

Researching Responsible and Trustworthy Natural Language Processing

What makes an NLP paper?

Emily Allaway

(With slides adapted from Frank Keller)

28 September 2026

School of Informatics
University of Edinburgh

eallaway@ed.ac.uk

Publishing in *CL

Anatomy of an (NLP) Paper

Publishing in *CL

Publication Process

Formatting

Anatomy of an (NLP) Paper

Publishing in *CL

Publication Process

Formatting

Anatomy of an (NLP) Paper

- ACL - Association for Computational Linguistics
 - *ACL - various associated conferences (ex: EACL)
 - ACL Anthology
- ARR - ACL Rolling Review

Venues and more acronyms

Mostly publish in conferences and workshops

- NLP specific conferences
 - EMNLP, ACL, NAACL, EACL, AACL, COLING, LREC, ...
 - Findings
- Broader ML/AI conferences where people also submit
 - COLM, AAAI, ICLR, NeurIPS
- NLP journals
 - TACL, CL
 - Journals can have benefits:
<https://ehudreiter.com/2018/12/11/publish-in-journals/>
- Many many workshops and shared tasks

Publication process

- | | |
|---|---------------------------|
| 1. Reviewing | [Reviewers] |
| 2. (sometimes) Author response | [Authors] |
| 3. Metareview | [Area Chair (AC)] |
| 4. Decision (accept/reject) | [Senior Area Chair (SAC)] |
| 5. Presentation format decision (oral/poster) | |

Steps 1-3 → ARR

Steps 4,5 → Individual venue

Cycle	Submission	Reviewer registration	Reviews due	Author Response	Meta-reviews release date	Cycle End
May 2025	May 19	May 21	June 18	June 26-July 2	July 23	July 27
July 2025	July 28	July 30	September 2*	September 9-15	October 2	October 5
October 2025	October 6	October 8	November 10	November 18-24	December 11	December 14
January 2026	January 5	January 7	TBA	TBA	TBA	March 15

Venue	Final ARR Submission Date	Commitment Date
NAACL 2025	October 15, 2024	December 16, 2024
ACL 2025	February 15, 2025	April 20, 2025
EMNLP 2025	May 19, 2025	July 31, 2025
AACL 2025	July 28, 2025	October 10, 2025
EACL 2026	October 6, 2025	December 14, 2025

Publishing in *CL

Publication Process

Formatting

Anatomy of an (NLP) Paper

Formatting & other guidelines

- Formatting is *not* just a suggestion
- Some issues can lead to automatic rejection
<https://aclrollingreview.org/authorchecklist>
- Detailed guidelines exist:
<https://acl-org.github.io/ACLPUB/formatting.html>
- Also reviewer guidelines:
<https://aclrollingreview.org/reviewerguidelines>

Be careful using color

× *Too much color* can be distracting

from the FactScore dataset (Min et al., 2023) and leverage the FactScore decomposition engine and verifier to evaluate the model's outputs.

Rationalization (Binary) We use three prompt datasets requiring binary responses with justification (Zhang et al., 2024): identifying prime numbers, finding a senator who represented a specific state and attended a specific college, and identifying if a flight sequence exists between any two cities. *Decomposition and Verification:* The correct answer is 'Yes' for primality testing and 'No' for senator search and graph connectivity; the opposite response and corresponding justification is considered hallucination.

Rationalization (Numerical) Prompts in this category ask the model to count entities satisfying a condition, providing a numerical answer followed by the list of entities. We generate 1014 prompts with unique correct answers. *Decomposition and Verification:* We use Llama-2-70B to extract listed entities and verify them against a gazetteer.

Scientific Attribution We investigate model hallucinations of scientific references for false claims.

We create prompts featuring inaccurate statements, misconceptions, incorrect answers to questions, and misleading claims, sourced from HeliNet (Himmelstein et al., 2017), TruthfulQA (Lin et al., 2022), COVID-19 Lies (Hossain et al., 2020), and SciFact (Wadden et al., 2020). *Decomposition and verification:* Model responses are decomposed into atomic units (reference titles), and verified against the S2 index (Kinney et al., 2023).

Historical Events We compile a list of 400 noteworthy individuals and extract 1500 pairs with non-overlapping lifespans, making meetings unlikely. *Decomposition and Verification:* We use Llama-2-70B to determine whether the response confirms or denies a meeting. Confirmations or failure to abstain are classified as hallucinations.

