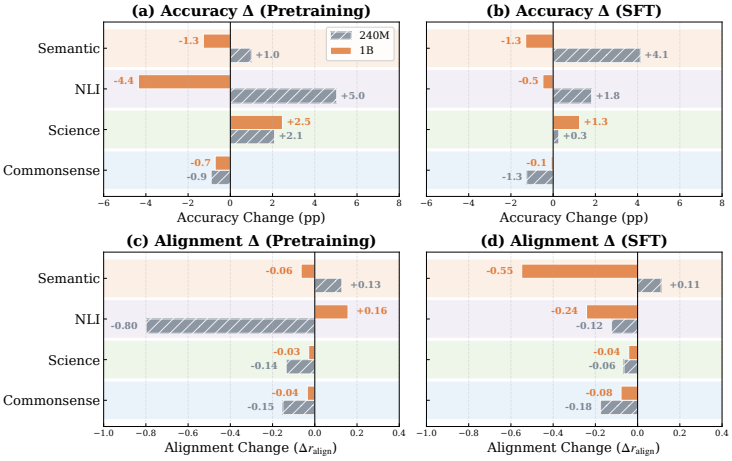


Figure 7: **FineWeb-Edu vs. RefinedWeb: Category-level effects at PT and SFT stages.** (a,b) Accuracy changes; (c,d) Alignment changes. Left column: Pretraining; Right column: SFT. The NLI accuracy reversal (240M: +5.0pp $\rightarrow$ 1B: -4.4pp) persists through SFT, though attenuated. At 1B SFT, Semantic alignment shows severe degradation ($\Delta = -0.55$), suggesting that FineWeb-Edu's effects on linguistic tasks become increasingly problematic at larger scales and later training stages.

Confidence as a Complementary Evaluation Signal. We demonstrate that confidence correlation provides information complementary to accuracy, with important scale-dependent dynamics. While confidence patterns transfer reliably across stages at smaller scales (240M) reflecting a relatively consistent calibration profile that persists through SFT, confidence transferability degrades substantially when the model scaled up to 1B parameters. This suggests that larger models develop more *nuanced, task-specific calibration profiles* during SFT, flipping the monolithic uncertainty patterns of smaller models.

Implication 2: *As models scale, practitioners cannot assume that well-calibrated pretraining confidence will carry over to post-SFT deployment. Task- and stage-specific calibration becomes increasingly critical at larger scales.*

Scale-Dependent Data Curation Effects. Our analysis reveals that aggressive educational content filtering (FineWeb-Edu) produces scale-dependent effects that challenge the common practice of using small proxy models for data curation decisions. For NLI, FineWeb-Edu yields +5.0pp accuracy gains at 240M but -4.4pp *degradation* at 1B—a complete reversal (Figure 7). Meanwhile, alignment effects also reverse: severe degradation at 240M ($\Delta = -0.80$) but slight improvement at 1B ($\Delta = +0.16$).

Implication 3: *Practitioners should exercise caution when extrapolating data curation decisions from small-scale experiments to larger models. For robust data curation, we recommend evaluating mixture effects at multiple scales before committing to production training.*

B Limitations and Future Work

Model Scale. Our experiments are conducted at 240M and 1B parameter scales, which may not fully capture phenomena that emerge at larger scales. The inverse scaling trends we observe for accuracy versus confidence transfer are particularly intriguing and warrant investigation at 7B+ scales to determine whether these dynamics persist, amplify, or reverse. We leave further scaling efforts beyond the *small language model* (SLM) scope for future work.

SFT Data and Protocol. We use a single SFT dataset (`tulu-v2-mix`) and training protocol across all experiments. Different instruction-tuning datasets or multi-stage post-training pipelines (e.g., incorporating RLHF or DPO) may exhibit different transfer dynamics. Our findings are most directly applicable to standard SFT workflows and may not generalize to more complex post-training regimes.

Benchmark Coverage. While we evaluate on 22 benchmarks spanning 8 capability categories, our analysis necessarily omits many important capabilities, including long-context reasoning, multi-turn dialogue, code generation, and safety-related behaviors. The transfer patterns we identify may not generalize to these domains, particularly those that are specifically targeted by instruction-tuning data.

Data Mixture Granularity. Our 9 data mixtures, while systematically varied, represent a sparse sampling of the possible mixture space. Finer-grained variations in code proportion, web source combinations, or the inclusion of additional data types may reveal patterns not captured in our analysis. Applying our proposed correlation protocols to study the impact of ***knowledge-enriched content*** such as textbooks (Gunasekar et al., 2023) or ***synthetic data*** (Kang et al., 2025; Chen et al., 2024; DatologyAI et al., 2025) across training stages is a promising direction for future work.

C Experimental Setup

C.1 Model Architecture

We train decoder-only transformer models at two parameter scales. Table 1 summarizes the architectural configurations.

Table 1: Model architecture configurations.

Hyperparameter	240M	1B
Model dimension	768	1680
Per head dimension	256	256
Number of heads	8	8
Number of layers	8	12
Expand factor	8	8
Context length	4096	4096
Vocabulary size	100,864	100,864
Total parameters	240M	1.0B

C.2 Training Hyperparameters

Pretraining. Table 2a details the pretraining hyperparameters for both model scales. We use the AdamW optimizer with a cosine learning rate schedule and linear warmup.

Supervised Fine-tuning. Following pretraining, we fine-tune each checkpoint on `tulu-v2-mix` (Ivison et al., 2023), a curated collection of instruction-following datasets including FLAN-v2, Open Assistant, ShareGPT, and others. Table 2b summarizes SFT hyperparameters.

