EKD: Entropy Knowledge Distillation for Efficient Language Model Compression

Almog Tavor *
Tel Aviv University
email@domain

Itay Ebenspander *
Tel Aviv University
email@domain

Neil Cnaan *
Tel Aviv University
email@domain

Abstract

When we'll finish.

1 Introduction

Large language models (LLMs) deliver state-of-the-art performance but are costly to serve and adapt. Knowledge distillation (KD) amortizes these costs by training a compact student to imitate a larger teacher (Hinton et al., 2015). While classic logit-based KD is effective for encoder-style models (e.g., DistilBERT (Sanh et al., 2019)), applying KD to auto-regressive LLMs raises two persistent obstacles: (i) **train–inference mismatch** in sequence generation, and (ii) **efficiency limits** from storing or querying full teacher logits.

To address (i), recent work proposes *on-policy* distillation that trains the student on prefixes it actually produces, with the teacher scoring those prefixes (Agarwal et al., 2024). To address (ii), sparse alternatives avoid caching the full distribution. However, deterministic top-K truncation of teacher logits (e.g., keeping only the largest K probabilities) yields biased supervision and degraded calibration, especially for small K (Anshumann et al., 2025; Shum et al., 2024). Storing more logits improves quality but erodes the efficiency goal.

We revisit *where* and *how* to apply teacher supervision for LLMs. Our thesis is that the KD budget should be spent *selectively at the token level* (where the teacher can add the most information) while keeping *unbiased* class-level supervision when it is applied. Concretely, we:

- introduce **token-selective distillation**: choose a subset of tokens for KD using scores such as teacher entropy, teacher–student KL, and student CE;
- estimate the teacher distribution at selected tokens with **Random-Sampling KD**—importance-sampling teacher classes and reweighting—so the gradient matches full KD in expectation while storing only a handful of logits (Anshumann et al., 2025);
- maintain standard cross-entropy (CE) on *all* tokens to stabilize training and calibration (Guo et al., 2017); and
- add simple curricula (increasing KD coverage over time) and optional EMA-based self-distillation for robustness.

Empirically, we compare against (a) full Random-Sampling KD at every token, (b) deterministic top-K logit truncation baselines including SLIM-style sparse logits (Raman et al., 2023), and (c) token/data selection schemes such as GLS for NMT (Wang et al., 2021), token-level uncertainty-aware post-training (Liu et al., 2025a), and progressive chain-of-thought distillation (Feng et al., 2024). We evaluate both in-distribution and under distribution shift using the ShiftKD protocol (Zhang et al., 2023). Our results indicate that token-selective, *unbiased* KD attains a stronger accuracy–efficiency Pareto frontier while preserving calibration.

Contributions. (1) A unified framework for *token-level* selection with *unbiased* class-level distillation; (2) practical scoring, curriculum, and bandit variants for allocating KD budget; (3) comprehensive evaluation vs. state-of-the-art sparse KD and selection baselines, including calibration and shift robustness.

2 Related Work

KD for LLMs. Foundational KD (Hinton et al., 2015) and encoder-model distillation (e.g., DistilBERT (Sanh et al., 2019)) have been extended

* Equal contribution.

to generative LMs: sequence-level KD for NMT (Kim and Rush, 2016), reverse-KL objectives for generation (MiniLLM) (Gu et al., 2023), and on-policy distillation (GKD) to mitigate exposure bias (Agarwal et al., 2024). Cross-tokenizer distillation has been addressed by Universal Logit Distillation (ULD) via optimal transport (Boizard et al., 2024).

Sparse logit KD and calibration. Deterministic top-K/percentile caching of teacher logits (e.g., SLIM (Raman et al., 2023) and top-5 variants) reduces storage but discards tail mass, inducing biased gradients and miscalibrated students. Random-Sampling KD (Anshumann et al., 2025) replaces truncation with importance sampling to provide *unbiased* estimates that match full-KD gradients in expectation with minimal overhead. Trustworthy distillation explicitly studies calibration and proposes processing the top-k teacher tokens to reduce miscalibration (Shum et al., 2024). We measure calibration via Expected Calibration Error (ECE) (Guo et al., 2017).

