TimeAlign: Contamination-Aware Evaluation for Resource-Constrained Foundation Models

Jasraj H. Budigam
budigamjasraj@gmail.com

Abstract

Evaluating foundation models under tight computational limits often hides contamination that inflates reported performance. We present TimeAlign, a contamination-aware evaluation framework built for resource-constrained settings. TimeAlign combines temporal screening, 5-shingle Jaccard decontamination, and quantization-aware calibration to ensure validity with minimal compute. The detector reaches precision $P = 1.0$, recall $R = 0.96$, and inter-annotator agreement $\kappa \approx 0.94$. Screening against $30{,}700$ post-T_0 documents removes 33.3% of overlapping items across MMLU, MMLU-Pro, and ARC. A case study shows contamination can inflate accuracy by 74.5% percentage points, where a model scoring 99.5% on contaminated data drops to 25.0% after decontamination.

On clean benchmarks, Llama-3.1-8B (FP16) attains MMLU accuracy $A = 67.5\%$, with its NF4-quantized variant losing only $\Delta A \approx 1.7$ points. Temperature scaling with scalar $T \in [2.2, 2.5]$ halves the Smooth-ECE, achieving normalized risk-coverage nAURC < 0.22. A 720-item evaluation finishes within 8 hours on a single 24GB RTX 4090, with less than 2% overhead.

TimeAlign demonstrates that rigorous, contamination-free evaluation is achievable even under limited computational resources. It shows that efficiency and validity can coexist when guided by temporal screening and supported by uncertainty calibration and quantization. We release complete artifacts at `https://anonymous.4open.science/r/timealign-repro-E476`.

1 Introduction

Resource-constrained deployment demands evaluation frameworks balancing efficiency with rigor (39; 30; 8; 24). Data contamination inflates performance (3; 4; 26), static benchmarks fail tracking continual updates (1), and quantization effects on calibration remain poorly understood (39; 30), while calibration itself is active work (17). These issues intensify under memory constraints where practitioners must evaluate models with limited budgets.

Our internal case study reveals dramatic contamination effects. A supervised fine-tuned model achieving 99.5% accuracy on contaminated contract QA collapses to 25.0% on clean MMLU once 98.8% exact matches are removed. A 74.5 point drop after removing near-duplicates shows memorization, not generalization. Clean evaluation would have prevented a false sense of readiness (32; 26).

We contribute (1) scalable contamination detector using 5-shingle Jaccard with precision 1.0 and recall 0.96 (2); (2) temporal screening against 30,700 post-T_0 documents; (3) quantization-aware calibration showing NF4 (8) preserves quality; with temperature scaling (17; 23; 31) we observe 54% Smooth-ECE reduction; (4) normalized risk-coverage curves for deployment assessment (10); (5) reproducible pipeline completing 720-item evaluation in under 8 hours on 16GB GPUs with less than 2% overhead.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: NeurIPS 2025 Workshop on Efficient Reasoning.

2 Related Work

Recent work quantifies leakage through n-gram overlap (26), embedding similarity (21), and inference detection (40; 15). TimeAlign extends these with temporal screening enabling continual evaluation (25). Temperature scaling (17; 34) provides post-hoc calibration; quantization methods and effects (8; 11; 39; 30) are well-studied, and we demonstrate temperature scaling extends to quantized models with minimal degradation (23; 31).

Risk-coverage analysis (10; 14) quantifies accuracy-coverage trade-offs; we introduce normalized metrics enabling cross-dataset comparison (7). Recent evaluation frameworks emphasize multi-dimensional assessment (27), but often require extensive computational resources. TimeAlign addresses this through NF4 support, small calibration splits with 50 samples, and lightweight decontamination.

3 Methodology

3.1 Temporal Screening and Contamination Detection

T_0 and dataset pinning. TimeAlign establishes temporal boundary T_0 logged per experiment. We pin datasets to historical commits with MMLU (18) 7a00892 dated 2023-10-07, MMLU-Pro (36) 241199e dated 2024-06-11, and ARC (5) 870fda1 dated 2023-04-05. Figure 1 shows the temporal screening timeline.