False Presuppositions Prompts ask a model to list N entities that satisfy a condition, where N is larger than the number of entities satisfying that condition. *Decomposition and Verification:* Hallucinated units are items not meeting the condition.

Verification Accuracy We examine the accuracy of verifiers that use LLMs in their pipeline. These

From Ravichander et al. (2025).

Be careful using color

✓ Use color to help comprehension

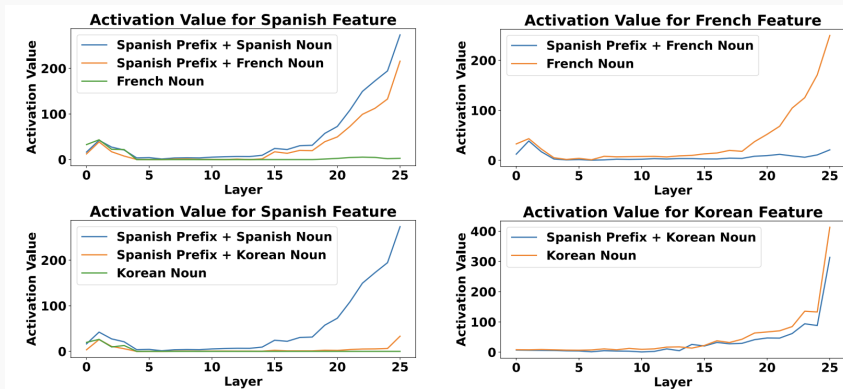
Humor	Martha walked into a pastry shop. After surveying all the pastries, she decided on a chocolate pie. "I'll take that one," Martha said to the attendant, "the whole thing." "Shall I cut it into four or eight pieces?" the attendant asked.	1. Correct Martha said, "Four pieces, please; I'm on a diet." 2. Literal Martha said: "Well, there are five people for dessert tonight, so eight pieces will be about right." 3. DistractorAssociative Martha said, "You make the most delicious sweet rolls in town." 4. DistractorFunny Then the attendant squirted whipped cream in Martha's face. 5. DistractorNeutral Martha said, "My leg is hurting so much."
Coherence	Mary's exam was about to start. Her palms were sweaty.	1. Correct Coherent 2. Incorrect Incoherent

Table 1: Sample item from each task in our evaluation. All items are originally curated by [Floyd et al. \(In prep\)](#).

task, with annotated answer options. Green labels indicate the target pragmatic interpretation.³ Blue labels indicate the literal interpretation. Red labels indicate incorrect non-literal interpretations, which are based on heuristics such as lexical similarity to the story, thus serving as distractor options.

Be aware of accessibility

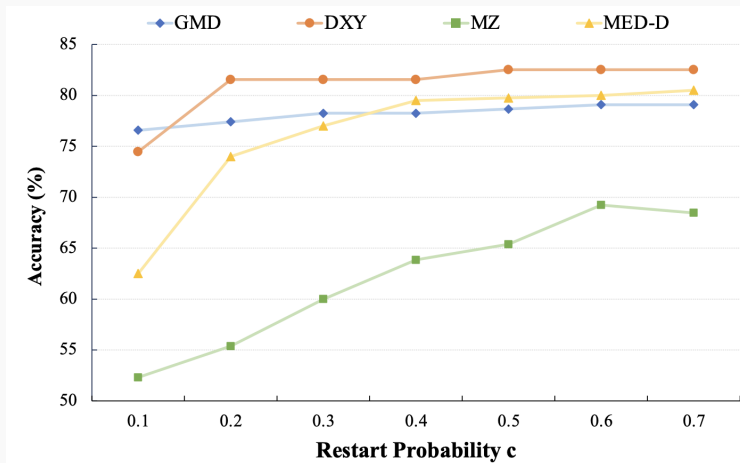
× *Unmarked lines* may be indistinguishable



From [Deng et al. \(2025\)](#).

Be aware of accessibility

✓ Use symbols to differentiate lines



- Writing advice
 - https://psresnik.github.io/writing_advice.html
- Research & publishing
 - <https://ehudreiter.com/2020/04/06/is-a-paper-scientifically-solid/>
 - <https://ehudreiter.com/2016/12/23/good-papers-are-hard-to-publish/>

Publishing in *CL

Anatomy of an (NLP) Paper

Title

Abstract

Introduction

Middle

Conclusion

Extras

Purpose: Inform

Most scientific writing has two specific purposes: *to inform and to persuade*. The level of persuasion varies: instructions require very little, a grant proposal requires a lot of persuasion.

