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

Anonymous Author(s)
Affiliation
Address
email

Abstract

Efficient evaluation of resource-constrained foundation models requires computational frugality and methodological rigor. We present TimeAlign, a contamination-aware framework combining temporal screening, textual decontamination, and uncertainty quantification. TimeAlign uses automated 5-shingle Jaccard detection with Cohen's $\kappa \approx 0.94$ and post-T0 temporal screening against 30,700 news documents. Evaluating Llama-3.1-8B and Qwen2.5-7B across MMLU, MMLU-Pro, and ARC under NF4 quantization, contamination inflates accuracy by 74.5 percentage points. Our 720-item clean evaluation completes in under 8 hours on 16GB GPUs. Temperature scaling reduces Smooth-ECE from 0.089 to 0.041 while NF4 introduces minimal degradation. 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-T0 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.

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].

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

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

T0 and dataset pinning. TimeAlign establishes temporal boundary T0 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.

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	T0 boundary	Temporal cutoff
2025-09+	WCEP-10, GDELT, CC-News	Post-T0 screening

Figure 1: Temporal screening timeline showing dataset commits, T0 boundary, and post-T0 screening sources.

Post-T0 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]. Adjudicators independently labeled matches as remove, flag, or keep without seeing model scores. Example with $J = 0.82$ shows eval "Explain photosynthesis in plants" matching training "Describe photosynthesis process in plant cells" with overlapping 5-shingles "photo", "synth", "plant", triggering removal. Full examples in Appendix A.

3.2 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].

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.

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.

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 1 shows contamination removal statistics.

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

Table 1: Contamination card showing removal counts per stage.

4 Results

Table 2 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].

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

Table 2: Performance collapse from contamination in internal case study.

Table 3 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, and temperature scaling mitigates miscalibration [17, 34].

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

Table 3: Clean evaluation after decontamination and temperature scaling.

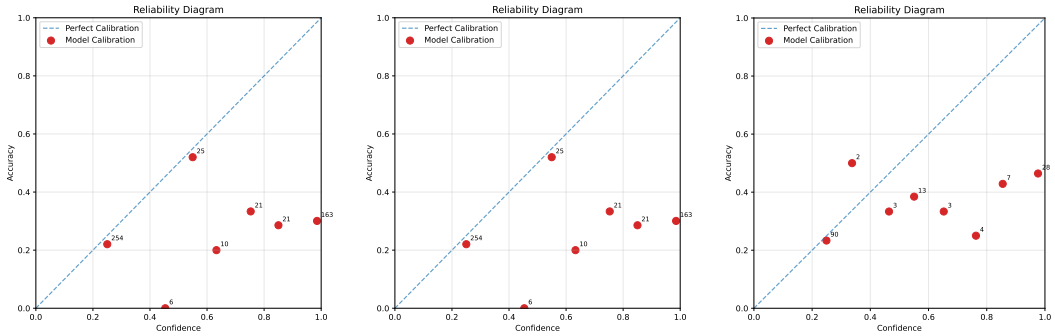
Figure 2 shows reliability diagrams. Uncalibrated models exhibit severe overconfidence; temperature scaling with $T \approx 2.2$ to 2.5 reduces Smooth-ECE by 46 to 54% [17]. Figure 3 presents risk-coverage curves with low nAURC from 0.12 to 0.22. For example, nAURC 0.18 on MMLU means selecting top 50% 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 abstention systems [20, 35].

5 Discussion

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].

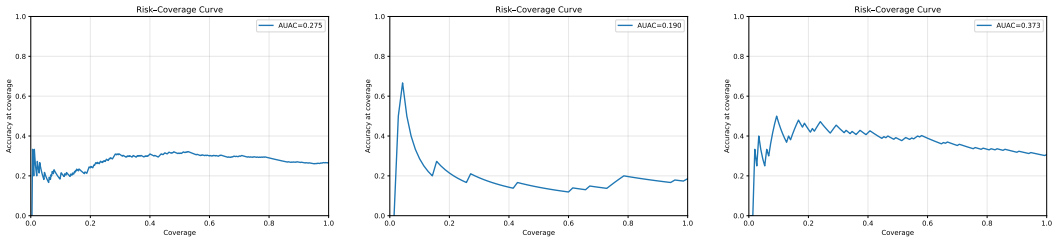
Small calibration splits address labeled data limitations. Practitioners obtain reliable calibration with minimal budget while reserving most data for testing. The contamination detection pipeline provides high-quality training data curation preventing memorization during continued pretraining [26, 21].