Table 2: Training hyperparameters for pretraining (left) and supervised fine-tuning (right).

(a) Pretraining

Hyperparameter	240M	1B
Batch size	256	256
Sequence length	4096	4096
Training steps	200,000	200,000
Learning rate	0.001	0.001
LR schedule	Cosine	Cosine
Warmup steps	2,000	2,000
Weight decay	0.001	0.001
End decay	0.1	0.1
Gradient clipping	1.0	1.0
Adam β_1	0.9	0.9
Adam β_2	0.95	0.95
Precision	bfloat16	bfloat16

(b) Supervised Fine-tuning

Hyperparameter	240M	1B
SFT dataset	`tulu-v2-mix`	
Training Steps	6,365	12,730
Epochs	5	10
Batch size	256	256
Sequence length	4096	4096
Learning rate	2e-5	1e-5
LR schedule	Cosine	Cosine
Weight Decay	0.001	0.001
Warmup steps	43	43
Precision	bfloat16	bfloat16

C.3 Pretraining Data Mixtures

We construct 9 pretraining data mixtures by systematically varying two orthogonal dimensions: (1) the *source* of web-crawled data, and (2) the *proportion* of code versus web content. This factorial design enables us to disentangle the effects of data quality from data composition.

Data Sources. Each mixture combines three categories of training data:

- **Web-crawled data**: General web text from one of three sources—RefinedWeb (Penedo et al., 2023), FineWeb-Edu (Lozhkov et al., 2024a), or DCLM (Li et al., 2025). These sources differ in filtering strategies: RefinedWeb applies minimal quality filtering, FineWeb-Edu aggressively filters for educational content, and DCLM uses model-based quality scoring.
- **Code data**: Programming-related content from either StarCoder (Li et al., 2023) or The Stack v2 (Lozhkov et al., 2024b), covering diverse programming languages and software documentation.
- **Curated knowledge**: High-quality reference material from RedPajama-v2 (Weber et al., 2024), including Wikipedia, ArXiv papers, StackExchange discussions, and books.

Mixture Naming Convention. Each mixture is denoted $\mathtt{V}x\mathtt{P}y$ where:

- $x \in \{1, 2, 3\}$ indicates the web data source: V1 = RefinedWeb, V2 = FineWeb-Edu, V3 = DCLM
- $y \in \{0, 1, 2\}$ indicates the proportion configuration: P0 = web-heavy, P1 = balanced, P2 = code-heavy

Proportion Configurations. The Py suffix controls the balance between data categories:

- **P0 (Web-heavy)**: 65% web, 25% code, 10% curated
- **P1 (Balanced)**: 45% web, 35% code, 20% curated
- **P2 (Code-heavy)**: 25% web, 45% code, 30% curated

Table 3 provides the complete specification of all 9 mixtures.

Table 3: Pretraining data mixture compositions. All mixtures include curated knowledge sources from RedPajama-v2 (Wikipedia, ArXiv, StackExchange, Books) at the specified proportion.

Mixture	Web (%)	Code (%)	Curated (%)
V1P0	RefinedWeb (65)	StarCoder (25)	10
V1P1	RefinedWeb (45)	StarCoder (35)	20
V1P2	RefinedWeb (25)	StarCoder (45)	30
V2P0	FineWeb-Edu (65)	Stack-v2 (25)	10
V2P1	FineWeb-Edu (45)	Stack-v2 (35)	20
V2P2	FineWeb-Edu (25)	Stack-v2 (45)	30
V3P0	DCLM (65)	Stack-v2 (25)	10
V3P1	DCLM (45)	Stack-v2 (35)	20
V3P2	DCLM (25)	Stack-v2 (45)	30

This design enables systematic analysis of: (1) *web source effects* by comparing V1 vs. V2 vs. V3 at fixed proportions, and (2) *code proportion effects* by comparing P0 vs. P1 vs. P2 within each web source.

C.4 Evaluation Benchmarks

We evaluate model performance across 20 benchmarks organized into 4 capability categories: *Commonsense*, *Science*, *NLI*, and *Semantic*. Table 4 provides complete benchmark details.

Table 4: Evaluation benchmarks organized by capability category. |Test| indicates evaluation set size. All benchmarks are evaluated in a multiple-choice format using log-probability scoring.

Category	Benchmark	|Test|	Description
Commonsense	HellaSwag (Zellers et al., 2019)	10,042	Sentence completion requiring physical commonsense
	PIQA (Bisk et al., 2019)	1,838	Physical intuition about everyday objects and actions
	COPA (Wang et al., 2020)	500	Causal reasoning about everyday events
	SIQA (Sap et al., 2019)	1,954	Social and emotional intelligence reasoning
	CommonsenseQA (Talmor et al., 2019)	1,221	General commonsense knowledge questions
	WinoGrande (Sakaguchi et al., 2019)	1,267	Pronoun resolution requiring commonsense
	BoolQ (Wang et al., 2020)	3,270	Yes/no questions requiring world knowledge
Science	ARC-Challenge (Clark et al., 2018)	1,172	Grade-school science questions (hard subset)
	ARC-Easy (Clark et al., 2018)	2,376	Grade-school science questions (easy subset)
	SciQ (Welbl et al., 2017)	1,000	Science exam questions with supporting passages
	OpenBookQA (Mihaylov et al., 2018)	500	Elementary science requiring external knowledge
NLI	MNLI (Williams et al., 2018)	9,815	Multi-genre natural language inference
	QNLI (Wang et al., 2019)	5,463	Question-answering NLI (from SQuAD)
	RTE (Wang et al., 2020)	277	Recognizing textual entailment
	CB (Wang et al., 2020)	56	CommitmentBank textual entailment
Semantic	QQP (Sharma et al., 2019)	40,430	Quora question paraphrase detection
	MRPC (Dolan & Brockett, 2005)	408	Microsoft paraphrase corpus
	WiC (Pilehvar & Camacho-Collados, 2019)	638	Word-in-context sense disambiguation
	WSC (Levesque et al., 2012)	104	Winograd schema coreference resolution
	MultiRC (Khashabi et al., 2018)	4,848	Multi-sentence reading comprehension

