Researching Responsible and Trustworthy Natural Language Processing

Scientific Writing in NLP

Emily Allaway
(Adapted from: Frank Keller)

29 September 2024

School of Informatics
University of Edinburgh
eallaway@ed.ac.uk

Overview

General Principles of Scientific Writing

Anatomy of an (NLP) Paper

Publishing in *CL

Please also look at Alley's web site, which has a lot of videos and additional materials: https://www.craftofscientificwriting.org/

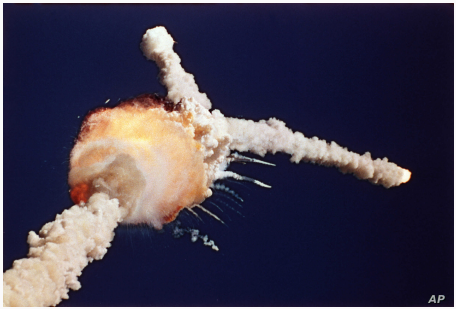
Overview

Writing Matters: Example from Alley (2018)

On January 28, 1986, the Space Shuttle *Challenger* took off from Cape Canaveral, Florida. On board were seven astronauts, including teacher Christa McAuliffe, the first civilian in space. Millions of school children across the US watched.

Writing Matters

The Shuttle exploded 73 seconds after takeoff, killing all seven astronauts on board.



A subsequent investigation found that the solid rocket boosters were the source of the explosion. Two O-rings, seals to prevent the fuel from escaping, had failed.

Writing Matters

Long before the fatal launch, engineers had reservations about the design of the O-rings on the boosters. NASA management requested they seek opinions from O-ring experts.

So NASA engineers visited two manufacturers of O-rings. They found that both manufacturers had serious concerns about the design of the O-rings.

The engineers wrote up a report about their visits. It contained strong warnings about the design of the O-rings and was entitled:

> *Subject: Visit to Precision Rubber Products Corporation and Parker Seal Company*

But no one responded to the report. NASA's paper trail ends here.

Exercise 1

The title of the reports on the Shuttle O-rings was:

> *Subject: Visit to Precision Rubber Products Corporation and Parker Seal Company*

Can you see anything wrong with this title?

Exercise 1

The title of the reports on the Shuttle O-rings was:

> *Subject: Visit to Precision Rubber Products Corporation and Parker Seal Company*

Can you see anything wrong with this title?

This is a weak title:

- It does not tell you what the report is about.
- The authors clearly haven't thought about their audience.

To write successfully, you need to understand your audience.

Exercise 1

To come up with a better title, let's think about the audience of this report:

- Who is the audience?
- Why is the audience reading?
- What does the audience know?

Based on this, what title would you suggest?

Lesson Plan

Audience, Purpose, Occasion

Before you write a scientific document, analyze:

- audience
- purpose
- occasion

These will greatly influence how you will write the document.

We will look at each aspect in turn.

Audience: Who

Who is the audience?

- conference paper: audience has very similar background to yourself; experts in your area
- journal article: audience is typically broader, and depending on the journal may include generalists
- grant proposal: mixture of experts (reviewers) and generalists (panel members)
- podcast: general audience with an interest in scientific issues

The broader or more mixed the audience is, the harder the document will be to write.

Audience: Why

Why is the audience reading?

- Once you know who your audience will be, ask what they want to get out of the document.
- Make sure this information is there, is detailed enough, and is structured so as to be easy to find and digest.
- For a grant proposal, look at the review form to see what the reviewers will look for; for a journal paper, look at other papers published in the same journal, etc.

Audience: Knowledge

What does the audience know? Thinking about this will tell you:

- how to arrange the content
- which terms to define
- what background to include

This tells you how to structure your document. Particularly hard if you have a mixed audience!

Think about your **primary** audience; maybe put content for your **secondary** audience in an appendix.

Lesson Plan

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.

Alley's analogy: 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: Inform

Image: Alley (2018)

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.

Lesson Plan

Occasion

The occasion for which you're writing the document determines its:

- form
- formality

Occasion: Form

Form refers to style and grammar, but also length and format of the document. We will discuss this later in the course.

Alley provides advice on grammar, punctuation, and usage in *Appendices A–C* of his book. Useful for both native and non-native speakers!

Important not to be prescriptive. Alley gives *advice,* not based on right/wrong, but on unsettles/distracts readers.

Note differences of British and American spelling. Important not which one you use, but to be consistent.

Occasion: Form

The *length of your document* has an obvious effect on how you write: a conference paper has half a page of literature review, a PhD thesis has a whole chapter.

The format of your document is often fixed (style file of conference or journal, thesis template). But Alley's *Appendix D* has general recommendations for formatting scientific documents ("for situations in which no graphic designer is available").

Occasion: Formality

A certain level of formality is expected in scientific writing. Examples:

Too informal	Accepted
a lot	*much* or *many*
get	*obtain*
contractions (*don't*)	written out (*do not*)
And ...	*Also, ...*
But ...	*However, ...*

Normally, don't address the reader with *you* (exception: instructions).

But what about writing for a general audience (like a twelve year old)?

Lesson Plan

Choose the Right Word

- Use the right technical terms: you wouldn't say *weight* when you mean *mass*.
- But everyday words have a precise meaning too (don't use fancy words like *plethora* unless you're sure what they mean).
- Take care with easily confused words such as *continuously* and *continually*.
- Alley has a whole list of them, Appendix D.

Synonyms

In creative writing, the use of synonyms is encouraged (they keep your prose interesting).

In scientific writing, synonyms are mostly not a good thing.

For example, *development set* and *validation set* are synonyms, but stick to one to avoid confusion. (The reader may wonder whether you are using two different sets to tune your model.)

Also, there's many *near*-synonyms, which can also cause confusion. For example *image descriptions* and *image captions* are closely related, but not exactly the same.

Don't hesitate to repeat a word if it's the right word!

Connotations, Exaggerations

Avoid words with negative connotations, e.g., *cheap*, *obvious*.

Avoid exaggerations, e.g., *countless activities*, *a thorough literature search*.

Be careful with words such as *prove*, *optimal*, and *significant*, which have precise meaning in most scientific fields.

Lesson Plan

Needless Complexity

Avoiding needless complexity is the most important advice to scientific writers (according to Alley). Avoid needlessly complex:

- paragraphs
- words
- phrases
- sentences

Consider the following paragraph, written by Niels Bohr (Nobel Prize in Physics, 1922).

Complex Paragraphs

The Correspondence Principle. So far as the principles of the quantum theory are concerned, the point which has been emphasized hitherto is the radical departure from our usual conceptions of mechanical and electrodynamical phenomena. As I have attempted to show in recent years, it appears possible, however, to adopt a point of view which suggests that the quantum theory may, nevertheless, be regarded as a rational generalization of ordinary conceptions. As may be seen from the postulates of the quantum theory, and particularly the frequency relation, a direct connection between the spectra and the motion of the kind required by the classical dynamics is excluded but at the same time, the form of these postulates leads us to another relation of a remarkable nature.

Complex Paragraphs

The Correspondence Principle. So far as the principles of the quantum theory are concerned, the point which has been emphasized hitherto is the radical departure from our usual conceptions of mechanical and electrodynamical phenomena. As I have attempted to show in recent years, it appears possible, however, to adopt a point of view which suggests that the quantum theory may, nevertheless, be regarded as a rational generalization of ordinary conceptions. As may be seen from the postulates of the quantum theory, and particularly the frequency relation, a direct connection between the spectra and the motion of the kind required by the classical dynamics is excluded but at the same time, the form of these postulates leads us to another relation of a remarkable nature.