Analogy from [Alley \(2018\)](#): a scientific document provides path that leads the reader up the mountain of your scientific expertise. If your purpose is merely to inform:

- you need to provide a path up the mountain
- it can be gentle (simple content) or steep (complex content)
- but you need to make sure readers can follow, break down the information, provide “vistas of understanding”

Purpose: Persuade

To inform, you need to answer *what, where, when, how*. To persuade, you also need to answer *why*. You need to build *credibility* with the audience:

- expend extra words to persuade; it's not about being maximally efficient
- you may not take most direct path up the mountain; it's more like navigating a boulder field
- the writing style changes: from lists (informative) to longer paragraphs (persuasive)

Persuasive writing explains why this is the right topic, research question, method, and technique.

Publishing in *CL

Anatomy of an (NLP) Paper

Title

Abstract

Introduction

Middle

Conclusion

Extras

How to write titles:

- “short and sweet” is great advice for the title of novels, but it doesn’t work well for titles of scientific papers
- the title should enable the audience to decide whether to read the paper or not
- in a bibliography or a web search readers will *only* see the title
- use the title to specify the scope of the document: identify the field of work and separate it from others in the field

Example titles from ACL 2025

- One-sentence summary of the result of the study
 - Statistical Deficiency for Task Inclusion Estimation
- Key result: details
 - Language Models Resist Alignment: Evidence From Data Compression
- Dataset name: description
 - AfriMed-QA: A Pan-African, Multi-Specialty, Medical Question-Answering Benchmark Dataset
- Method name: description
 - AGrail: A Lifelong Agent Guardrail with Effective and Adaptive Safety Detection
- Catchy phrase + description
 - "Yes, My LoRD." Guiding Language Model Extraction with Locality Reinforced Distillation
 - Whose Boat Does it Float? Improving Personalization in Preference Tuning via Inferred User Personas

Publishing in *CL

Anatomy of an (NLP) Paper

Title

Abstract

Introduction

Middle

Conclusion

Extras

Why is the Abstract Important?

For readers:

- find out what the paper is about (beyond the information in the title)
- get a summary of the most important findings
- decide whether they want to read the paper or not

For reviewers:

- find out whether the paper is within their area of expertise
- decide whether they want to review the paper or not
- form a first opinion about the quality of the paper

Many readers will *only read the title and the abstract*. So this is the one chance to get your message across!

Why is the Abstract Important?

The abstract can also be a tool for the writer:

- helps you decide what your most important points are
- helps you clarify the overall argumentation of the paper
- provides a way of repeating important information
- allows you to influence who will read (and review!) the paper

Content of the Abstract

Zobel (1997) offers the following practical advice:

- the abstract is typically a single paragraph of about 50–200 words
- it presents a summary of the paper's aims, scope, and conclusions
- do not use acronyms, mathematics, abbreviations, citations (the abstract should be self-contained!)
- be as specific as possible (instead of *we improve the state of the art*, write things like *we improve the state of the art by 3.5%*)
- but only include *important* details.

Organization of the Abstract

Zobel (1997) suggests to start by writing one sentence on each of the following:

1. A general statement introducing the broad research area.
2. An explanation of the specific problem to be solved.
3. A review of existing solutions and their limitations.
4. An outline of the proposed new solution.
5. A summary of how the solution was evaluated and the result of the evaluation.

So you start with five sentences, but then you can add additional sentences, re-write the ones you have, merge them, etc.

Exercise 1: writing an abstract

First, consider your own pitch. Swap pitches

- Write the 5 sentence abstract suggested by Zobel
 - For sentence 5, just include how it was evaluated

Swap pitches. For your partner's pitch

- Try to write down the 5 abstract sentences
- What information are you missing?

Publishing in *CL

Anatomy of an (NLP) Paper

Title

Abstract

Introduction

Middle

Conclusion

Extras

Acts as an overview of the entire paper, sets out the *story* of the paper

- What is the topic?
- Why is the work important?
- How is the work different from prior work?
- What are the work's contributions?