Evaluation Protocol. All benchmarks are evaluated using a multiple-choice format with log-probability scoring. For each example, we compute the log-probability of each candidate answer conditioned on the prompt and select the answer with the highest probability. We report two metrics:

- **Accuracy**: The proportion of examples where the model selects the correct answer.
- **Confidence**: The probability assigned to the selected answer (correct or incorrect), averaged across the evaluation set. This measures the model's certainty in its predictions.

D Additional Results

D.1 Cross-Stage Correlation Results

This section presents complete cross-benchmark correlation heatmaps for both accuracy and confidence metrics across training stages and model scales.

D.1.1 Accuracy Correlations

Figures 8–10 show pairwise benchmark correlations during pretraining, after SFT, and across stages (PT→SFT).

Key observations: During pretraining (Figure 8), the 240M model exhibits widespread negative correlations—particularly between Commonsense benchmarks and other categories—indicating that data mixtures optimizing one capability often degrade others. The 1B model shows more coherent positive structure, especially within the Science block, suggesting that larger models learn more generalizable representations that benefit multiple benchmarks simultaneously.

After SFT (Figure 9), correlation patterns largely persist but with notable changes: the NLI block becomes more internally coherent at 1B, while cross-category relationships weaken at 240M. This suggests SFT primarily preserves rather than reorganizes the benchmark relationships established during pretraining.

The PT→SFT transfer heatmap (Figure 10) reveals which pretraining benchmarks predict which SFT outcomes. HellaSwag emerges as a robust predictor across both scales, while WiC and MNLI show inconsistent or negative transfer, particularly at 240M.

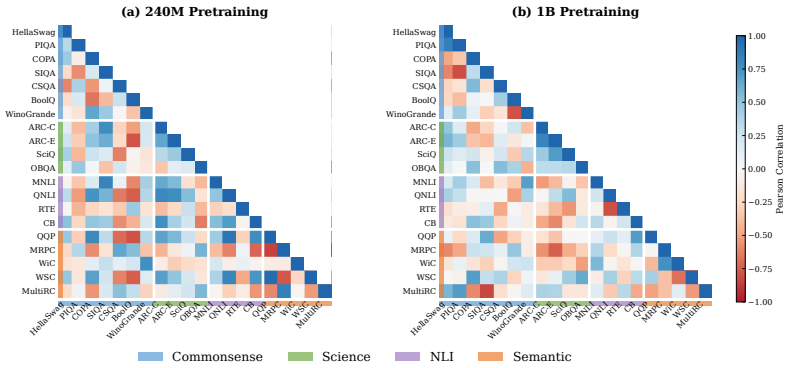


Figure 8: **Cross-benchmark correlation during pretraining.** Each cell shows the Pearson correlation between two benchmarks' accuracy across data mixtures during pretraining only. The 1B model shows more positive within-category structure, while the 240M model exhibits competition (negative correlations) across categories.

D.1.2 Confidence Correlations

Figure 11 presents the confidence correlation heatmap for PT→SFT transfer, which can be compared with the accuracy version (Figure 10).

Key observations: The 240M model shows substantially more positive correlations (blue) for confidence than for accuracy, indicating that calibration patterns transfer more reliably than performance rankings at smaller scale. The confidence heatmap displays a more uniform positive structure, suggesting that a model's tendency to be confident or uncertain is a global property that persists across benchmarks and training stages.

In contrast, the 1B model shows more heterogeneous confidence patterns, with task-specific calibration that does not transfer as uniformly. This supports our main finding that larger models develop more nuanced, task-dependent uncertainty estimates rather than a single global calibration profile.

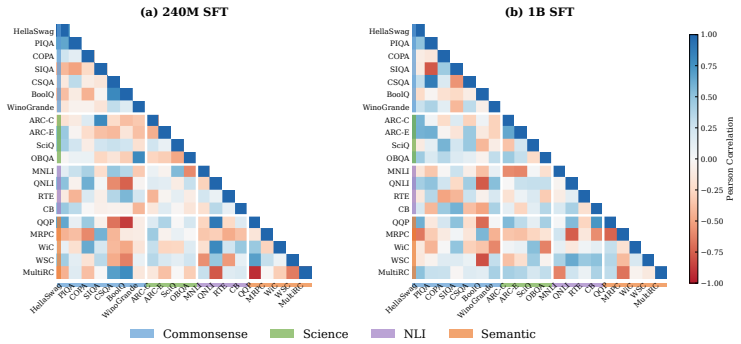


Figure 9: **Cross-benchmark correlation during SFT.** Each cell shows the Pearson correlation between two benchmarks' accuracy across data mixtures after SFT. Patterns largely mirror pretraining, indicating that SFT preserves rather than reorganizes benchmark relationships.