Complex words

Complex Paragraphs

The Correspondence Principle. So far as the principles of the quantum theory are concerned, the point which has been emphasized hitherto is the radical departure from our usual conceptions of mechanical and electrodynamical phenomena. As I have attempted to show in recent years, it appears possible, however, to adopt a point of view which suggests that the quantum theory may, nevertheless, be regarded as a rational generalization of ordinary conceptions. As may be seen from the postulates of the quantum theory, and particularly the frequency relation, a direct connection between the spectra and the motion of the kind required by the classical dynamics is excluded but at the same time, the form of these postulates leads us to another relation of a remarkable nature.

Complex words

Complex sentences: on average 40 words per sentence

Complex Paragraphs

The Correspondence Principle. Many people have stated that the quantum theory is a radical departure from classical mechanics and electrodynamics. However, the quantum theory may be regarded as nothing more than a rational extension of classical concepts. Although no direct connection exists between quantum theory and classical dynamics, the form of the quantum theory's postulates, particularly the frequency relation, leads us to another kind of relation, one that is remarkable.

This revised version is shorter and less complex.

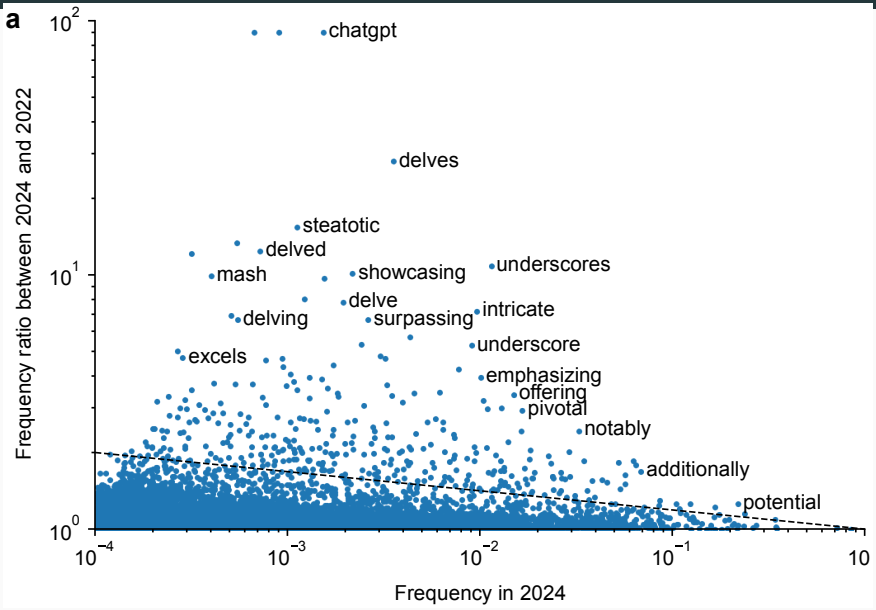
Complex Words

Avoid words that are long and infrequent, but don't add precision and clarity, e.g.:

- *elucidate:* use *show*, *reveal* instead
- many *-ize* words: *prioritize* or *utilize*; use *rank* and *use* instead
- some *-ize* words have precise meaning: *minimize* or *maximize*

Individual word substitutions may not make a difference, but overall, the effect can be substantial.

Excess Vocabulary in ChatGPT: Kobak et al. (2024)



Complex Words

Consider an excerpt from Gagné and Soulie-Fogelman (2020):

In the foreseeable future courtesy AI economies will start reaping rich benefits because of cost advantages in labor and time. AI will penetrate more broadly because of the ML (Machine Learning) processes, wherein systems progressively learn and improve their performance over time. Thus, government and the private sector need to actively support innovation and adoption, in ways that support equitable growth. However, AI businesses are exhibiting unique challenges, in part related to intense competition and potentially lower margins in AI than in some legacy IT sectors.

Complex Words

Consider an excerpt from Gagné and Soulie-Fogelman (2020):

In the foreseeable future courtesy AI economies will start reaping rich benefits because of cost advantages in labor and time. AI will penetrate more broadly because of the ML (Machine Learning) processes, wherein systems progressively learn and improve their performance over time. Thus, government and the private sector need to actively support innovation and adoption, in ways that support equitable growth. However, AI businesses are exhibiting unique challenges, in part related to intense competition and potentially lower margins in AI than in some legacy IT sectors.

Complex Words

Other sources of complexity that should be used sparingly or avoided:

- abbreviations: use as sparingly as possible
- all caps for names: avoid if possible
- slashed terms: replace by a single, better term, or a conjunction

Complex Phrases

The excerpt from Gagné and Soulie-Fogelman (2020) also contains complex phrases:

In the foreseeable future courtesy AI economies will start reaping rich benefits because of cost advantages in labor and time. AI will penetrate more broadly because of the ML (Machine Learning) processes, wherein systems progressively learn and improve their performance over time. Thus, government and the private sector need to actively support innovation and adoption, in ways that support equitable growth. However, AI businesses are exhibiting unique challenges, in part related to intense competition and potentially lower margins in AI than in some legacy IT sectors.

Complex Phrases

The excerpt from Gagné and Soulie-Fogelman (2020) also contains complex phrases:

In the foreseeable future courtesy AI economies will start reaping rich benefits because of cost advantages in labor and time. AI will penetrate more broadly because of the ML (Machine Learning) processes, wherein systems progressively learn and improve their performance over time. Thus, government and the private sector need to actively support innovation and adoption, in ways that support equitable growth. However, AI businesses are exhibiting unique challenges, in part related to intense competition and potentially lower margins in AI than in some legacy IT sectors.

Complex Sentences

And another example from Gagné and Soulie-Fogelman (2020):

The Global Partnership on AI (GPAI) was created as an international and multistakeholder initiative with the mandate to guide the responsible development and use of AI in a way that is consistent with human rights, fundamental freedoms, and shared democratic values, as reflected in the OECD Principles on Artificial Intelligence.

This sentence is long (50 words) and it tries to communicate multiple ideas at once.

Instead, try to use short sentences (in the teens). And express *one idea per sentence.*

Complex Sentences

Also, the sentence contains nine prepositional phrases and four conjunctions. This makes it hard to for the reader to figure out when the sentence will end.

Rewritten version with two, simpler sentences:

The Global Partnership on AI (GPAI) is an international initiative involving multiple stakeholders. It aims to guide the development and use of AI in a way that respects human rights, fundamental freedoms, and democratic values, in line with the OECD Principles on Artificial Intelligence.

Exercise 2: Bender et al. (2021)

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 

Emily M. Bender*
ebender@uw.edu
University of Washington
Seattle, WA, USA

Timnit Gebru*
timnit@blackinai.org
Black in AI
Palo Alto, CA, USA

Angelina McMillan-Major
aymm@uw.edu
University of Washington
Seattle, WA, USA

Shmargaret Shmitchell
shmargaret.shmitchell@gmail.com
The Aether

ABSTRACT

The past 3 years of work in NLP have been characterized by the development and deployment of ever larger language models, especially for English. BERT, its variants, GPT-2/3, and others, most recently Switch-C, have pushed the boundaries of the possible both through architectural innovations and through sheer size. Using these pretrained models and the methodology of fine-tuning them

alone, we have seen the emergence of BERT and its variants [39, 70, 74, 113, 146], GPT-2 [106], T-NLG [112], GPT-3 [25], and most recently Switch-C [43], with institutions seemingly competing to produce ever larger LMs. While investigating properties of LMs and how they change with size holds scientific interest, and large LMs have shown improvements on various tasks (§2), we ask whether enough thought has been put into the potential risks associated with developing them and strategies to mitigate these risks

Exercise 2: Twelve year old child

- Why is the audience reading?
- What does the audience know?
- What purpose do you want to achieve?