Selecting where to supervise. Beyond "how" to form the target, several works decide "where" to apply supervision. In NMT, Wang et al. (2021) select *words* by cross-entropy using batch-level and global (FIFO) queues (GLS). For post-training, token-level uncertainty-aware objectives select *tokens* by loss/entropy and add self-distillation to prevent OOD overfitting (Liu et al., 2025a). In reasoning, KPOD learns keypoint weights and a progressive schedule within chain-of-thought rationales (Feng et al., 2024). Data-centric selection via Selective Reflection Distillation curates training instances with a curriculum (Liu et al., 2025b). Our work differs by coupling *token-level* selection with *unbiased* logit sampling at selected positions, yielding efficiency without the calibration drawbacks of deterministic truncation.

Self-distillation and EMA teachers. Self-distillation regularizes students via a teacher derived from the student itself, e.g., Born-Again Networks (Furlanello et al., 2018) and Mean Teacher (EMA) consistency (Tarvainen and Valpola, 2017). We adopt an optional EMA self-distill term as a light regularizer complementary to external-teacher KD.

Evaluation under shift. We report in-distribution and distribution-shift results following ShiftKD (Zhang et al., 2023), which benchmarks KD methods under diversity and correlation shifts.

Command	Output	Command	Output
{\"a}	ä	{\c c}	ç
{\^e}	ê	{\u g}	ğ
{\`i}	ì	{\l}	ł
{\.I}	İ	{\~n}	ñ
{\o}	ø	{\H o}	ő
{\'u}	ú	{\v r}	ř
{\aa}	å	{\ss}	ß

Table 1: Example commands for accented characters, to be used in, *e.g.*, BibTEX entries.

Summary. Prior work either improves the *where* (token/word/sample selection, curricula) *or* the *how* (on-policy, reverse-KL, cross-tokenizer, sparse logits) of KD. We unify both: allocate KD budget to high-value tokens and deliver unbiased supervision there via Random-Sampling KD.

3 Introduction

These instructions are for authors submitting papers to ACL 2023 using LATEX. They are not self-contained. All authors must follow the general instructions for *ACL proceedings, as well as guidelines set forth in the ACL 2023 call for papers. This document contains additional instructions for the LATEX style files. The templates include the LATEX source of this document (`acl2023.tex`), the LATEX style file used to format it (`acl2023.sty`), an ACL bibliography style (`acl_natbib.bst`), an example bibliography (`custom.bib`), and the bibliography for the ACL Anthology (`anthology.bib`).

4 Engines

To produce a PDF file, pdfLATEX is strongly recommended (over original LATEX plus dvips+ps2pdf or dvipdf). XeLATEX also produces PDF files, and is especially suitable for text in non-Latin scripts.

5 Preamble

The first line of the file must be

```
\documentclass[11pt]{article}
```

To load the style file in the review version:

```
\usepackage[review]{ACL2023}
```

http://acl-org.github.io/ACLPUB/formatting.html

https://2023.aclweb.org/calls/main_conference/

Output	natbib command	Old ACL-style command
(Cooley and Tukey, 1965)	`\citep`	`\cite`
Cooley and Tukey, 1965	`\citealp`	no equivalent
Cooley and Tukey (1965)	`\citet`	`\newcite`
(1965)	`\citeyearpar`	`\shortcite`
Cooley and Tukey's (1965)	`\citeposs`	no equivalent
(FFT; Cooley and Tukey, 1965)	`\citep[FFT;][]`	no equivalent

Table 2: Citation commands supported by the style file. The style is based on the natbib package and supports all natbib citation commands. It also supports commands defined in previous ACL style files for compatibility.

For the final version, omit the `review` option:

```
\usepackage{ACL2023}
```

To use Times Roman, put the following in the preamble:

```
\usepackage{times}
```

(Alternatives like txfonts or newtx are also acceptable.) Please see the LaTeX source of this document for comments on other packages that may be useful. Set the title and author using `\title` and `\author`. Within the author list, format multiple authors using `\and` and `\And` and `\AND`; please see the LaTeX source for examples. By default, the box containing the title and author names is set to the minimum of 5 cm. If you need more space, include the following in the preamble:

```
\setlength\titlebox{<dim>}
```

where `<dim>` is replaced with a length. Do not set this length smaller than 5 cm.