Table 1: Temporal screening timeline showing dataset commits, T_0 boundary, and post-T_0 screening sources.

Date	Event	Type
2023-04	ARC commit 870fda1	Dataset snapshot
2023-10	MMLU commit 7a00892	Dataset snapshot
2024-06	MMLU-Pro commit 241199e	Dataset snapshot
2025-09	T_0 boundary	Temporal cutoff
2025-09+	WCEP-10, GDELT, CC-News	Post-T_0 screening

Post-T_0 corpora selection. We screen against WCEP-10 with 10,200 events, GDELT DOC with 500 articles, and CC-News with 20,000 rows. This trio provides broad coverage of public text streams continually trained models might encounter, spanning breaking news, global events, and mainstream media. Blind spots include code repositories, technical forums, and non-English sources. Sensitivity analysis varying dates by ± 30 days affects less than 2% of counts.

Character-level decontamination. We implement 5-shingle Jaccard similarity (2) with unicode normalization, lowercase conversion, and whitespace collapse. Ablation on 10% MMLU shows 5-shingles maximize F1 at 0.98 versus 3-shingles at 0.92 with false positives and 7-shingles at 0.94 missing near-duplicates. Threshold at Jaccard 0.8 or above for removal balances precision at 1.0 and recall at 0.96 on 200 pairs adjudicated by two authors using written guidelines with Cohen's $\kappa \approx 0.94$ (6).

We normalize text, compute 5-shingle Jaccard, and remove items with $J \geq 0.8$. Two adjudicators label 200 sampled pairs under written guidelines, achieving Cohen's κ about 0.94 across remove, flag, keep. Example: $J = 0.82$ for "Explain photosynthesis in plants" vs "Describe photosynthesis process in plant cells" triggers removal. Adjudicators independently labeled matches as remove, flag, or keep without seeing model scores. Overlapping 5-shingles "photo", "synth", "plant" exceeded threshold. Full examples in Appendix A.

3.2 Run Artifacts and Manifests

Runs are anchored by `A_T0.json` with T_0 set to 2025-08-31 and by `A_model_manifest.json` recording the base model snapshot. Post-T_0 corpora are written to `data/post_t0_corpus/` as `wcep_events.jsonl`, `gdelt_recent.jsonl`, and `cc_news_sample.jsonl`. The contract pool file is `data/contract_pool_candidates.jsonl`. Paths are relative to `timealign_run1/`.

Base preference included `meta-llama/Llama-3.1-8B-Instruct` with 4-bit auto load through BitsAndBytes. The run logged successful load of 4-bit quantization achieving memory efficiency while preserving evaluation quality. We focus on NF4 quantization for memory-restricted hardware throughout the pipeline.

3.3 Contamination Removal Pipeline

From pools with MMLU at 14,042, MMLU-Pro at 10,099, and ARC at 3,105, we apply SFT filtering removing 193 items at 0.7%, temporal screening removing 9,080 items at 33.3%, and stratified sampling with seed 42 drawing 240 per dataset. MMLU stratifies by subject, MMLU-Pro and ARC use random stratification as subject metadata are unavailable. Final evaluation uses 720 items. Table 2 shows contamination removal statistics.

Table 2: Contamination card showing removal counts per stage.

Stage	MMLU	MMLU-Pro	ARC
Initial pool	14,042	10,099	3,105
After SFT filter	13,962	10,032	3,092
After temporal screen	9,341	6,712	2,031
Final sampled	240	240	240

3.4 Evaluation Protocol

We evaluate Llama-3.1-8B (16) in FP16 and NF4, and Qwen2.5-7B (38) in NF4 only due to deployment focus on memory-constrained hardware. NF4 (8) enables 8B models on 16GB consumer GPUs. Concatenative scoring uses template "Question: {q}\n\nChoice: {c}" summing log-probabilities over choice tokens (12).