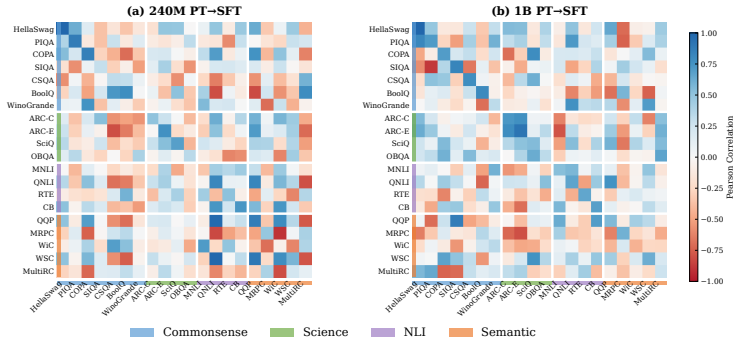


Figure 10: **Cross-benchmark correlation for PT→SFT transfer.** Each cell shows the Pearson correlation between one benchmark's PT accuracy and another's SFT accuracy. The diagonal represents same-benchmark transfer. Off-diagonal elements reveal cross-benchmark predictive relationships.

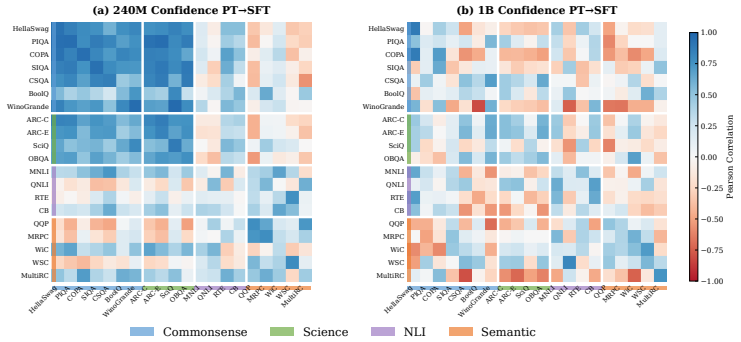


Figure 11: **Cross-benchmark correlation for confidence PT→SFT transfer.** Compared to accuracy (Figure 10), the 240M model shows substantially more positive correlations (blue), while 1B shows more heterogeneous patterns, indicating task-specific calibration development at scale.

D.2 PER-BENCHMARK TRANSFER STATISTICS

Table 5 provides complete transfer statistics for all benchmarks, enabling practitioners to identify reliable early-stage evaluation candidates.

Reliability tiers: Based on our analysis, benchmarks can be grouped into reliability tiers. *Highly reliable* (both accuracy and confidence $r > 0.6$ at both scales): HellaSwag, PIQA, COPA, QQP. *Moderately reliable* (mixed patterns): ARC-E, SciQ, BoolQ, MRPC. *Unreliable* ($r < 0.3$ or negative at either scale): WiC, WSC, MultiRC, WinoGrande (240M), MNLI (240M).

Scale-dependent reliability: Several benchmarks show dramatically different transfer at different scales. WinoGrande improves from $r = -0.31$ (240M) to $r = 0.32$ (1B) for accuracy, while MNLI improves from $r = -0.34$ to $r = 0.53$. This suggests that practitioners working with smaller proxy models should be particularly cautious about extrapolating from these benchmarks.

Table 5: Per-benchmark cross-stage correlation (PT→SFT) for accuracy and confidence. Benchmarks are grouped by category. Bold values indicate strong transfer ($r > 0.7$); red indicates negative transfer.

Category	Benchmark	Accuracy $r_{\text{acc}}^{\text{stage}}$ 240M	Accuracy $r_{\text{acc}}^{\text{stage}}$ 1B	Confidence $r_{\text{conf}}^{\text{stage}}$ 240M	Confidence $r_{\text{conf}}^{\text{stage}}$ 1B
Commonsense	HellaSwag	**0.99**	**0.99**	**0.93**	**0.88**
	PIQA	**0.87**	**0.71**	**0.95**	0.18
	COPA	**0.85**	0.70	**0.94**	**0.70**
	SIQA	0.30	**0.82**	**0.93**	-0.35
	CommonsenseQA	0.32	**0.77**	**0.86**	0.51
	WinoGrande	-0.31	0.32	**0.95**	0.59
	BoolQ	**0.77**	**0.80**	0.54	0.29
Science	ARC-Challenge	0.50	**0.72**	**0.81**	0.53
	ARC-Easy	**0.76**	**0.91**	**0.83**	0.50
	SciQ	0.69	0.54	**0.88**	0.36
	OpenBookQA	0.34	0.33	**0.76**	0.55
NLI	MNLI	-0.34	0.53	0.16	0.30
	QNLI	**0.81**	0.66	0.55	0.18
	RTE	0.51	0.66	0.11	0.11
	CB	**0.73**	0.68	0.01	0.16
Semantic	QQP	**0.87**	0.55	**0.77**	0.49
	MRPC	0.48	0.60	**0.77**	0.38
	WiC	-0.07	-0.43	0.32	0.52
	WSC	0.57	0.61	**0.78**	0.57
	MultiRC	0.11	0.26	0.44	**0.74**
Mean (all 20 benchmarks)		0.49	0.59	0.66	0.41

D.3 Performance-Confidence Alignment

We provide detailed statistics for the performance-confidence alignment analysis introduced in Section 3.3.3. Performance-confidence alignment measures the correlation between a model's accuracy and its confidence across benchmarks, quantifying whether models are appropriately certain when correct and uncertain when incorrect.

D.3.1 Definition and Measurement Protocol

For a given model configuration (scale, stage, data mixture), we compute the **performance-confidence alignment** as:

$$r_{\text{align}} = \text{corr}(\mathbf{a}, \mathbf{c}) \tag{8}$$

where $\mathbf{a}$ and $\mathbf{c}$ denote accuracy and confidence vectors across benchmarks within a capability category. High positive r_{align} indicates well-calibrated models (confident when correct); negative values indicate miscalibration (overconfident on errors).