6 Document Body

6.1 Footnotes

Footnotes are inserted with the `\footnote` command.

This is a footnote.

6.2 Tables and figures

See Table 1 for an example of a table and its caption. **Do not override the default caption sizes.**

6.3 Hyperlinks

Users of older versions of LaTeX may encounter the following error during compilation:

```
\pdfendlink ended up in different
nesting level than \pdfstartlink.
```

This happens when pdfLaTeX is used and a citation splits across a page boundary. The best way to fix this is to upgrade LaTeX to 2018-12-01 or later.

6.4 Citations

Table 2 shows the syntax supported by the style files. We encourage you to use the natbib styles. You can use the command `\citet` (cite in text) to get "author (year)" citations, like this citation to a paper by Gusfield (1997). You can use the command `\citep` (cite in parentheses) to get "(author, year)" citations (Gusfield, 1997). You can use the command `\citealp` (alternative cite without parentheses) to get "author, year" citations, which is useful for using citations within parentheses (e.g. Gusfield, 1997).

6.5 References

The LaTeX and BibTeX style files provided roughly follow the American Psychological Association format. If your own bib file is named `custom.bib`, then placing the following before any appendices in your LaTeX file will generate the references section for you:

```
\bibliographystyle{acl_natbib}
\bibliography{custom}
```

You can obtain the complete ACL Anthology as a BibTeX file from `https://aclweb.org/anthology/anthology.bib.gz`. To include both the Anthology and your own .bib file, use the following instead of the above.

```
\bibliographystyle{acl_natbib}
\bibliography{anthology,custom}
```

Please see Section 7 for information on preparing BibTeX files.

6.6 Appendices

Use `\appendix` before any appendix section to switch the section numbering over to letters. See Appendix A for an example.

7 BibTeX Files

Unicode cannot be used in BibTeX entries, and some ways of typing special characters can disrupt BibTeX's alphabetization. The recommended way of typing special characters is shown in Table 1.

Please ensure that BibTeX records contain DOIs or URLs when possible, and for all the ACL materials that you reference. Use the `doi` field for DOIs and the `url` field for URLs. If a BibTeX entry has a URL or DOI field, the paper title in the references section will appear as a hyperlink to the paper, using the hyperref LaTeX package.

Limitations

ACL 2023 requires all submissions to have a section titled "Limitations", for discussing the limitations of the paper as a complement to the discussion of strengths in the main text. This section should occur after the conclusion, but before the references. It will not count towards the page limit. The discussion of limitations is mandatory. Papers without a limitation section will be desk-rejected without review.

While we are open to different types of limitations, just mentioning that a set of results have been shown for English only probably does not reflect what we expect. Mentioning that the method works mostly for languages with limited morphology, like English, is a much better alternative. In addition, limitations such as low scalability to long text, the requirement of large GPU resources, or other things that inspire crucial further investigation are welcome.

Ethics Statement

Scientific work published at ACL 2023 must comply with the ACL Ethics Policy. We encourage all authors to include an explicit ethics statement on the broader impact of the work, or other ethical considerations after the conclusion but before the references. The ethics statement will not count toward the page limit (8 pages for long, 4 pages for short papers).

