

The evaluation and restoration of PTQ models’ conversational capabilities

Anonymous ACL submission

Abstract

This document is a supplement to the general instructions for *ACL authors. It contains instructions for using the \LaTeX style files for ACL conferences. The document itself conforms to its own specifications, and is therefore an example of what your manuscript should look like. These instructions should be used both for papers submitted for review and for final versions of accepted papers.

1 Introduction

Quantization Survey: Evaluating Quantized Large Language Models (Li et al., 2024), Give me BF16 or give me death (Kurtic et al., 2024), A Comprehensive Evaluation of Quantization Strategies for Large Language Models (Jin et al., 2024) An Empirical Study of LLaMA3 Quantization: From LLMs to MLLMs(Huang et al., 2024)

Evaluation Benchmarks: Base of RoPE Bounds Context Length (Xu et al., 2024), Can perplexity reflect large language model’s ability in long text understanding?(Hu et al., 2024)

Quantization Methods: GPTQ (Frantar et al., 2022), QLoRA (Dettmers et al., 2023), SmoothQuant (Xiao et al., 2023), AWQ (Lin et al., 2024), OWQ (Lee et al., 2024a), OmniQuant (Shao et al., 2024), LRQuant(Zhao et al., 2024),SpQR (Dettmers et al., 2024),LLM.int8() (Dettmers et al., 2022)

Reinforcement learning: Improving Conversational Abilities of Quantized Large Language Models via Direct Preference Alignment (Lee et al., 2024b), REFT: Reasoning with REinforced Fine-Tuning (Trung et al., 2024), Understanding Reinforcement Learning-Based Fine-Tuning of Diffusion Models: A Tutorial and Review (Uehara et al., 2024)

- Mixed Precision Quantization. (1) Magnitude-based. (2) Saliency-based.

- Ultra low-bit Quantization.

These instructions are for authors submitting papers to *ACL conferences using \LaTeX . They are not self-contained. All authors must follow the general instructions for *ACL proceedings,¹ and this document contains additional instructions for the \LaTeX style files.

The templates include the \LaTeX source of this document (acl_latex.tex), the \LaTeX style file used to format it (acl.sty), an ACL bibliography style (acl_natbib.bst), an example bibliography (custom.bib), and the bibliography for the ACL Anthology (anthology.bib).

2 Engines

To produce a PDF file, pdf \LaTeX is strongly recommended (over original \LaTeX plus dvips+ps2pdf or dvi2pdf). Xe \LaTeX also produces PDF files, and is especially suitable for text in non-Latin scripts.

3 Preamble

The first line of the file must be

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

To load the style file in the review version:

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

For the final version, omit the review option:

```
\usepackage{acl}
```

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

```
\usepackage{times}
```

¹<http://acl-org.github.io/ACL/PUB/formatting.html>

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

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

(Alternatives like txfonts or newtx are also acceptable.)

Please see the L^AT_EX 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 L^AT_EX 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.

4 Document Body

4.1 Footnotes

Footnotes are inserted with the `\footnote` command.²

4.2 Tables and figures

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

As much as possible, fonts in figures should conform to the document fonts. See Figure 1 for an example of a figure and its caption.

Using the `graphicx` package `graphics` files can be included within figure environment at an appropriate point within the text. The `graphicx` package supports various optional arguments to control the appearance of the figure. You must include it explicitly in the L^AT_EX preamble (after the `\documentclass` declaration and before `\begin{document}`) using `\usepackage{graphicx}`.

²This is a footnote.



Figure 1: A figure with a caption that runs for more than one line. Example image is usually available through the mwe package without even mentioning it in the preamble.

4.3 Hyperlinks

Users of older versions of L^AT_EX may encounter the following error during compilation:

```
\pdfendlink ended up in different nest-
ing level than \pdfstartlink.
```

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

4.4 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).

A possessive citation can be made with the command `\citeposs`. This is not a standard natbib command, so it is generally not compatible with other style files.