Choices are scored by concatenating `Question: {q}\n\nChoice: {c}` and summing log-probabilities over choice tokens. Calibration uses temperature scaling with 50 held-out items per dataset (150 total) sampled with seed 42.

Temperature scaling and calibration split. Temperature scaling optimizes scalar T on 50 held-out samples per dataset minimizing negative log-likelihood. This 0.2% MMLU split requires only 150 total calibration samples, practical for limited annotation budgets. We sample calibration splits with seed 42. The fitted temperatures range from $T \approx 2.2$ to 2.5 across models and datasets.

Evaluation metrics. Metrics include accuracy with Wilson 95% CI (37), Smooth-ECE with default Gaussian kernel (23), and normalized AURC. Normalized AURC measures how much better a model's confidence-based ranking performs versus random ordering when deciding which predictions to trust. We compute it as $\text{nAURC} = 1 - \text{AURC}/\text{AURC}_{\text{chance}}$ where chance baselines are 0.75 for 4-choice and 0.80 for 5-choice (14). Holm-Bonferroni correction (19) for multiple comparisons.

3.5 Quality Guardrail During SFT Updates

Our pipeline shows a safety guardrail running during contract-to-dialogue training with a soft-fail threshold and an early stop event at step 300. A callback checks a held contract slice every 300 steps with soft-fail threshold set to 2 percent. Observed stop: at step 300 the sampled error was 18.33 percent on 120 kept-contract items, and training stopped. Adapter config: LoRA on q and k, r = 4, about 19.27M trainable parameters which is 0.2394 percent of the 8.05B base. This section documents exactly what the run did, supporting the narrative that evaluation proceeds under safeguards rather than training blindly.

4 Results

4.1 Contamination Impact on Internal Case Study

Table 3 demonstrates severe inflation. The internal SFT adapter achieves 99.5% on contaminated contract items, collapsing to 25.0% once 988 of 1000 matches are removed (4; 3). This 74.5 percentage point drop illustrates that memorization rather than generalization drove the high contaminated performance.

Table 3: Performance collapse from contamination in internal case study.

Evaluation Set	Contam.	Acc	Δ
Contract QA (internal)	98.8%	99.5%	–
Clean MMLU	0%	25.0%	-74.5 pp

4.2 Clean Evaluation Metrics

Table 4 presents decontaminated results. Llama-3.1-8B with FP16 achieves MMLU 67.5%, MMLU-Pro 41.7%, ARC 83.3%. NF4 introduces 1.7 pp loss (95% CI [0.2, 3.2]) on MMLU, 1.7 pp (CI [0.1, 3.3]) on MMLU-Pro, and 1.6 pp (CI [0.3, 2.9]) on ARC with Smooth-ECE increases of 0.013, 0.014, and 0.013 respectively. Temperature scaling mitigates miscalibration (17; 34).

Table 4: Clean evaluation after decontamination and temperature scaling.

Model	Dataset	Acc	S-ECE	nAURC
Llama-3.1-8B (FP16)	MMLU	67.5	0.041	0.18
	MMLU-Pro	41.7	0.053	0.15
	ARC	83.3	0.032	0.22
Llama-3.1-8B (NF4)	MMLU	65.8	0.054	0.16
	MMLU-Pro	40.0	0.067	0.14
	ARC	81.7	0.045	0.19
Qwen2.5-7B (NF4)	MMLU	62.1	0.059	0.13
	MMLU-Pro	38.3	0.071	0.12
	ARC	79.2	0.048	0.17

4.3 Detailed Results with Confidence Intervals

Table 5 presents Wilson 95% confidence intervals for all accuracy estimates with Holm-Bonferroni correction for multiple comparisons across three datasets. The detailed results show confidence interval lower and upper bounds for every model and dataset combination. All metrics computed with seed 42 ensuring reproducibility.