Exercise 2: CEO of a tech startup

- Why is the audience reading?
- What does the audience know?
- What purpose do you want to achieve?

Exercise 3

Now let's discuss a real-life NLP paper:

Attention Is All You Need

This is the title of Vaswani et al. (2017), the paper that introduced the transformer architecture.

It's one of the most famous papers in the NLP literature and currently has 195,580 citations on Google Scholar.

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.0 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature.

Exercise 3

Let's look at some text from the intro of Vaswani et al. (2017) (next page):

- Does it use complex words, phrases, sentences?
- What about synonyms, exaggerations, abbreviations, needlessly complex verbs?
- What is the overall balance of precision and clarity?

Would you re-write these paragraphs? How?

Exercise 3

Recurrent neural networks, long short-term memory [13] and gated recurrent [7] neural networks in particular, have been firmly established as state of the art approaches in sequence modeling and transduction problems such as language modeling and machine translation [35, 2, 5]. Numerous efforts have since continued to push the boundaries of recurrent language models and encoder-decoder architectures [38, 24, 15].

Recurrent models typically factor computation along the symbol positions of the input and output sequences. Aligning the positions to steps in computation time, they generate a sequence of hidden states h_t, as a function of the previous hidden state h_{t-1} and the input for position t. This inherently sequential nature precludes parallelization within training examples, which becomes critical at longer sequence lengths, as memory constraints limit batching across examples.

Lesson Plan

Graphs, Diagrams, and Tables

Visual material is an important part of a paper:

- diagrams illustrate complex ideas, processes, or models
- graphs show trends or relationships in data
- tables present results or regularities in data
- textual panels present algorithms or mathematical formulas

Such materials attracts the reader's attention; some readers will only look at figures and tables, they will not read (all of) the text.

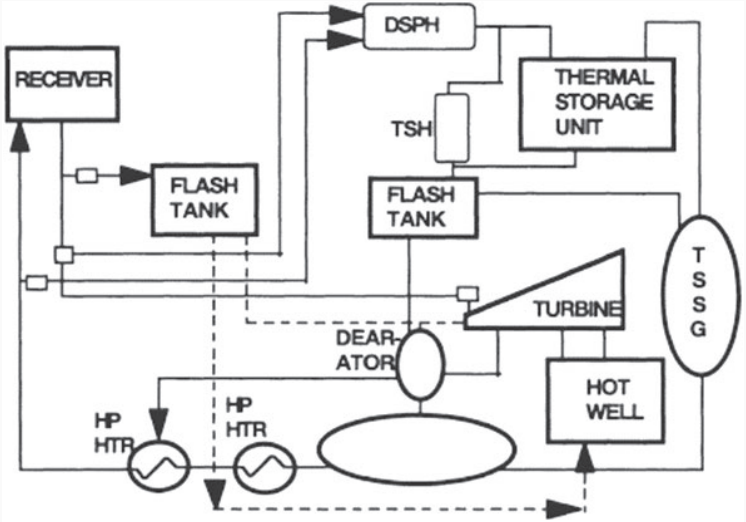


Figure 2-1. Thermal storage system.

Illustrations

Balance precision and clarity also in your illustrations:

- Don't use figures that are more complex than the text used to explain them.
- Use figures to illustrate the most important aspects of what you want to explain, leave out unnecessary details.
- When a figure provides information that's not in the text, it needs to be explained (or be self-explanatory).

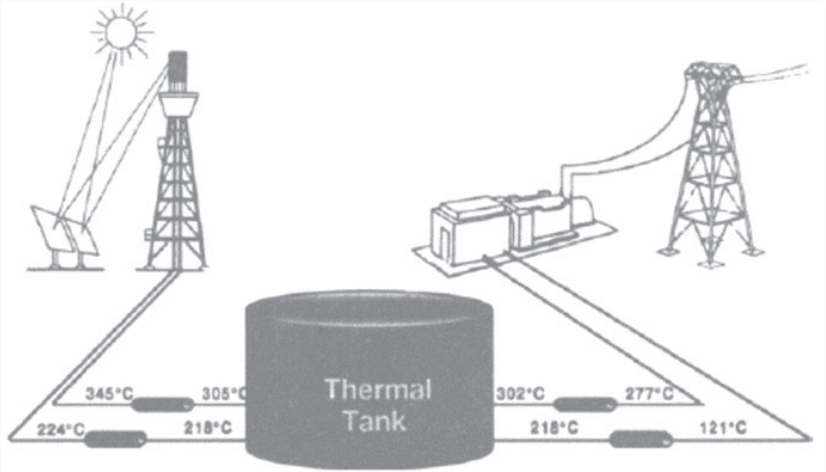


Figure 2-2. Thermal storage system. This storage system takes excess energy from the solar receiver and stores it for later use when the sun is no longer providing solar radiation to the mirrors.

Captions

Every figure needs a caption:

- Reader are automatically drawn to figures, and will try to understand them, often before reading the main text.
- The caption needs to contain everything that's required to understand the figure.
- Start with a phrase that identifies the illustration; formulate it using the same consideration as for document titles.
- Then explain what the figure shows in more detail, expand any abbreviations, label all the parts, etc.

Captions

- captions should fully describe the major elements of a figure or table
- together with its caption, the figure or table should be self-contained, i.e., understandable without referring to the text
- captions should assist a reader who's only skimming the paper, or who is going back to re-read parts of a longer paper
- normally, the caption appears above a table, but below a figure
- if you use abbreviations or symbols in a figure or table, then these need to be explained in the caption
- the caption can also contain additional detail that would interrupt the flow of the main text

Graphs

- graphs can make behaviors and trends obvious that a hard to discern from a table
- keep graphs simple, avoid both clutter and unnecessary whitespace
- for elements such as secondary ticks, legends, gridlines, boxes, ask if you really need them
- use the same fonts in graphs and tables as in the main text
- sometimes logarithmic axes are appropriate
- a table of results can often be represented as a bar graph
- if you use multiple graphs to display the same quantity, use the same axis (same range) in all of them

Graphs

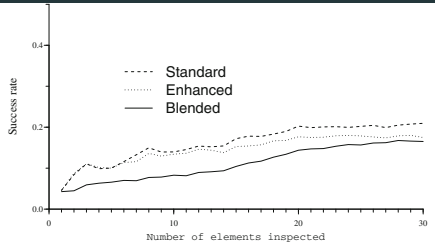


FIGURE 7. *Success rate as the number of inspected items is increased. It is clear that blending is not effective.*

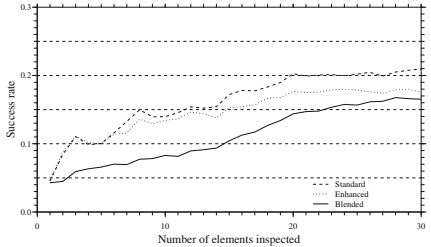


FIGURE 7. *Success rate as the number of inspected items is increased. It is clear that blending is not effective.*

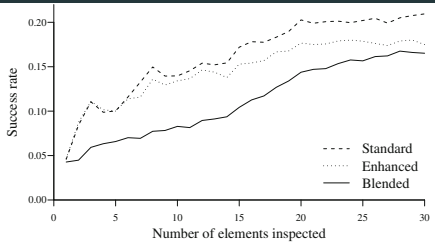


FIGURE 7. *Success rate as the number of inspected items is increased. It is clear that blending is not effective.*