One potential outline of an introduction

- ¶1. Define central topic
 - Try to define why this is important → hook
 - Potentially use an example → often a figure
- ¶2-3. Explain why the topic is important and/or challenging
 - End with a way to clearly transition to what you did
- ¶4-5. Summarise methods & experiments
 - Your research questions
 - If relevant
 - Basic dataset information
 - Experiments overview + key findings
- ¶6. Contributions list

The first paragraph

Often the most difficult one to write

- Give the audience a reason to keep reading.
- The importance of the work can derive from its applications, but also from pure curiosity.
- In that case, you need to instill this curiosity in the reader!
- This determines how you need to justify the importance of your work.

The classic “Figure 1”

Many NLP papers include a first-page figure. It can do a lot of work for you.

- Include an example. This helps
 - Explain terminology or
 - Illustrate a task
- Can illustrate methodology
 - Visually simplifies

Title + figure $\Rightarrow$ illustrate what the paper is about

Example 1: starting paragraph & figure

- ✓ Good: illustration of example & importance
- ✗ *Not good*: isn't clear what the paper does

1 Introduction

Stance detection, automatically identifying positions on a specific topic in text (Mohammad et al., 2017), is crucial for understanding how information is presented in everyday life. For example, a news article on crime may also implicitly take a position on immigration (see Table 1).

There are two typical approaches to stance detection: *topic-specific* stance (developing topic-

Topic: immigration **Stance:** against

Text: The jury's verdict will ensure that another violent criminal alien will be removed from our community for a very long period ...

Table 1: Example snippet from Fox News describing a crime but taking a stance *against* immigration. Phrases indicating stance are highlighted.

From Allaway and McKeown (2020).

Example 2: starting paragraph & figure

✓ Good: illustration of what was done

✗ *Not good*: not using the figure

1 Introduction

Text comprehension at the discourse level entails understanding higher level structures between sentences and paragraphs, beyond the meaning of individual words. The linguistic framework of Questions Under Discussion (QUD) (Van Kuppevelt, 1995; Roberts, 2012; Benz and Jasinskaja, 2017) views the progression of discourse as continuously posing (implicit) questions (i.e., QUDs) and answering them; thus each sentence in a monologue is an answer, or part of an answer, to a QUD. In Figure 2, the third sentence answers the implicit QUD, “What does Glenn say about the situation?”, elicited from sentence 2. The advent of large language models (LLMs) makes viewing discourse in a question generation and answering fashion increasingly

From Wu et al. (2023).

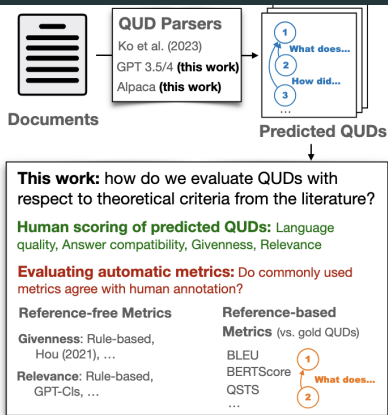


Figure 1: An overview of this work. Given QUD parser outputs, we (a) design a protocol for evaluation; (b) collect human judgments using this protocol; (c) evaluate automatic evaluation metrics. This work additionally contributes new LLM-based QUD parsers.

Example 3: starting paragraph & figure

Given a snapshot observation of an event, people can easily anticipate and reason about unobserved causes and effects in relation to the observed event: what might have happened just before, what might happen next as a result, and how different events are chained through causes and effects. For instance, if we observe an event “X repels Y’s attack” (Figure 1), we can immediately infer various plausible facts surrounding that event. In terms of the *plausible motivations* behind the event, X probably wants to protect herself. As for the *plausible pre-conditions* prior to the event, X may have been trained in self-defense to successfully fend off Y’s attack. We can also infer the *plausible characteristics* of X; she might be strong, skilled, and brave. As a *result* of the event, X probably feels angry and might want to file a police report. Y, on the other hand, might feel scared of getting caught and want to run away.

The examples above illustrate how day-to-day commonsense reasoning can be operationalized through a densely connected collection of inferential knowledge. It is through



Figure 1: A tiny subset of ATOMIC, an atlas of machine commonsense for everyday events, causes, and effects.

From Sap et al. (2019).