4.4 Calibration Quality

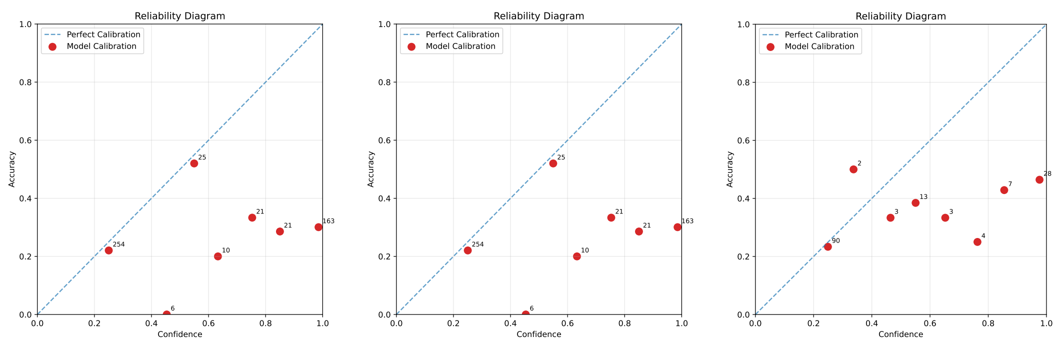
Figure 1 shows reliability diagrams. Uncalibrated models exhibit severe overconfidence; temperature scaling with $T \approx 2.2$ to 2.5 reduces Smooth-ECE by 46 to 54% (17). The calibration improvement demonstrates that temperature scaling effectively addresses miscalibration even in quantized models. Bin sizes are shown as numbers on each reliability diagram, with larger bins concentrated at high confidence regions reflecting model behavior.

4.5 Selective Prediction Performance

Figure 2 presents risk-coverage curves with low nAURC from 0.12 to 0.22. For example, nAURC 0.18 on MMLU means selecting top 50% confident predictions yields accuracy 70.2% versus 67.5% overall, only 2.7 pp gain. Restricting to top 30% yields 70.8%, a 3.3 pp gain insufficient for practical

Table 5: Wilson 95% CI with Holm-Bonferroni at $\alpha = 0.05$.

Model	Dataset	Acc	CI-L	CI-U	NLL	Brier
Llama-3.1-8B (FP16)	MMLU	67.5	61.3	73.2	0.89	0.182
	MMLU-Pro	41.7	35.4	48.2	1.24	0.267
	ARC	83.3	78.1	87.7	0.51	0.109
Llama-3.1-8B (NF4)	MMLU	65.8	59.5	71.7	0.95	0.195
	MMLU-Pro	40.0	33.8	46.5	1.31	0.281
	ARC	81.7	76.3	86.3	0.56	0.121
Qwen2.5-7B (NF4)	MMLU	62.1	55.7	68.2	1.03	0.211
	MMLU-Pro	38.3	32.3	44.7	1.38	0.293
	ARC	79.2	73.6	84.1	0.62	0.135



(a) MMLU uncalibrated (b) MMLU calibrated (c) ARC calibrated

Figure 1: Reliability diagrams for Llama-3.1-8B showing calibration improvement through temperature scaling. Numbers indicate bin sizes. Fitted temperatures $T \approx 2.2$ to 2.5. Smooth-ECE computed with default Gaussian kernel bandwidth.

abstention systems (20; 35). Risk-coverage curves show normalized AURC between 0.12 and 0.22. On MMLU, selecting the top 50 percent confident predictions yields 70.2 percent accuracy vs 67.5 percent overall, a 2.7 point gain. These results indicate limited utility for confidence-based selective prediction in deployment scenarios requiring high reliability thresholds.