4.5 References

The L^AT_EX 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 L^AT_EX file will generate the references section for you:

```
\bibliography{custom}
```

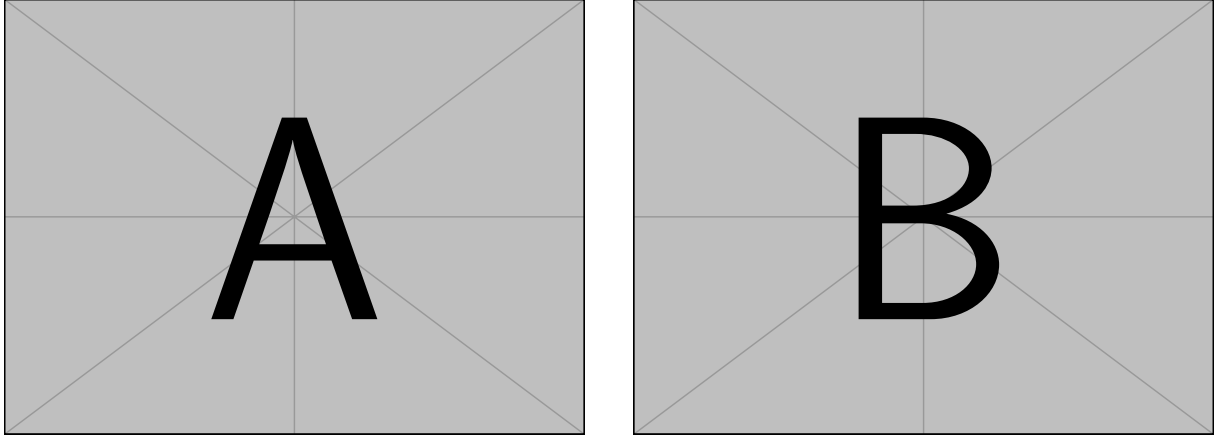


Figure 2: A minimal working example to demonstrate how to place two images side-by-side.

Output	natbib command	ACL only command
(Gusfield, 1997)	<code>\citep</code>	
Gusfield, 1997	<code>\citealp</code>	
Gusfield (1997)	<code>\citet</code>	
(1997)	<code>\citeyearpar</code>	
Gusfield’s (1997)		<code>\citeposs</code>

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.

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.

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

Please see Section 5 for information on preparing BibTeX files.

4.6 Equations

An example equation is shown below:

$$A = \pi r^2 \quad (1)$$

Labels for equation numbers, sections, subsections, figures and tables are all defined with the `\label{label}` command and cross references to them are made with the `\ref{label}` command.

This an example cross-reference to Equation 1.

4.7 Appendices

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

5 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 L^AT_EX package.

Acknowledgments

This document has been adapted by Steven Bethard, Ryan Cotterell and Rui Yan from the instructions for earlier ACL and NAACL proceedings, including those for 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-