Example 4: Starting paragraphs

1 Introduction

Generalisations allow researchers to extrapolate from concrete findings to broader implications (Bunge, 1960; Peters et al., 2022). As a result, they are abundant in scientific communication, including in research articles (DeJesus et al., 2019). However, by obscuring variability, these generalisations may also exaggerate or overgeneralise findings beyond what is actually supported by evidence, leading to potential misinterpretations.

Generalisations can take many forms in language (see Fig 1), with varying semantic interpretations (Haigh et al., 2020; Bowker, 2023). Notably,

From Diao et al. (2026).

1 Introduction

In our daily lives, we make decisions based on what we think will and will not happen. This knowledge about event typicality (i.e., which events are likely or unlikely to happen) helps us navigate the intricacies of human conversation, interpreting meaning beyond what speakers have literally said (Grice, 1975; Bergen et al., 2016).

From Chi et al. (2026).

Example 5: more starting figures



[Although the proposed approach does bring WSD improvements, it is rather incremental. This is probably not a "weakness" per se, just that the paper is not an eye-opener]. [The evaluation is done on German data only, which leaves some doubts about other languages.]

Figure 1: Illustration of lazy thinking in ARR-22 reviews sourced from NLPEER (Dycke et al., 2023). The first review segment belongs to the class ‘The results are not novel.’ The last segment pertains to, ‘The approach is tested only on [not English], so unclear if it will generalize to other languages.’ as per ARR-22 guidelines.

From Purkayastha et al. (2025)



Figure 1: Schematic of our literature review table generation task: (1) synthesize multiple input papers into a table with both (2) a schema (columns) and (3) values. Each row corresponds to an input paper.

From Newman et al. (2024).

Why is the work important? How is it different?

- Contrast with *key* prior works
 - Show that your work is novel, unique
 - What is the gap you are addressing
 - Leave extensive discussion for later section
- Quick summary of why *epistemically rigorous*

What exactly is the work?

- Describe the scope and limitations of the work.
- State any underlying theoretical or methodological assumptions.
 - Methodological & conceptual rigor
- More detailed summary of what you did than the abstract

State your contributions

Do this explicitly

- Why?
 - Consider your audience
 - Reviewers → need to assess these
 - Other readers → want to know quickly what you did
- Isn't this hand-holding?
 - Yes! But it's necessary

Example 6: contribution paragraphs

× *Beware*: being too concise, smashing contributions together

Our contributions are as follows: (1) we develop a new dataset, VAST, for zero-shot and few-shot stance detection and (2) we propose a new model for stance detection that improves performance on a number of challenging linguistic phenomena (e.g., sarcasm) and relies less on sentiment cues (which often lead to errors in stance classification). We make our dataset and models available for use: <https://github.com/>

From [Allaway and McKeown \(2020\)](#).

Our key contributions are:

1. Three AI review systems: (i) OpenReviewer for automatic peer review with LLMs; (ii) Papers with Reviews an online paper review platform; and (iii) Reviewer Arena for evaluating reviewers by preferences.
2. Four evaluation methods: (i) Human evaluation; (ii) Automatic LLM evaluation; (iii) Automatic LLM prediction of human preferences; and (iv) Automatic discovery of LLM review limitations, using synthesis and analysis to map errors and shortcomings in LLM based reviewing.
3. Role playing: Dialogue between LLMs playing different roles in the review process.
4. User feedback: Evaluating quality and trustworthiness.

From [Tyser et al. \(2024\)](#).

Example 7: contribution paragraphs

✓ Good explanation of experiments & findings

Contributions. We make the following contributions: **(1)** Introduce LAZYREVIEW, a dataset annotated with *lazy thinking* classes for a new task in model development and evaluation. **(2)** Enhance annotation guidelines to improve both automated and human annotations. **(3)** Demonstrate that positive examples boost annotation quality and in-context learning. **(4)** Show that instruction tuning on our dataset enhances model performance. **(5)** Find that annotated *lazy thinking* classes improve review quality.

From [Purkayastha et al. \(2025\)](#).

Our contributions are: **(i)** We create two new large-scale, high-quality corpora for studying event typicality, EventTyp and EventTypGold; **(ii)** We conduct large-scale human experiments and find evidence for three CUES that can lead to inferences about event typicality; and **(iii)** We conduct extensive experiments with LLMs and show that LLMs often lack event commonsense knowledge, which undermines their ability to make event-typicality inferences.²

From [Chi et al. \(2026\)](#).