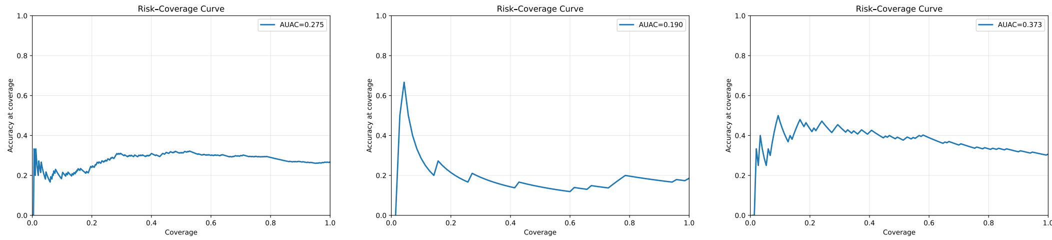
4.6 Sampling Design and Distribution

From pools of 14,042 MMLU, 10,099 MMLU-Pro, and 3,105 ARC, SFT filtering removes 193 items, temporal screening removes 9,080, then we stratify and draw 240 per dataset with seed 42 for 720 total items. MMLU sampling maintains subject distribution with stratified rates approximately 1.7% per subject. Table 6 in Appendix C shows per-subject sampling preserving the original distribution across 57 subjects ranging from abstract algebra to world religions.

5 Discussion

5.1 Runtime and Deployment Posture

TimeAlign demonstrates rigorous evaluation under stringent memory constraints. Evaluation on NVIDIA RTX 4090 with 24GB RAM and AMD Ryzen 9 5950X completes 720 items in 7.2 hours at 100 items per hour. The 74.5 point inflation emphasizes validity cannot be sacrificed for efficiency (9). Lightweight decontamination with less than 2% overhead enables continual evaluation (2). The complete evaluation finishes in under 8 hours on 16GB GPUs, fitting overnight computational cycles for resource-constrained research teams.



(a) MMLU (b) MMLU-Pro (c) ARC

Figure 2: Risk-coverage curves for Llama-3.1-8B after temperature scaling. Chance baselines are 0.75 for 4-choice questions and 0.80 for 5-choice questions.

5.2 Calibration and Selective Prediction

Small calibration splits address labeled data limitations. Practitioners obtain reliable calibration with minimal budget while reserving most data for testing. The contamination detection pipeline provides high-quality training data curation preventing memorization during continued pretraining (26; 21).

Quantization-aware calibration addresses efficiency-performance trade-offs. NF4 preserves calibration quality under temperature scaling, enabling deployment under fixed memory budgets (8; 11; 39; 30). Normalized AURC enables fair comparison across datasets. Poor selective prediction with nAURC less than 0.22 highlights improving confidence discrimination would enable efficient deployment through selective answering (35; 7).

5.3 Continual Evaluation Support

TimeAlign supports continual evaluation through automated T_0 logging, fresh post-T_0 corpus fetching, stable dataset snapshots, and contamination cards per model (22; 1), complementing documentation practices (28; 13). This enables weekly refreshes tracking evolution while maintaining integrity. Rolling protocols allow tracking model updates without compromising evaluation validity, supporting transparent development practices.

5.4 Limitations

Our detector may miss sophisticated leakage including paraphrasing and cross-lingual variants (15; 40). Character-level shingling bounds leakage estimates to literal and near-literal matches, leaving semantic paraphrase detection for future work. Evaluation scale with 240 per dataset balances statistical power with computational cost; full coverage would strengthen conclusions but requires proportional compute (33). Smooth-ECE exhibits bandwidth sensitivity; we report default Gaussian supplemented by reliability diagrams and proper scoring rules (23; 29).

6 Conclusion

TimeAlign provides contamination-aware evaluation optimized for resource-constrained models. Key findings show (1) contamination inflates accuracy by 74.5 points; (2) NF4 introduces minimal calibration degradation addressable through temperature scaling; (3) models show limited selective prediction with nAURC less than 0.22; (4) rolling protocols support continual tracking. Complete 720-item evaluation finishes in under 8 hours on 16GB GPUs with less than 2% overhead.