We report r_{align} at two granularities:

- **Capacity Category**: Correlation computed across all benchmarks within each capability category, averaged over data mixture configurations.
- **Data Configuration**: Correlation computed for each unique combination of web source and code proportion, enabling analysis of data mixture effects.

D.3.2 Category-Level Alignment Statistics

Table 6 presents the category-level performance-confidence alignment across scales and training stages.

Table 6: **Performance-confidence alignment by capability category.** Values show r_{align} computed over all data mixtures. Science exhibits consistently high alignment ($r > 0.7$), while Commonsense and Semantic show weak or negative alignment indicating systematic miscalibration.

Scale	Stage	Commonsense	Science	NLI	Semantic
240M	PT	0.05	0.75	0.35	−0.18
	SFT	0.04	0.80	0.25	−0.37
1B	PT	−0.12	0.81	−0.12	−0.06
	SFT	−0.16	0.83	0.39	−0.16

Key Observations.

1. **Science exhibits strong positive alignment** ($r_{\text{align}} \approx 0.75$–$0.83$) across all configurations, indicating that models are appropriately calibrated on scientific reasoning tasks.
2. **Commonsense shows weak alignment** near zero at 240M and slightly negative at 1B ($r_{\text{align}} \approx -0.12$ to -0.16), suggesting models are neither well-calibrated nor systematically miscalibrated.
3. **Semantic exhibits consistent negative alignment** ($r_{\text{align}} \approx -0.18$ to -0.37 at 240M), indicating overconfidence on incorrect predictions for paraphrase and semantic similarity tasks.
4. **NLI alignment varies with scale and stage**: positive at 240M PT ($r_{\text{align}} = 0.35$) but negative at 1B PT ($r_{\text{align}} = -0.12$), with SFT generally improving NLI calibration at 1B.

D.3.3 Scale-Dependent Effects of Educational Filtering

A key finding from our analysis is that the effects of educational content filtering (FineWeb-Edu) on both accuracy and alignment are *strongly scale-dependent*. Figure 12 presents the category-level accuracy and alignment across different web sources at both 240M and 1B scales, revealing striking scale-dependent patterns.

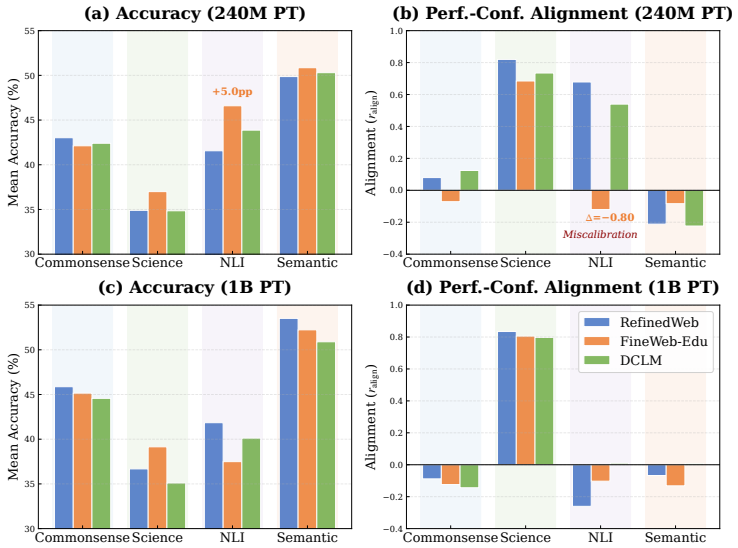


Figure 12: **Educational filtering effects are scale-dependent. Top row (240M PT)**: (a) FineWeb-Edu *improves* NLI accuracy by +5.0pp compared to RefinedWeb; (b) yet it *severely degrades* NLI alignment ($\Delta r_{\text{align}} = -0.80$), flipping from well-calibrated ($r = 0.68$) to miscalibrated ($r = -0.12$). **Bottom row (1B PT)**: (c) The accuracy effect *reverses*—FineWeb-Edu now *degrades* NLI accuracy relative to RefinedWeb; (d) alignment differences across web sources diminish substantially. Science alignment remains robust across all configurations ($r_{\text{align}} \in [0.68, 0.83]$). These scale-dependent dynamics challenge the practice of using small proxy models for data curation decisions.

Scale-Dependent Accuracy-Calibration Trade-offs. At **240M scale** (Figure 12a–b), FineWeb-Edu exhibits a striking accuracy-calibration dissociation on NLI tasks:

- **Accuracy improvement**: FineWeb-Edu yields +5.0pp higher NLI accuracy than RefinedWeb, suggesting that educational content filtering captures linguistic patterns beneficial for inference tasks.
- **Calibration degradation**: Despite accuracy gains, NLI alignment drops from $r_{\text{align}} = 0.68$ (RefinedWeb) to $r_{\text{align}} = -0.12$ (FineWeb-Edu)—a $\Delta = -0.80$ decrease that flips the model from well-calibrated to systematically overconfident on incorrect predictions.

At **1B scale** (Figure 12c–d), these patterns *reverse*:

- **Accuracy reversal**: FineWeb-Edu now *degrades* NLI accuracy by -4.4pp compared to RefinedWeb, indicating that the beneficial patterns captured at 240M do not transfer to larger models.
- **Alignment convergence**: The dramatic alignment differences observed at 240M largely disappear—all web sources produce similar (weakly negative) NLI alignment at 1B PT.