Acknowledgements

This document has been adapted by Jordan Boyd-Graber, Naoaki Okazaki, Anna Rogers from the style files used for earlier ACL, EMNLP and NAACL proceedings, including those for EACL 2023 by Isabelle Augenstein and Andreas Vlachos, EMNLP 2022 by Yue Zhang, Ryan Cotterell and Lea Frermann, ACL 2020 by Steven Bethard, Ryan Cotterell and Rui Yan, ACL 2019 by Douwe Kiela and Ivan Vulić, NAACL 2019 by Stephanie Lukin and Alla Roskovskaya, ACL 2018 by Shay Cohen, Kevin Gimpel, and Wei Lu, NAACL 2018 by Margaret Mitchell and Stephanie Lukin, BibTeX suggestions for (NA)ACL 2017/2018 from Jason Eisner, ACL 2017 by Dan Gildea and Min-Yen Kan, NAACL 2017 by Margaret Mitchell, ACL 2012 by Maggie Li and Michael White, ACL 2010 by Jing-Shin Chang and Philipp Koehn, ACL 2008 by Johanna D. Moore, Simone Teufel, James Allan, and Sadaoki Furui, ACL 2005 by Hwee Tou Ng and Kemal Oflazer, ACL 2002 by Eugene Charniak and Dekang Lin, and earlier ACL and EACL formats written by several people, including John Chen, Henry S. Thompson and Donald Walker. Additional elements were taken from the formatting instructions of the *International Joint Conference on Artificial Intelligence* and the *Conference on Computer Vision and Pattern Recognition*.

References

Rishabh Agarwal, Nino Vieillard, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. 2024. On-policy distillation of language models: Learning from self-generated mistakes. *arXiv preprint arXiv:2306.13649*.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L_1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Piyush Anshumann, Rishabh Agarwal, and Olivier Bachem. 2025. Sparse logit sampling: Accelerating knowledge distillation in llms. *arXiv preprint arXiv:2503.16870*.

Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance detection with bidirectional conditional encoding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 876–885, Austin, Texas. Association for Computational Linguistics.

Nicolas Boizard, Pierre Colombo, Pablo Piantanida, and Chloé Clavel. 2024. Towards cross-tokenizer

https://www.aclweb.org/portal/content/acl-code-ethics

distillation: the universal logit distillation loss for llms. *arXiv preprint arXiv:2402.12030*.

James W. Cooley and John W. Tukey. 1965. An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, 19(90):297–301.

Kaituo Feng, Changsheng Zhang, Ye Wu, and Guoren Zhang. 2024. Keypoint-based progressive chain-of-thought distillation for llms. *arXiv preprint arXiv:2405.16064*.

Tommaso Furlanello, Zachary C. Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. 2018. Born-again neural networks. In *Proceedings of Machine Learning Research*, volume 80, pages 1607–1616.

James Goodman, Andreas Vlachos, and Jason Naradowsky. 2016. Noise reduction and targeted exploration in imitation learning for Abstract Meaning Representation parsing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1–11, Berlin, Germany. Association for Computational Linguistics.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2023. Knowledge distillation of large language models. *arXiv preprint arXiv:2306.08543*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. *arXiv preprint arXiv:1706.04599*.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

Mary Harper. 2014. Learning from 26 languages: Program management and science in the babel program. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, page 1, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. *arXiv preprint arXiv:1606.07947*.

Jiahao Liu, Tongxu Chen, Chengming Zhang, Hongwei Liu, Haochen Wang, and Nenghai Peng. 2025a. Token-level uncertainty-aware objective for language model post-training. *arXiv preprint arXiv:2503.16511*.

Ming Liu, Jie Qi, Yicheng Wang, Jiawei Han, and Lei Chen. 2025b. Selective reflection-tuning: Student-selected data recycling for llm instruction-tuning. *arXiv preprint*.

Neeraj Raman, Siddharth Vare, Apurva Srinivasan, Vignesh Chandra, and Kush Khandelwal. 2023. For distillation, tokens are not all you need. In *OpenReview*.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Hubert Shum, Linlin Wang, Jiajun Chen, and Liang Zheng. 2024. On the calibration and trustworthiness of knowledge distillation. *arXiv preprint*.

Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in Neural Information Processing Systems*, 30.

Fusheng Wang, Jianhao Xu, Fandong Zhao, Jianfeng Jia, and Jie Zhou. 2021. Selective knowledge distillation for neural machine translation. *arXiv preprint arXiv:2105.12967*.

Jinghan Zhang, Rishabh Agarwal, Igor Babuschkin, and Olivier Bachem. 2023. Benchmarking knowledge distillation under distribution shift. *arXiv preprint arXiv:2312.16242*.

A Example Appendix

This is a section in the appendix.