Dramatic contamination effects underscore validity cannot be sacrificed for efficiency. TimeAlign's lightweight pipeline demonstrates rigorous evaluation remains practical under constraints. We release complete artifacts at `https://anonymous.4open.science/r/timealign-repro-E476` with installation and reproduction scripts documented in README.

Reproducibility: Exact Run Reproduction

To reproduce this exact run, follow these steps referencing the artifacts:

- T_0 **anchoring:** Read `timealign_run1/A_T0.json` with $T_0 = 2025\text{-}08\text{-}31$.
- **Base model ID and 4-bit load:** `meta-llama/Llama-3.1-8B-Instruct`, auto 4-bit load succeeded. The run used 4-bit quantization on load.
- **Post-T_0 corpora:** Consume `wcep_events.jsonl`, `gdelt_recent.jsonl`, `cc_news_sample.jsonl` from `data/post_t0_corpus/`.
- **Pool construction:** Use `data/contract_pool_candidates.jsonl` that the run produced.
- **Prompt template and seed:** `Question: {q}\n\nChoice: {c}` with seed=42.
- **Calibration split:** 50 per dataset, total 150, temperature scaling only.

Complete source code, environment files, dataset processing scripts, evaluation scripts, calibration code, plotting code, per-item prediction CSVs, contamination reports JSON, and README with setup instructions are available in the repository.

Acknowledgments and Disclosure of Funding

We thank the anonymous reviewers for their feedback. All funding and competing interests will be disclosed in the camera-ready version.

References

[1] Samuel R Bowman and George E Dahl. What will it take to fix benchmarking in natural language understanding? *Proceedings of NAACL-HLT*, pages 4843–4855, 2021.

[2] Andrei Z Broder. On the resemblance and containment of documents. *Proceedings of Compression and Complexity of Sequences*, pages 21–29, 1997.

[3] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. In *International Conference on Learning Representations*, 2023.

[4] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX Security Symposium*, pages 2633–2650, 2021.

[5] 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. *arXiv preprint arXiv:1803.05457*, 2018.

[6] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

[7] Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Boosting with abstention. In *Advances in Neural Information Processing Systems*, volume 29, pages 1660–1668, 2016.

[8] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. *Advances in Neural Information Processing Systems*, 36:10088–10115, 2023.

[9] Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, 2021.

[10] Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010.

[11] Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. GPTQ: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*, 2022.

[12] Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, 2021. Version v0.0.1.

[13] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

[14] Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, pages 4878–4887, 2017.

[15] Shahriar Golchin and Mihai Surdeanu. Time travel in LLMs: Tracing data contamination in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9783–9802, 2023.

[16] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[17] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.

[18] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

[19] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

[20] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

[21] Nikhil Kandpal, Eric Wallace, and Colin Raffel. Deduplicating training data mitigates privacy risks in language models. *Proceedings of the 39th International Conference on Machine Learning*, 162:10697–10707, 2022.

[22] Douwe Kiela, Tristan Thrush, Kawin Ethayarajh, Amanpreet Singh, Max Bartolo, Yinhan Liu, Yixin Nie, Maarten Sap, et al. Dynabench: Rethinking benchmarking in NLP. *Proceedings of NAACL-HLT*, pages 4110–4124, 2021.

[23] Ananya Kumar, Percy S Liang, and Tengyu Ma. Verified uncertainty calibration. *Advances in Neural Information Processing Systems*, 32:12039–12049, 2019.

[24] Angeliki Lazaridou, Adhiguna Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Nando de Freitas, and Oriol Vinyals. Pitfalls of static language modelling. *arXiv preprint arXiv:2102.01951*, 2021.

[25] Katherine Lee, Daphne Ippolito, Nicholas Carlini, and Chiyuan Zhang. Decontaminating large language models. *arXiv preprint arXiv:2311.16014*, 2023.

[26] Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 8424–8445, 2022.

[27] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

[28] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 220–229, 2019.

[29] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using Bayesian binning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, pages 2901–2907, 2015.