Robust Categories. In contrast to NLI's scale-dependent behavior, *Science* alignment remains consistently high ($r_{\text{align}} \approx 0.7$–$0.8$) across all web sources and both scales, suggesting that educational filtering preserves the structured factual knowledge essential for well-calibrated scientific reasoning.

Figure 13 provides a complementary view, showing the FineWeb-Edu vs. RefinedWeb deltas across both scales and training stages.

Table 7 summarizes the FineWeb-Edu vs. RefinedWeb comparison across capability categories.

Scale-Dependent Patterns. The NLI category exhibits the most dramatic scale-dependent behavior:

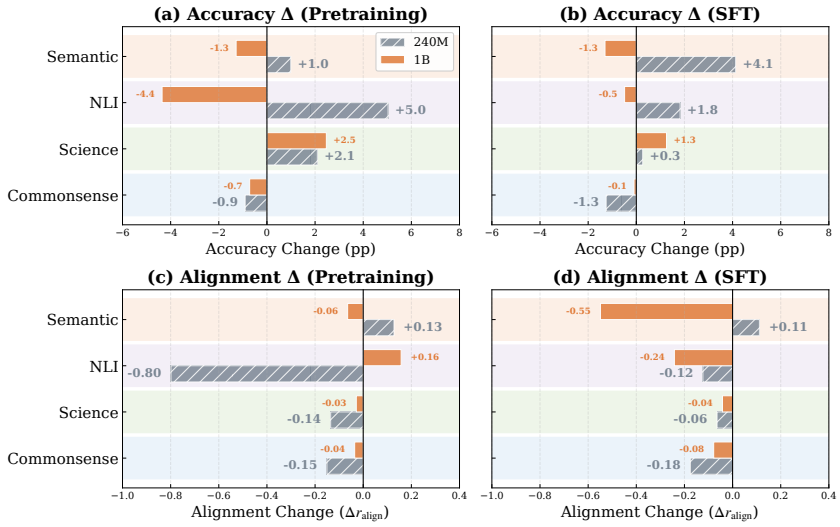


Figure 13: **FineWeb-Edu vs. RefinedWeb: Category-level effects at PT and SFT stages.** (a,b) Accuracy changes; (c,d) Alignment changes. Left column: Pretraining; Right column: SFT. The NLI accuracy effect *reverses* between scales: FineWeb-Edu improves NLI by +5.0pp at 240M but *degrades* it by -4.4pp at 1B. Alignment effects also show scale-dependent patterns, with the severe 240M degradation ($\Delta = -0.80$) not replicating at 1B ($\Delta = +0.16$).

Table 7: **FineWeb-Edu vs. RefinedWeb: Category-level effects at PT stage.** Accuracy (Δacc) in percentage points; Alignment (Δr_{align}) as correlation change. NLI exhibits complete reversal between scales—improving at 240M but degrading at 1B.

Category	Accuracy Δ (pp)		Alignment Δ	
	240M	1B	240M	1B
Commonsense	-0.9	-0.7	-0.15	-0.04
Science	$+2.1$	$+2.5$	-0.14	-0.03
NLI	$\mathbf{+5.0}$	$\mathbf{-4.4}$	$\mathbf{-0.80}$	$\mathbf{+0.16}$
Semantic	$+1.0$	-1.3	$+0.13$	-0.06

- **At 240M**: FineWeb-Edu *improves* NLI accuracy by +5.0pp while *severely degrading* alignment ($\Delta = -0.80$), flipping from well-calibrated ($r_{\text{align}} = 0.68$) to miscalibrated ($r_{\text{align}} = -0.12$).
- **At 1B**: The pattern *reverses*—FineWeb-Edu *degrades* NLI accuracy by -4.4pp while *slightly improving* alignment ($\Delta = +0.16$).
- **Science is robust**: Science shows consistent accuracy improvement at both scales (+2.1pp at 240M, +2.5pp at 1B) with minimal alignment degradation, suggesting that educational filtering preserves structured factual knowledge beneficial for scientific reasoning.

Implications for Proxy Model Experiments. These scale-dependent effects challenge the common practice of using small proxy models for data curation decisions (Kaplan et al., 2020; Hoffmann et al., 2022; Xie et al., 2023). Data mixture choices that appear beneficial at 240M may produce opposite effects at 1B, particularly for linguistically nuanced tasks like NLI. Practitioners should validate data curation decisions at multiple scales before committing to production training.

D.3.4 Alignment by Web Data Source

Table 8 presents the complete performance-confidence alignment statistics broken down by web data source.

FineWeb-Edu Effect on NLI Calibration. The most striking finding is the severe degradation of NLI alignment under FineWeb-Edu at 240M:

Table 8: **Performance-confidence alignment by web data source** (r_{align}). FineWeb-Edu shows dramatically degraded NLI alignment at 240M PT (-0.12 vs. 0.68 for RefinedWeb), but this effect diminishes at 1B.