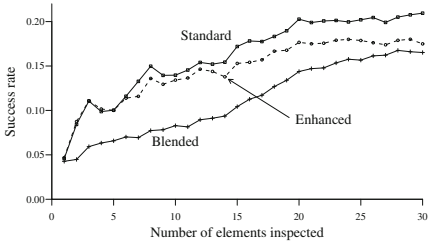


FIGURE 7. *Success rate as the number of inspected items is increased. It is clear that blending is not effective.*

Graphs

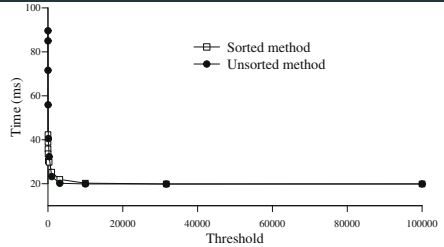


FIGURE 6. *Evaluation time (in milliseconds) for bulk insertion methods as threshold is varied.*

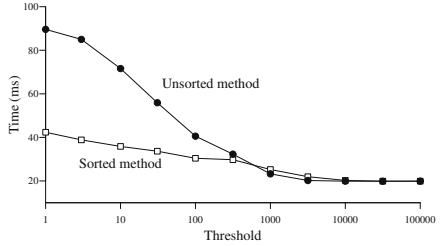


FIGURE 6. *Evaluation time (in milliseconds) for bulk insertion methods as threshold is varied.*

Graphs

Data set	Method	
	A	B
Small, random	11.5	11.6
Large, random	27.9	17.1
Small, clustered	9.7	8.2
Large, clustered	24.0	13.5
All documents	49.4	60.1
First 1000	21.1	35.4
Last 1000	1.0	5.5

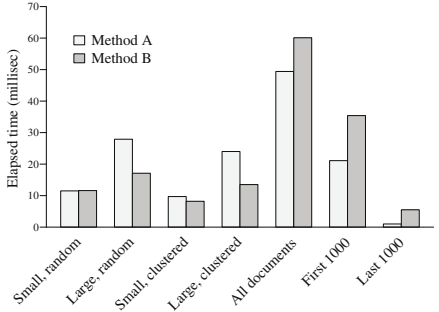


FIGURE 2. *Elapsed time (milliseconds) for methods A and B applied to data sets 1–7.*

Diagrams

- diagrams show architectures, structures, processes, relationships, or states
- typically, the diagram should just show *one* of the things; an attempt to combine them often makes the diagram less clear
- it's a good idea to sketch the diagram by hand first, check layout, proportions, use of space, sizes of elements
- focus on the concept being illustrated, avoid clutter and unnecessary detail
- use pictorial elements consistently (arrows or boxes of the same kind always have the same meaning, etc.)
- don't expect to get it right first time, revise your diagrams as you would revise your text

Diagrams

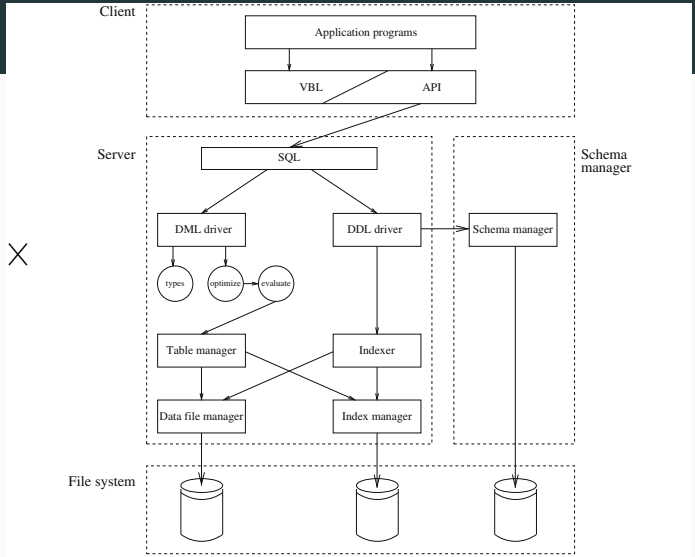


FIGURE 1.3. *System architecture, showing the relationship between the major components. Each component is an independent process. Note the lack of a single interface to the file system.*

Diagrams

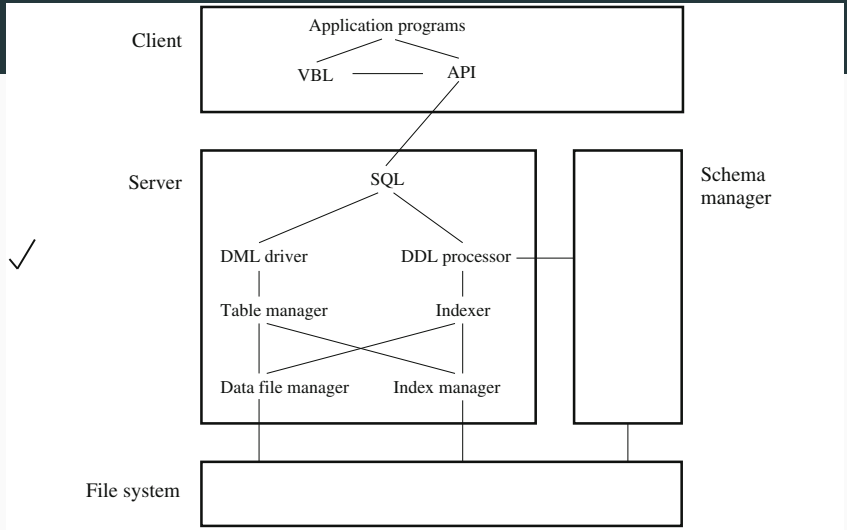


FIGURE 1.3. *System architecture, showing the relationship between the major components. Each component is an independent process. Note the lack of a single interface to the file system.*

Diagrams

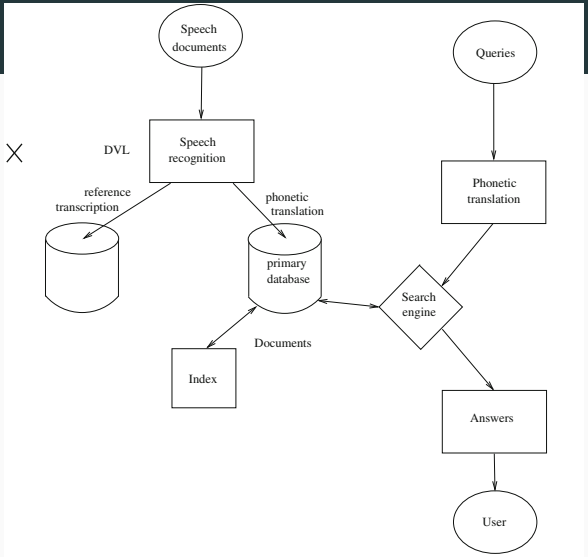


FIGURE 7. *The* QUIRK *system for matching written queries to speech. Each input document is translated into a string of phonemes and then stored. Queries are also translated into phonemes, which can be matched to the documents. Answers are returned to the user.*

