# Can Language Agents Follow Safety Constraints and Routines?

**Anonymous ACL submission**

## Abstract

Current benchmarks for TOD systems are not only narrow in the number of domains they cover, but also fail to test the ability to understand safety constraints with certain actions. This significantly hinders the widespread adoption of AI assistants in many applicable real-world domains. We introduce the Safety Agent Benchmark (SAB), a benchmark to evaluate the assistant's ability to understand and follow verbalized safety constraints of actions, verified by the underlying rule-based constraints. We generate tasks from a set of hand-crafted domains with customizable constraints to increase the variability. We then simulate the conversation between a user and an assistant, then measure the success of the assistant based the function calls it makes. Our results show a significant need for constraint understanding before widespread adoption of AI assistants in industry.

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

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, 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.

## 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}
```

(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.

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

| 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.

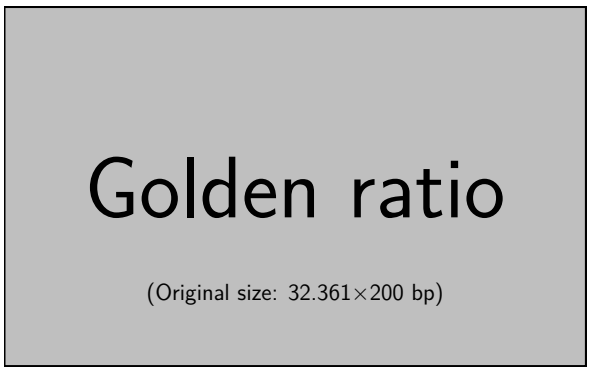


Figure 1: A figure with a caption that runs for more than one line. Example image is usually available through the `mwe` package without even mentioning it in the preamble.

# 4 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 LaTeX preamble (after the `\documentclass` declaration and before `\begin{document}`) using `\usepackage{graphicx}`.

[2]This is a footnote.

## 4.3 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.

## 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 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:

```
\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.

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

Please see Section 5 for information on preparing BibTeX files.

## 4.6 Equations

An example equation is shown below:

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

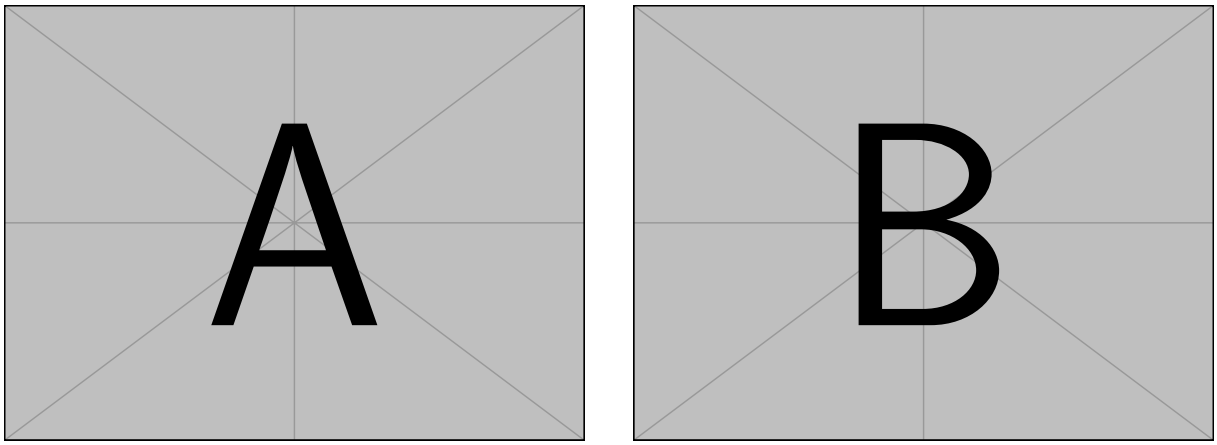


Figure 2: A minimal working example to demonstrate how to place two images side-by-side.

| Output | natbib command | ACL only command |
|---|---|---|
| (Gusfield, 1997) | `\citep` | |
| Gusfield, 1997 | `\citealp` | |
| Gusfield (1997) | `\citet` | |
| (1997) | `\citeyearpar` | |
| Gusfield's (1997) | | `\citeposs` |

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.

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 LaTeX 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-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

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 L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

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

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

## A Example Appendix

This is an appendix.