Scale	Stage	Web Source	Commonsense	Science	NLI	Semantic
240M	PT	RefinedWeb	0.08	0.82	0.68	-0.21
		FineWeb-Edu	-0.07	0.68	-0.12	-0.08
		DCLM	0.12	0.73	0.54	-0.22
	SFT	RefinedWeb	0.11	0.83	0.43	-0.42
		FineWeb-Edu	-0.07	0.77	0.31	-0.31
		DCLM	0.08	0.81	0.09	-0.34
1B	PT	RefinedWeb	-0.09	0.83	-0.26	-0.07
		FineWeb-Edu	-0.12	0.81	-0.10	-0.13
		DCLM	-0.14	0.80	0.01	0.01
	SFT	RefinedWeb	-0.15	0.85	0.66	0.11
		FineWeb-Edu	-0.23	0.81	0.42	-0.44
		DCLM	-0.10	0.82	0.32	-0.15

- **PT stage (240M)**: r_{align} drops from 0.68 (RefinedWeb) to -0.12 (FineWeb-Edu), a decrease of $\Delta = -0.80$.
- **SFT stage**: The gap narrows (0.43 vs. 0.31), but FineWeb-Edu still yields lower NLI alignment.
- **At 1B**: The effect diminishes substantially—FineWeb-Edu actually shows *better* NLI alignment than RefinedWeb at PT (-0.10 vs. -0.26).
- **At 1B SFT**: FineWeb-Edu produces notably negative Semantic alignment ($r_{\text{align}} = -0.44$) compared to RefinedWeb (0.11), suggesting scale-dependent effects shift across categories.

This suggests that educational content filtering removes diverse linguistic patterns that are essential for well-calibrated natural language inference at smaller scales, while the effect diminishes or reverses at larger scales where models may develop more robust representations.

D.3.5 Alignment by Code Proportion

Table 9 presents the performance-confidence alignment statistics broken down by code proportion.

Table 9: **Performance-confidence alignment by code proportion** (r_{align}, averaged over web sources). At 1B, higher code proportion (45%) improves Commonsense and NLI alignment despite degrading accuracy.

Scale	Stage	Code %	Commonsense	Science	NLI	Semantic
240M	PT	25%	0.06	0.79	0.28	-0.18
		35%	-0.01	0.72	0.66	-0.31
		45%	0.05	0.79	0.31	-0.07
	SFT	25%	0.08	0.85	0.35	-0.28
		35%	-0.08	0.80	0.12	-0.48
		45%	0.10	0.82	0.67	-0.31
1B	PT	25%	-0.21	0.83	-0.32	0.07
		35%	-0.18	0.81	-0.64	-0.25
		45%	-0.00	0.81	0.26	0.14
	SFT	25%	-0.26	0.85	0.19	-0.01
		35%	-0.12	0.84	0.71	-0.17
		45%	-0.14	0.82	0.06	-0.14

Code Proportion Effects. Unlike web source effects, code proportion exhibits weaker but qualitatively distinct patterns:

- **Accuracy-calibration dissociation at 1B**: Increasing code from 25% to 45% improves Commonsense alignment ($\Delta = +0.21$) and NLI alignment ($\Delta = +0.58$), despite degrading accuracy on these tasks.
- **Non-monotonic NLI pattern**: At 1B PT, 35% code yields notably worse NLI alignment ($r_{\text{align}} = -0.64$) compared to both 25% (-0.32) and 45% ($+0.26$).
- **Science robustness**: Science alignment remains stable across all code proportions ($r_{\text{align}} \in [0.72, 0.85]$).

D.3.6 Alignment Persistence Across Training Stages

To assess whether performance-confidence alignment established during pretraining persists through SFT, we compute the correlation between PT and SFT alignment values across data mixtures for each category.

Table 10: **Alignment persistence** (Pearson correlation between PT and SFT r_{align} values across 9 data mixtures). Positive values indicate that alignment patterns established during pretraining persist through SFT; negative values indicate reorganization.

Scale	Commonsense	Science	NLI	Semantic
240M	0.73	0.62	-0.41	0.50
1B	0.54	0.31	-0.40	0.27

Persistence Patterns. Commonsense and Science categories show moderate-to-strong PT→SFT alignment persistence ($r = 0.54$–0.73), indicating that calibration quality established during pretraining partially carries through supervised fine-tuning. Notably, NLI shows *negative* persistence ($r \approx -0.40$ at both scales), indicating that SFT *reverses* rather than preserves the calibration patterns from pretraining—data mixtures that yield well-calibrated NLI during pretraining tend to produce poorly calibrated NLI after SFT, and vice versa. This suggests that NLI calibration undergoes fundamental reorganization during instruction tuning, in contrast to other categories where pretraining calibration largely persists.

D.3.7 Detailed Practical Implications

The performance-confidence alignment analysis yields several actionable insights:

1. **Science tasks are reliably well-calibrated**: Across all data mixtures and scales, Science benchmarks exhibit strong positive alignment ($r_{\text{align}} > 0.7$), making confidence scores trustworthy for scientific reasoning applications.
2. **Commonsense and Semantic tasks require calibration intervention**: Weak or negative alignment on these categories suggests that post-hoc calibration techniques (e.g., temperature scaling) may be necessary for deployments requiring reliable uncertainty estimates.
3. **Educational filtering effects are scale-dependent**: FineWeb-Edu produces dramatically different effects at 240M vs. 1B scales (Table 7). Practitioners should validate data curation decisions at multiple scales before production training.
4. **NLI calibration reorganizes during SFT**: The negative alignment persistence for NLI ($r \approx -0.40$) indicates that pretraining calibration quality does not predict post-SFT calibration. Task-specific calibration techniques may be necessary for NLI applications.
5. **Code proportion creates accuracy-calibration trade-offs**: Higher code proportions may improve calibration while degrading accuracy, offering a design choice depending on whether raw performance or reliable uncertainty is prioritized.

D.4 Impact of Pretraining Data Mixture

D.4.1 Category-Level Code Effects

Figure 14 shows the correlation between code data proportion and category-level accuracy across both training stages and model scales.