Diagrams

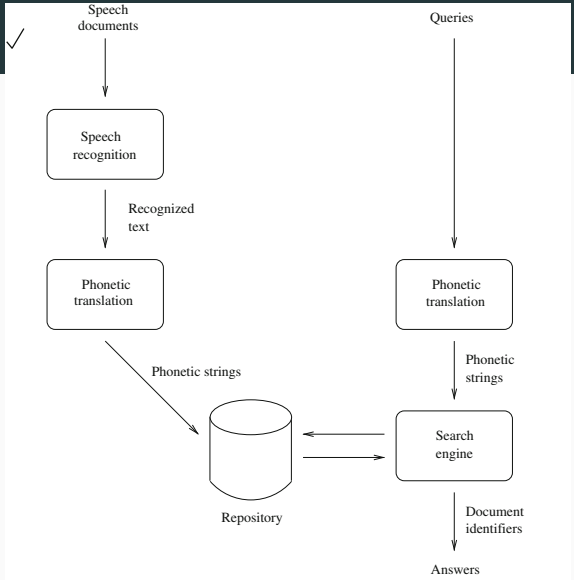


FIGURE 7. *The* QUIRK *system for matching written queries to speech. Each input document is translated into a string of phonemes and then stored. Queries are also translated into phonemes, which can be matched to the documents.*

Tables

- some information cannot be presented easily in graphs or diagrams
- in some cases, the exact numeric values are important
- tables are more suitable than graphs if only a small number of values need to be displayed
- tables can have a hierarchical structure: columns and rows can be partitioned or have internal structure
- the structure needs to be indicated by headings, labels, dividers
- limit the use of horizontal rules; vertical rules should be avoided; tables should contain sufficient whitespace
- don't make a table too big; instead, use two tables or a graph

TABLE 6. *Statistics of text collections used in experiments.*

STATISTICS	SMALL	LARGE
Characters	18,621	1,231,109
Words	2,060	173,145
After stopping	1,200	98,234
Index size	1.31 Kb	109.0 Kb

TABLE 6. *Statistics of text collections used in experiments.*

	Collection	
	Small	Large
File size (Kb)	18.2	1,202.3
Index size (Kb)	1.3	109.0
Number of words	2,060	173,145
— after stopping	1,200	98,234

TABLE 11. *Resources used during compression and indexing. Only the vocabulary is constructed in the first pass; the other structures are built in the second pass.*

Pass	Output	Size Mb	Size %	CPU Hr:Min	Mem Mb
Pass 1:					
Compression	Model	4.2	0.2	2:37	25.6
Inversion	Vocabulary	6.4	0.3	3:02	18.7
Overhead				0:19	2.5
Total		10.6	0.5	5:58	46.8
Pass 2:					
Compression	Text	605.1	29.4	3:27	25.6
	Doc. map	2.8	0.1		
Inversion	Index	132.2	6.4	5:25	162.1
	Index map	2.1	0.1		
	Doc. lens	2.8	0.1		
	Approx. lens	0.7	0.0		
Overhead				0:23	2.5
Total		745.8	36.3	9:15	190.2
Overall		756.4	36.8	15:13	190.2

TABLE 11. *Resources used during compression and indexing. Only the vocabulary is constructed in the first pass; the other structures are built in the second pass.*

Task	Size (Mb)	CPU (Hr:Min)	Memory (Mb)
Pass 1:			
Compression	4.2	2:37	25.6
Inversion	6.4	3:02	18.7
Overhead	—	0:19	2.5
Total	10.6	5:58	46.8
Pass 2:			
Compression	607.9	3:27	25.6
Inversion	137.8	5:25	162.1
Overhead	—	0:23	2.5
Total	745.8	9:15	190.2
Overall	756.4	15:13	190.2

Exercise 4

Let's look at an illustration from Vaswani et al. (2017) (next page):

- Does the figure balance clarity and precision?
- Does the caption contain a meaningful title?
- Are figure and caption taken together self-contained?

Exercise 4

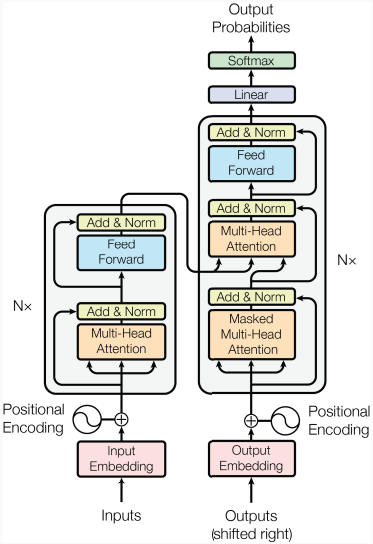


Figure 1: The Transformer - model architecture.

Exercise 4

1. Re-write the title of Vaswani et al. (2017) paper.
2. Write a proper caption for their Figure 1.

Exercise 5

Let's return to Lee et al. (2024) and Chen et al. (2024), the two papers on explainable multimodal NLP that we look at last time.

The following page show to examples of diagrams from these papers.

- Are the diagrams well-designed?
- Does they have the right level of complexity?
- Are the captions appropriate?

How would you modify the diagrams and captions to improve them?

Exercise 5: Lee et al. (2024)

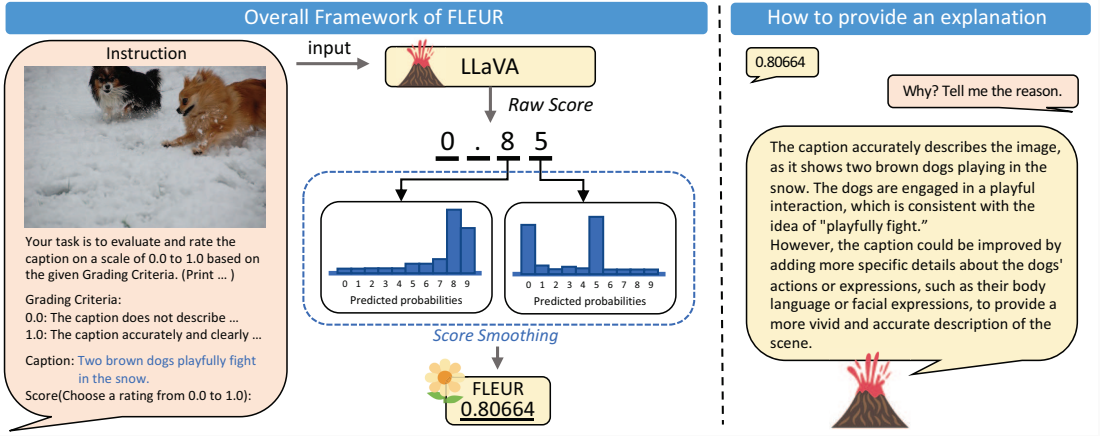


Figure 2: The overall framework of FLEUR. **Left**: When feeding LLaVA with the prompt containing the grading criteria, image, and the candidate caption for evaluation, FLEUR takes a weighted sum of probabilities of tokens (0 to 9) as the final score. **Right**: When prompted by the user for the rationale behind the given score, FLEUR provides explanations in a language understandable to humans.