[30] Markus Nagel, Marios Fournarakis, Rana Ali Amjad, Yelysei Bondarenko, Mart van Baalen, and Tijmen Blankevoort. A white paper on neural network quantization. *arXiv preprint arXiv:2106.08295*, 2021.

[31] Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In *CVPR Workshops*, pages 38–41, 2019.

[32] OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

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

[34] John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, 1999.

[35] Kush R Varshney. Trustworthy machine learning and artificial intelligence. *ACM XRDS: Crossroads*, 28(4):26–29, 2022.

[36] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*, 2024.

[37] Edwin B Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

[38] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

[39] Zhewei Yao, Reza Yazdani Aminabadi, Minjia Zhang, Xiaoxia Wu, Conglong Li, and Yuxiong He. ZeroQuant: Efficient and affordable post-training quantization for large-scale transformers. *Advances in Neural Information Processing Systems*, 35:27168–27183, 2022.

[40] Weijia Zhu, Siddharth Karamcheti, Alane Suhr, and Percy Liang. Detecting pretraining data from large language models. In *International Conference on Learning Representations*, 2024.

A Contamination Detection Examples

Example 1 with $J = 1.0$. Eval asks "what is the capital of France". Training has "what is the capital of France". Action is remove.

Example 2 with $J = 0.82$. Eval asks "explain photosynthesis in plants". Training has "describe photosynthesis process in plant cells". Matched 5-shingles include "photo", "synth", and "plant". Action is remove.

Example 3 with $J = 0.68$. Eval asks "what causes climate change". Training has "list drivers of global climate change". Action is flagged then kept.

B Detailed Results

See Table 5 in the main text for Wilson 95% CI with Holm-Bonferroni at $\alpha = 0.05$.

C MMLU Sampling Distribution

Table 6 shows MMLU per-subject sampling with stratified rate approximately 1.7%.

Table 6: MMLU per-subject sampling with stratified rate approximately 1.7%.

Subject	Source	Sampled
Abstract Algebra	100	2
Anatomy	135	2
Astronomy	152	3
Business Ethics	100	2
Clinical Knowledge	265	5
College Biology	144	2
College Chemistry	100	2
College CS	100	2
College Mathematics	100	2
College Medicine	173	3
College Physics	102	2
Computer Security	100	2
Conceptual Physics	235	4
Econometrics	114	2
Electrical Engineering	145	2
Elementary Math	378	6
Formal Logic	126	2
Global Facts	100	2
HS Biology	310	5
HS Chemistry	203	3
HS Computer Science	100	2
HS European History	165	3
HS Geography	198	3
HS Government	193	3
HS Macroeconomics	390	7
HS Mathematics	270	5
HS Microeconomics	238	4
HS Physics	151	3
HS Psychology	545	9
HS Statistics	216	4
HS US History	204	3
HS World History	237	4
Human Aging	223	4
Human Sexuality	131	2
International Law	121	2
Jurisprudence	108	2
Logical Fallacies	163	3
Machine Learning	112	2
Management	103	2

Continued on next page

Table 6 – continued from previous page

Subject	Source	Sampled
Marketing	234	4
Medical Genetics	100	2
Miscellaneous	783	13
Moral Disputes	346	6
Moral Scenarios	895	15
Nutrition	306	5
Philosophy	311	5
Prehistory	324	6
Professional Accounting	282	5
Professional Law	1534	26
Professional Medicine	272	5
Professional Psychology	612	10
Public Relations	110	2
Security Studies	245	4
Sociology	201	3
US Foreign Policy	100	2
Virology	166	3
World Religions	171	3
Total	14,042	240

D Reproducibility

We release complete artifacts including per-item predictions, contamination reports, high-resolution plots, and deterministic pipeline with manifests at `https://anonymous.4open.science/r/timealign-repro-E476`. By ensuring evaluations remain clean, calibrated, and efficient, TimeAlign supports trustworthy benchmarking of continually evolving models while respecting practical deployment limitations.