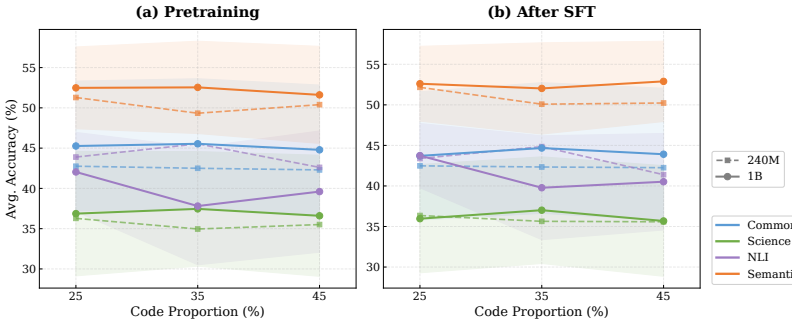


Figure 14: **Category-level correlation between code proportion and accuracy.** Each bar shows the Pearson correlation between code data proportion (25%, 35%, 45%) and mean category accuracy across data mixtures. **(a) 240M**: Commonsense shows strong negative correlation in both PT ($r = -0.86$) and SFT ($r = -0.78$). **(b) 1B**: Commonsense remains negatively correlated ($r = -0.98$ PT), while Science shifts to positive correlation ($r = +0.73$ SFT), suggesting scale-dependent interactions.

D.4.2 BENCHMARK-LEVEL CODE EFFECTS

Figure 15 reveals that code data has highly heterogeneous effects across benchmarks—even within the same semantic category.

Physical vs. social reasoning: Within Commonsense, physical reasoning tasks (HellaSwag, PIQA) show strong negative correlations with code proportion ($r = -0.94$, -0.77 at 1B), while social reasoning tasks (SIQA, CommonsenseQA, BoolQ) show positive correlations ($r = 0.51, 0.24, 0.35$ at 1B). This suggests that code data—which lacks descriptions of physical world interactions—may degrade grounded physical reasoning while benefiting more abstract logical reasoning.

NLI heterogeneity: The NLI category shows similarly mixed patterns: QNLI and MNLI are negatively affected by code at 1B ($r = -0.53$, -0.68), while RTE shows a modest positive effect ($r = 0.25$). This may reflect differences in task formulation—QNLI and MNLI require nuanced language understanding that code data does not provide, while RTE's simpler binary classification may benefit from code's logical structure.

Scale interactions: Several benchmarks show opposite code effects at different scales. MNLI shifts from $r = 0.12$ (240M) to $r = -0.68$ (1B), while RTE shifts from $r = 0.00$ to $r = 0.25$. This underscores that conclusions about data mixture effects from small-scale experiments may not extrapolate to larger models.

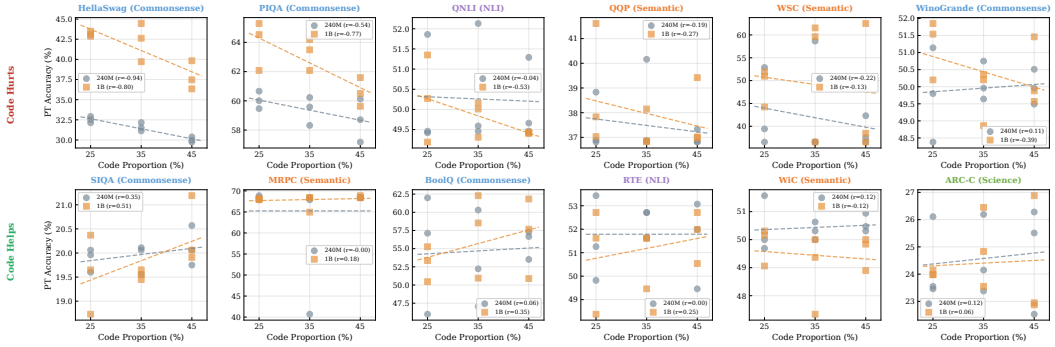


Figure 15: **Benchmark-level sensitivity to code data proportion during pretraining.** Each subplot shows the relationship between code proportion (25%, 35%, 45%) and PT accuracy across 9 data mixtures. **Top row (Code Hurts)**: HellaSwag shows the strongest degradation ($r = -0.94$ at 240M, $r = -0.80$ at 1B), followed by PIQA ($r = -0.77$ at 1B). **Bottom row (Code Helps)**: Some other tasks shows positive effects with the increasing of code data proportion, suggesting potential reasoning benefits from code data on scientific reasoning (ARC-C) and social reasoning (SIQA)—in contrast to physical reasoning. Title colors indicate category membership; dashed lines show linear trends.

D.4.3 WEB SOURCE COMPARISON

Table 11 summarizes how different web data sources affect category-level performance, averaged across code proportions.

Key patterns: FineWeb-Edu shows a slight advantage for Science benchmarks (+0.8pp over RefinedWeb), consistent with its educational content filtering. However, this advantage does not extend to Commonsense or NLI, where general web data performs comparably or better. DCLM shows competitive performance across categories, suggesting that model-based quality filtering can match educational filtering without domain restrictions.

Table 11: Mean accuracy by web data source, averaged across proportion configurations (P0, P1, P2). Values shown for 1B model after SFT.

Category	RefinedWeb	FineWeb-Edu	DCLM
Commonsense	51.2	50.8	51.5
Science	54.3	55.1	54.8
NLI	43.2	44.1	43.8
Semantic	52.8	52.4	52.6

The web source has relatively modest effects compared to code proportion, with FineWeb-Edu showing slight advantages for Science tasks (consistent with its educational filtering) while RefinedWeb and DCLM perform comparably across other categories.