Exercise 5: Chen et al. (2024)

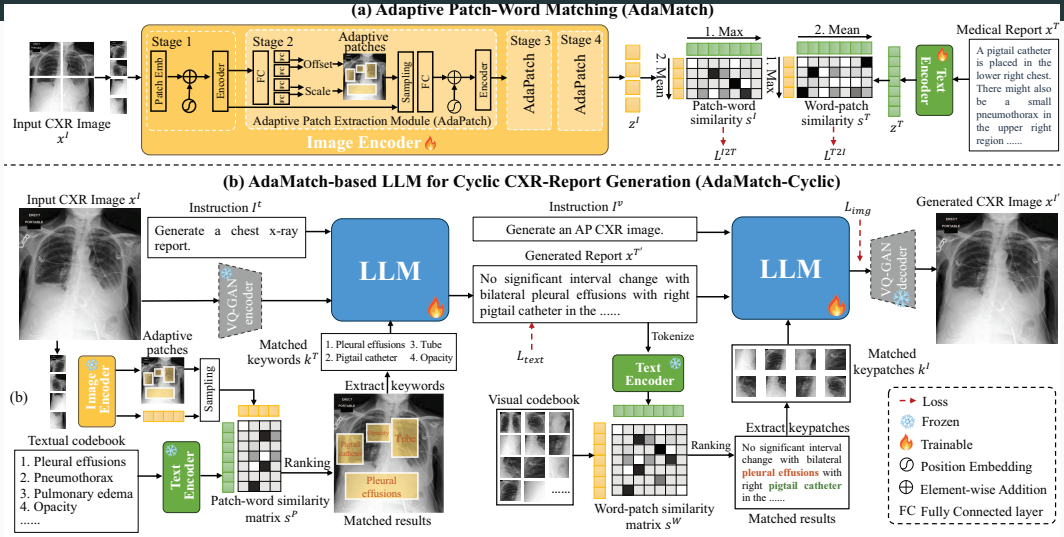


Figure 2: The overview of the proposed methods. (a) Adaptive patch-word Matching (AdaMatch) model. (b) AdaMatch-based bidirectional large language model (LLM) for cyclic CXR-report generation (AdaMatch-Cyclic).

Exercise 6

Here are some tables from Lee et al. (2024) and Chen et al. (2024).

- Is the table layout good? How about the use of whitespace?
- Can you decode the hierarchical structure of these tables?
- Should they maybe have used a graph instead?
- Are the captions appropriate?

How would you modify the tables and captions to improve them?

Exercise 6: Lee et al. (2024)

Type	Exp	Metric	Flickr8k		COM	Pascal-50S (Accuracy ↑)				
			EX (τ_c ↑)	CF (τ_b ↑)	(τ_c ↑)	HC	HI	HM	MM	Avg
reference -based		BLEU-4	30.8	16.9	30.6	53.0	92.4	86.7	59.4	72.9
		ROUGE-L	32.3	19.9	32.4	51.5	94.5	92.5	57.7	74.1
		METEOR	41.8	22.2	38.9	56.7	97.6	94.2	63.4	78.0
		CIDEr	43.9	24.6	37.7	53.0	98.0	91.5	64.5	76.8
		SPICE	44.9	24.4	40.3	52.6	93.9	83.6	48.1	69.6
		BERTScore	39.2	22.8	30.1	65.4	96.2	93.3	61.4	79.1
	✓	CLAIR[4]	48.3	–	61.0	52.4	99.5	89.8	73.0	78.7
		TIGEr	49.3	–	45.4	56.0	**99.8**	92.8	74.2	80.7
		ViLBERTScore-F	50.1	–	52.4	49.9	99.6	93.1	75.8	79.6
		RefCLIPScore	53.0	36.4	55.4	64.5	99.6	95.4	72.8	83.1
		RefPAC-S	55.9	37.6	57.3	67.7	99.6	96.0	75.6	84.7
		Polos	**56.4**	37.8	57.6	**70.0**	99.6	97.4	**79.0**	**86.5**
	✓	RefFLEUR (Ours)	51.9	**38.8**	**64.2**	68.0	**99.8**	**98.0**	76.1	85.5
reference -free		CLIPScore	51.2	34.4	53.8	56.5	99.3	96.4	70.4	80.7
		PAC-S	54.3	36.0	55.7	60.6	99.3	96.9	72.9	82.4
		InfoMetIC+[5]	**55.5**	36.6	59.3	–	–	–	–	–
	✓	FLEUR (Ours)	53.0	**38.6**	**63.5**	**61.3**	**99.7**	**97.6**	**74.2**	**83.2**

Table 1: Overall correlation and accuracy comparison with human judgment on Flickr8k-Expert (Flickr8k-EX), Flickr8k-CF, COMPOSITE (COM), and Pascal-50S datasets. Bold indicates the best result in each type. 'Exp' stands for 'explainable' and checkmarks are applied only to the corresponding metrics. FLEUR is the only metric satisfying both explainable and reference-free. All results except for ours are reported results from prior works.

Exercise 6: Chen et al. (2024)

Table 1: Comparison of CXR-to-report generation performance on the MIMIC-CXR and the OpenI datasets.

	MIMIC-CXR						OpenI					
Methods	B-1	B-2	B-3	B-4	M	R-L	B-1	B-2	B-3	B-4	M	R-L
R2Gen	0.3553	0.2232	0.1523	0.1038	0.1412	0.2784	0.3992	0.2407	0.1518	0.0973	0.1390	0.3052
R2GenCMN	0.3719	0.2332	0.1538	0.1053	0.1501	0.2827	0.4091	0.2493	0.1594	0.1045	0.1509	0.3181
Joint-TriNet	0.3585	0.2266	**0.1550**	0.1021	0.1425	0.2788	0.3833	0.2409	0.1598	0.1078	0.1457	0.3293
XProNet	0.3532	0.2212	0.1498	0.1052	0.1415	0.2811	0.4114	0.2502	0.1598	0.1045	0.1457	0.3240
ITHN	0.3623	0.2128	0.1402	0.0992	0.1488	0.2622	0.2661	0.1516	0.0976	0.0663	0.1561	0.2617
M2KT	0.3661	0.2192	0.1465	0.1044	0.1528	0.2673	0.2559	0.1381	0.0819	0.0523	0.1468	0.2439
AdaMatch-Cyclic	**0.3793**	**0.2346**	0.1540	**0.1060**	**0.1625**	**0.2859**	**0.4161**	**0.3002**	**0.2073**	**0.1446**	**0.1621**	**0.3656**

Overview

Organizing Content

Scientific documents are almost always structured as:

- Title
- Abstract
- Introduction
- Middle
- Conclusion

The structure of the middle can vary. In experimental papers, it's typically:

- Method
- Results
- Discussion

There are also extra parts: figures and tables, appendices, etc.

Lesson Plan

Title

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

Title: Examples

The Bluest Eye.

Sula.

Song of Solomon.

Tar Baby.

Beloved.

Jazz.

Paradise.

Love.

A Mercy.

Home.

God Help the Child.

Title: Examples

The Bluest Eye.
Sula.
Song of Solomon.
Tar Baby.
Beloved.
Jazz.
Paradise.
Love.
A Mercy.
Home.
God Help the Child.

Titles of novels of Tony Morrison (Nobel Prize in Literature, 1993).

Title: Examples from ACL 2024

Titles don't have to be phrases, they can be sentences:

Speech language models lack important brain-relevant semantics

This title is a one-sentence summary of the result of the study. In some fields (e.g., medicine), this is the conventional way to write titles.

It's also common to combine a main title and a subtitle, separated by a colon:

Self-Alignment for Factuality: Mitigating Hallucinations in LLMs via Self-Evaluation

Here, the main title introduces the method, and the subtitle the task.