178	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 <i>International Joint Conference on Artificial Intelligence</i> and the <i>Conference on Computer Vision and Pattern Recognition</i> .	
189	Limitations	
190	References	
191	Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. <i>Journal of Machine Learning Research</i> , 6:1817–1853.	
192		
193	Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In <i>Proceedings of the 24th International Conference on Machine Learning</i> , pages 33–40.	
194		
195	Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale. <i>Advances in Neural Information Processing Systems</i> , 35:30318–30332.	
196		
197	Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms . In <i>Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023</i> .	
198		
199	Tim Dettmers, Ruslan Svirschevski, Vage Egiazarian, Denis Kuznedelev, Elias Frantar, Saleh Ashkboos, Alexander Borzunov, Torsten Hoefer, and Dan Alistarh. 2024. Spqr: A sparse-quantized representation for near-lossless LLM weight compression . In <i>The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024</i> . OpenReview.net.	
200		
201	Elias Frantar, Saleh Ashkboos, Torsten Hoefer, and Dan Alistarh. 2022. Gptq: Accurate post-training quantization for generative pre-trained transformers. <i>arXiv preprint arXiv:2210.17323</i> .	
202		
203	Dan Gusfield. 1997. <i>Algorithms on Strings, Trees and Sequences</i> . Cambridge University Press, Cambridge, UK.	
204		
205	Yutong Hu, Quzhe Huang, Mingxu Tao, Chen Zhang, and Yansong Feng. 2024. Can perplexity reflect large language model’s ability in long text understanding? In <i>The Second Tiny Papers Track at ICLR 2024, Tiny Papers @ ICLR 2024, Vienna, Austria, May 11, 2024</i> . OpenReview.net.	
206		
207	Wei Huang, Xingyu Zheng, Xudong Ma, Haotong Qin, Chengtao Lv, Hong Chen, Jie Luo, Xiaojuan Qi, Xianglong Liu, and Michele Magno. 2024. An empirical study of llama3 quantization: From llms to mllms. <i>arXiv preprint arXiv:2404.14047</i> .	232
208		233
209		234
210		235
211		236
212	Renren Jin, Jiangcun Du, Wuwei Huang, Wei Liu, Jian Luan, Bin Wang, and Deyi Xiong. 2024. A comprehensive evaluation of quantization strategies for large language models . In <i>Findings of the Association for Computational Linguistics: ACL 2024</i> , pages 12186–12215, Bangkok, Thailand. Association for Computational Linguistics.	237
213		238
214		239
215		240
216		241
217		242
218		243
219	Eldar Kurtic, Alexandre Marques, Shubhra Pandit, Mark Kurtz, and Dan Alistarh. 2024. "give me bf16 or give me death"? accuracy-performance trade-offs in llm quantization. <i>arXiv preprint arXiv:2411.02355</i> .	244
220		245
221		246
222		247
223	Changhun Lee, Jungyu Jin, Taesu Kim, Hyungjun Kim, and Eunhyeok Park. 2024a. OWQ: outlier-aware weight quantization for efficient fine-tuning and inference of large language models . In <i>Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada</i> , pages 13355–13364. AAAI Press.	248
224		249
225		250
226		251
227		252
228		253
229		254
230		255
231		256
232		257
233	Janghwan Lee, Seongmin Park, Sukjin Hong, Minsoo Kim, Du-Seong Chang, and Jungwook Choi. 2024b. Improving conversational abilities of quantized large language models via direct preference alignment . In <i>Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 11346–11364, Bangkok, Thailand. Association for Computational Linguistics.	258
234		259
235		260
236		261
237		262
238		263
239		264
240		265
241	Shiyao Li, Xuefei Ning, Luning Wang, Tengxuan Liu, Xiangsheng Shi, Shengen Yan, Guohao Dai, Huazhong Yang, and Yu Wang. 2024. Evaluating quantized large language models . In <i>Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024</i> . OpenReview.net.	266
242		267
243		268
244		269
245		270
246		271
247		272
248	Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. 2024. AWQ: activation-aware weight quantization for on-device LLM compression and acceleration . In <i>Proceedings of the Seventh Annual Conference on Machine Learning and Systems, MLSys 2024, Santa Clara, CA, USA, May 13-16, 2024</i> . mlsys.org.	273
249		274
250		275
251		276
252		277
253		278
254		279
255		280
256	Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser . <i>Computing Research Repository</i> , arXiv:1503.06733. Version 2.	281
257		282
258		283
259		284
260	Wenqi Shao, Mengzhao Chen, Zhaoyang Zhang, Peng Xu, Lirui Zhao, Zhiqian Li, Kaipeng Zhang, Peng Gao, Yu Qiao, and Ping Luo. 2024. Omniquant: Omnidirectionally calibrated quantization for large	285
261		286
262		287
263		288

language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Luong Trung, Xinbo Zhang, Zhanming Jie, Peng Sun, Xiaoran Jin, and Hang Li. 2024. [ReFT: Reasoning with reinforced fine-tuning](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7601–7614, Bangkok, Thailand. Association for Computational Linguistics.

Masatoshi Uehara, Yulai Zhao, Tommaso Biancalani, and Sergey Levine. 2024. Understanding reinforcement learning-based fine-tuning of diffusion models: A tutorial and review. *arXiv preprint arXiv:2407.13734*.

Guangxuan Xiao, Ji Lin, Mickaël Seznec, Hao Wu, Julien Demouth, and Song Han. 2023. [Smoothquant: Accurate and efficient post-training quantization for large language models](#). In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 38087–38099. PMLR.

Mingyu Xu, Xin Men, Bingning Wang, Qingyu Zhang, Hongyu Lin, Xianpei Han, et al. 2024. Base of rope bounds context length. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Jiaqi Zhao, Miao Zhang, Chao Zeng, Ming Wang, Xuebo Liu, and Liqiang Nie. 2024. [LRQuant: Learnable and robust post-training quantization for large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2240–2255, Bangkok, Thailand. Association for Computational Linguistics.

A Example Appendix

This is an appendix.