Quantization-aware calibration addresses efficiency-performance trade-offs. NF4 preserves calibration quality under temperature scaling, enabling deployment under fixed memory bud-



(a) MMLU uncal (b) MMLU cal (c) ARC cal

Figure 2: Reliability diagrams showing calibration improvement. Fitted temperatures $T \approx 2.2$ to 2.5. Bin sizes shown as numbers. Smooth-ECE computed with default Gaussian kernel bandwidth.



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

Figure 3: Risk-coverage curves showing limited selective prediction. Chance at 0.75 for 4-choice, 0.80 for 5-choice.

gets [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].

TimeAlign supports continual evaluation through automated T0 logging, fresh post-T0 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.

Our detector may miss sophisticated leakage including paraphrasing and cross-lingual variants [15, 40]. Evaluation scale with 240 per dataset balances power with 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.

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 equals 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 equals 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 equals 0.68. Eval asks what causes climate change. Training has list drivers of global climate change. Action is flagged then kept.

B Detailed Results

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

Table 4: Wilson 95% CI with Holm-Bonferroni at alpha equals 0.05.

C MMLU Sampling Distribution

Table 5: 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

continued on next page

Table 5 continued

Subject	Source	Sampled
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
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.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We state contamination detection with 74.5 pp inflation, temporal screening with 30,700 documents, quantization-aware calibration with 54% ECE reduction, and rolling evaluation with specific numbers matching my results.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss limitations in the Discussion covering detector coverage gaps with paraphrasing and cross-lingual variants, evaluation scale trade-offs with 240 items per dataset, Smooth-ECE bandwidth sensitivity, and computational requirements.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: My work is empirical without theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide model IDs with Llama-3.1-8B and Qwen2.5-7B, dataset commits with exact hashes and dates, quantization settings with NF4, scoring protocol using concatenative method, calibration procedure with 50 samples optimizing NLL, and metric definitions including Wilson CI and Holm-Bonferroni. The repository at anonymous.4open.science contains config files, environment specs, and random seed 42.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open Access to Data and Code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We release the repository at https://anonymous.4open.science/r/timealign-repro-E476 containing 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.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We specify Llama-3.1-8B in FP16 and NF4, Qwen2.5-7B in NF4, concatenative scoring with newline-Choice delimiter, temperature scaling with 50 samples per dataset minimizing NLL, Wilson 95% CI with Holm-Bonferroni, Smooth-ECE with default Gaussian kernel, SFT filtering removing 193 items, temporal screening removing 9,080 items, and stratified sampling drawing 240 items per dataset.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We report Wilson 95% confidence intervals for all accuracy estimates with Holm-Bonferroni correction for multiple comparisons across three datasets. The detailed results table in the appendix shows confidence interval lower and upper bounds for every model and dataset combination.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We state NF4 quantization enables 8B parameter evaluation on consumer GPUs with 16GB memory in the Discussion. We also specify evaluation throughput exceeding 100 items per hour and complete 720-item assessment completing in under 8 hours fitting overnight cycles.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: My work evaluates existing open models including Llama-3.1-8B and Qwen2.5-7B on public benchmarks including MMLU, MMLU-Pro, and ARC with transparent contamination detection using 5-shingle Jaccard similarity.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We address positive impacts including honest evaluation preventing resource misallocation, rolling protocols supporting transparency, and efficient evaluation enabling broader access in the Discussion. We also address limitations including detector coverage gaps and computational requirements.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: My work releases evaluation methodology including contamination detection code and benchmark results including per-item predictions, not pretrained models or high-risk datasets.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for Existing Assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite Llama-3.1-8B from Dubey et al 2024 with Apache 2.0 license, Qwen2.5-7B from Yang et al 2024 with Apache 2.0 license, MMLU from Hendrycks et al 2021 with MIT license at commit 7a00892, MMLU-Pro from Wang et al 2024 with MIT license at commit 241199e, and ARC from Clark et al 2018 with CC BY-SA 4.0 at commit 870fda1.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We release artifacts at https://anonymous.4open.science/r/timealign-repro-E476 including per-item prediction CSVs with confidence scores, contamination reports JSON with matched examples, high-resolution diagnostic PDFs including reliability diagrams and risk-coverage curves, pipeline code with environment manifests, and README with setup documentation.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

Justification: My work does not involve crowdsourcing or human subjects research.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: My work does not involve human subjects research.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM Usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research?

Answer: [NA]

Justification: LLMs including Llama-3.1-8B and Qwen2.5-7B are evaluation subjects being benchmarked, not methodological components used for developing my contamination detection, temporal screening, or calibration methods. My core methodology uses character-level shingling, temporal filtering, and temperature scaling without using LLMs as tools.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.