Title: Examples from ACL 2025

GAPO: Learning Preferential Prompt through Generative Adversarial Policy Optimization

FACT-AUDIT: An Adaptive Multi-Agent Framework for Dynamic Fact-Checking Evaluation of Large Language Models

Statistical Deficiency for Task Inclusion Estimation

"Yes, My LoRD." Guiding Language Model Extraction with Locality Reinforced Distillation

Self-Instructed Derived Prompt Generation Meets In-Context Learning: Unlocking New Potential of Black-Box LLMs

Whose Boat Does it Float? Improving Personalization in Preference Tuning via Inferred User Personas

AGrail: A Lifelong Agent Guardrail with Effective and Adaptive Safety Detection

Lesson Plan

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

According to Alley (who calls it summary), the abstract should:

- contain the most important points of the paper
- contain *only* the important points
- only include material that occurs elsewhere in the paper (verbatim or paraphrased)
- be self-contained, i.e., the reader should be able to understand the abstract without having to read anything else
- this means unusual terms, techniques, etc. need to be explained in the abstract
- don't assume all readers will be specialists in the topic of the paper; assume a broad readership.

Content of the Abstract

Alley distinguishes:

- *informative summary:* describes the most important results of a paper;
- *descriptive summary:* states what kind of information the paper provides (like a table of contents), but doesn't give the actual results.

The abstract of a conference or journal paper is a mixture of both: it provides signposting (which information to expect in the paper), but also summarizes the results.

Content of the Abstract

Zobel 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 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.

Organization of the Abstract

My own experience shows:

- longer documents may require longer abstracts: the abstract of a journal paper is somewhat longer than that of a conference paper
- the abstract of a PhD thesis is typically a whole page; it should summarize each (content) chapter
- abstracts can contain sentences extracted from the main body of the text (you may need to edit them for coherence)
- but: it is sometimes a good strategy to write the abstract *before* writing the paper – helps planning the overall argumentation, deciding what to focus on
- and then once the paper is finished, you need to completely re-write the abstract!

Lesson Plan

Introduction

The introduction prepares the reader for the main content of the document, answering the following questions:

- What exactly is the work?
- Why is the work important?
- What is needed to understand the work?
- How will the work be presented?

Not all question may be present, and not always in this order.

What exactly is the work?

- Describe the scope and limitations of the work.
- Provide more detail than the abstract.
- State any underlying theoretical or methodological assumptions.

Why is the work important?

- Give the audience a reason to starting reading, and continue reading, the document.
- If the document is a proposal: why should this be funded?
- 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!
- Bear in mind who your *audience* is: experts, general readers, funders, managers.
- This determines how you need to justify the importance of your work.

Why is the work important?

In size, density, and composition, Ganymede and Callisto (Jupiter's two largest moons) are near twins: rock-loaded snowballs. These moons are about 5000 km in diameter and contain 75 percent water by volume. The one observable difference between them is their albedo: Callisto is dark all over, while Ganymede has dark patches separated by broad light streaks. This paper discusses how these two similar moons evolved so differently.

Here curiosity is the main motivation for the work.

What is needed to understand the work?

Present the background required to understand the main part of your document:

- review literature and related work
- show that your work is novel, unique
- identify gaps in the literature (respectfully!)
- tailor this to your audience and what they know
- boost your credibility as an author

Can be a separate section or chapter, or just a part of the intro.

In a thesis, you can have both an upfront background chapter and a background section for each content chapter.

What is needed to understand the work?

A table is sometimes a good way to present the related literature:

Dataset	Task	#L	#V	Obj	Imgs	Sen	Des	Cln	ML	Resource	Example Labels
Ikizler (Ikizler et al., 2008)	AC	6	6	0	467	N	N	Y	N	–	running, walking
Sports Dataset (Gupta et al., 2009)	AC	6	6	4	300	N	N	Y	N	–	tennis serve, cricket bowling
Willow (Delaitre et al., 2010)	AC	7	6	5	986	N	N	Y	Y	–	riding bike, photographing
PPMI (Yao and Fei-Fei, 2010)	AC	24	2	12	4.8k	N	N	Y	N	–	play guitar, hold violin
Stanford 40 Actions (Yao et al., 2011)	AC	40	33	31	9.5k	N	N	Y	N	–	cut vegetables, ride horse
PASCAL 2012 (Everingham et al., 2015)	AC	11	9	6	4.5k	N	N	Y	Y	–	riding bike, riding horse
89 Actions (Le et al., 2013)	AC	89	36	19	2k	N	N	Y	N	–	ride bike, fix bike
MPII Human Pose (Andriluka et al., 2014)	AC	410	–	66	40.5k	N	N	Y	N	–	riding car, hair styling
TUHOI (Le et al., 2014)	HOI	2974	–	189	10.8k	N	N	Y	Y	–	sit on chair, play with dog
COCO-a (Ronchi and Perona, 2015)	HOI	–	140	80	10k	N	Y	Y	Y	VerbNet	walk bike, hold bike
Google Images (Ramanathan et al., 2015)	AC	2880	–	–	102k	N	N	N	N	–	riding horse, riding camel
HICO (Chao et al., 2015)	HOI	600	111	80	47k	Y	N	Y	Y	WordNet	ride#v#1 bike; hold#v#2 bike
VCOCO-SRL (Gupta and Malik, 2015)	VSRL	–	26	48	10k	N	Y	Y	Y	–	verb: hit; instr: bat; obj: ball
imSitu (Yatskar et al., 2016)	VSRL	–	504	11k	126k	Y	N	Y	N	FrameNet WordNet	verb: ride; agent: girl#n#2 vehicle: bike#n#1; place: road#n#2
VerSe (Gella et al., 2016)	VSD	163	90	–	3.5k	Y	Y	Y	N	OntoNotes	ride.v.01, play.v.02
Visual Genome (Krishna et al., 2016)	VRD	42.3k	–	33.8k	108k	N	N	Y	Y	–	man playing frisbee

From Gella and Keller (2017).

Lesson Plan

Middle

The middle of a document presents the work in a logical and persuasive fashion. To achieve this, choose a *strategy* for presenting the material:

- chronological
- spatial
- classification and division
- cause-effect
- comparison-contrast

The strategy you choose must be suitable for the audience. Use it to group the work in sections, to make it more digestible.

Descriptive Headings

Section headings should be descriptive, make the strategy of the document obvious, serve as a roadmap for the reader.

Sections provide *whitespace,* which allows readers:

- pause and reflect on what they've read
- jump to the information that interests them, skip those parts that don't

Don't over-section the text, consider if a paragraph break is better than introducing a subsection.

Descriptive Headings

Stylistic considerations for section headings:

- avoid cryptic one-word titles also for sections
- use parallelism (all subsections are titled using noun phrases, participles, etc)
- Don't use dangling subsections (1, 2, 2.1, 3, 3.1, 3.2, 4)

Look at the **table of content** to see if your sectioning works.

Put `\tableofcontent` at the beginning of your paper (even if the final version won't have a ToC).

Descriptive Headings

Weak Headings	Strong Headings
Introduction	Introduction
Debris Recovered	Completed Work
Cataloguing	Recovering Debris
Interpretation	Cataloguing Debris
Results	Interpreting the Debris
Placement	Preliminary Results of Work
Bomb Makeup	Placement of Bomb
Work to be Done	Construction of Bomb
Interpretation	Future Work

Depth

Depth is the level of detail in your document. Affected by:

- **occasion:** determines length, e.g., conference paper vs. journal article
- **audience:** satisfy the readers interest, anticipate their questions, don't raise questions the document doesn't answer
- **purpose:** if you want to persuade, you will need to discuss advantages and disadvantages, rebut objections, etc.

Adapt paragraph and section length depending on the depth of your document.

Avoid very long or very short paragraphs (exception: instructions).

Lesson Plan

Conclusion

Summarizes the middle and provides a future perspective:

- provide an analysis of the most important results
- analyze the results overall, not individually (do that in the middle, in a Discussion section)
- do not present new evidence or new results here
- provide a future perspective:
  - recommendations that derive from your work
  - future direction of you work
  - re-iterate the scope and limitations of your work (already in the intro)
- ACL conferences now require a separate section on limitations

Conclusion ties together loose ends, provides closure.

Exercise 8: Evaluating an Abstract

Look at the abstracts on the next pages. Investigate the following questions:

1. Can you identify a structure that these abstracts follow? Is Zobel right?
2. Which audience do the abstracts target? Is Alley right (general reader)?
3. Are the abstracts self-contained, i.e., they are understandable independent of the paper?
4. Do they use acronyms, mathematics, abbreviations, citations (Alley and Zobel)?
5. Is enough detail provided, and is all the detail important?

Both abstracts are from papers that appeared at ACL 2024.

Exercise 8: Evaluating an Abstract

Abstract of Lee et al. (2024):

Most existing image captioning evaluation metrics focus on assigning a single numerical score to a caption by comparing it with reference captions. However, these methods do not provide an explanation for the assigned score. Moreover, reference captions are expensive to acquire. In this paper, we propose FLEUR, an explainable reference-free metric to introduce explainability into image captioning evaluation metrics. By leveraging a large multimodal model, FLEUR can evaluate the caption against the image without the need for reference captions, and provide the explanation for the assigned score. We introduce score smoothing to align as closely as possible with human judgment and to be robust to user-defined grading criteria. FLEUR achieves high correlations with human judgment across various image captioning evaluation benchmarks and reaches state-of-the-art results on Flickr8k-CF, COMPOSITE, and Pascal-50S within the domain of reference-free evaluation metrics. Our source code and results are publicly available at: https://github.com/Yebin46/FLEUR.

Exercise 8: Evaluating an Abstract

Abstract of Chen et al. (2024):

Fine-grained vision-language models (VLM) have been widely used for inter-modality local alignment between the predefined fixed patches and textual words. However, in medical analysis, lesions exhibit varying sizes and positions, and using fixed patches may cause incomplete representations of lesions. Moreover, these methods provide explainability by using heatmaps to show the general image areas potentially associated with texts rather than specific regions, making their explanations not explicit and specific enough. To address these issues, we propose a novel Adaptive patch-word Matching (AdaMatch) model to correlate chest X-ray (CXR) image regions with words in medical reports and apply it to CXR-report generation to provide explainability for the generation process. AdaMatch exploits the fine-grained relation between adaptive patches and words to provide explanations of specific image regions with corresponding words. To capture the abnormal regions of varying sizes and positions, we introduce an Adaptive Patch extraction (AdaPatch) module to acquire adaptive patches for these regions adaptively. Aiming to provide explicit explainability for the CXR-report generation task, we propose an AdaMatch-based bidirectional LLM for Cyclic CXR-report generation (AdaMatch-Cyclic). It employs AdaMatch to obtain the keywords for CXR images and 'keypatches' for medical reports as hints to guide CXR-report generation. Extensive experiments on two publicly available CXR datasets validate the effectiveness of our method and its superior performance over existing methods. Source code will be released.

Fine-Grained Image-Text Alignment in Medical Imaging Enables Explainable Cyclic Image-Report Generation

Wenting Chen[1] **Linlin Shen**[2] **Jingyang Lin**[4] **Jiebo Luo**[4]
Xiang Li[5*] **Yixuan Yuan**[2*]
[1]City University of Hong Kong [2]The Chinese University of Hong Kong
[3]Shenzhen University [4]University of Rochester
[5]Massachusetts General Hospital and Harvard Medical School
[1]wentichen7-c@my.cityu.edu.hk [2]yxyuan@ee.cuhk.edu.hk [3]llshen@szu.edu.cn
[4]{jluo@cs, jlin81@ur}.rochester.edu [5]xli60@mgh.harvard.edu

Abstract

Fine-grained vision-language models (VLM) have been widely used for inter-modality local alignment between the predefined fixed patches and textual words. However, in medical analysis, lesions exhibit varying sizes and positions, and using fixed patches may cause incomplete representations of lesions. Moreover, these methods provide explainability by using heatmaps to show the general image areas potentially associated with texts rather than specific regions, making their explanations not explicit and specific enough. To address these issues, we propose a novel Adaptive patch-word Matching (AdaMatch) model to correlate chest X-ray (CXR) image regions with words in medical reports and apply it to CXR-report generation to provide explainability for the generation process. AdaMatch exploits the fine-grained relation between adaptive patches and words to provide explanations of specific image regions with corresponding words. To capture the abnormal regions of varying sizes and positions, we introduce an Adaptive Patch extraction (AdaPatch) module to acquire adaptive patches for these regions adaptively. Aiming to provide explicit explainability for the CXR-report generation task, we propose an AdaMatch-based bidirectional LLM for Cyclic CXR-report generation (AdaMatch-Cyclic). It employs AdaMatch to obtain the keywords for CXR images and 'keypatches' for medical reports as hints to guide CXR-report generation. Extensive experiments on two publicly available CXR datasets validate the effectiveness of our method and its superior performance over existing methods.

1 Introduction

Inter-modality alignment, such as vision and language, has been an important task with growing interests in the field of computer vision, especially with the recent advancement in representation learning (Radford et al., 2021). Technologies like contrastive learning and self-supervised learning have dramatically improved state-of-the-art alignment performance. Recent vision-language models (VLMs) demonstrate two approaches: global contrastive alignment, which integrates images and texts at a global level (Radford et al., 2021; Jia et al., 2021; Jang et al., 2023; Wang et al., 2023; Yang et al., 2022), and local alignment, focusing on detailed connections between visual objects and textual words (Chen et al., 2020a; Li et al., 2020b,a; Zhan et al., 2021; Kim et al., 2021; Yao et al., 2021), as illustrated in Fig. 1.

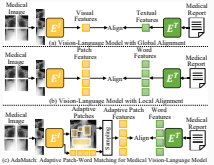


Figure 1: Current vision-language models (VLM) achieve (a) global alignment and (b) local alignment by matching overall visual with textual features, and aligning patches with word features, respectively. (c) To exploit the relation between textual words and abnormal patches with varied sizes, our AdaMatch obtains adaptive patch features and aligns them with word features.

Current VLMs with local alignment either adopt the pre-trained object detector to extract region-of-interest (ROI) features from images and match the corresponding object features with textual words (Chen et al., 2020a; Li et al., 2020b,a; Zhan et al., 2021), or align the visual token from each

*Xiang Li and Yixuan Yuan are corresponding authors.

Exercise 9: Shortening an Abstract

In addition to the problems already discussed, the abstract of Chen et al. (2024) is simply too long!

- It takes up a large chunk of the first page (254 words).
- Not all readers will have the patience for this.
- And it's space you could use elsewhere in the paper! (Remember, abstracts shouldn't contain new information.)

Your task: shorten the abstract to 125 words.

Exercise 9: Shortening an Abstract

Let's ask GPT: "Could you please shorten the following abstract to 125 words: [...]"

Fine-grained vision-language models (VLMs) often align fixed image patches with textual words, but this approach may miss critical details in medical images, where lesions vary in size and position. To improve this, we propose AdaMatch, an Adaptive patch-word Matching model, for correlating regions in chest X-rays (CXRs) with words in medical reports, enhancing explainability. AdaMatch uses adaptive patches, captured by our Adaptive Patch extraction (AdaPatch) module, to represent abnormal regions accurately. For CXR-report generation, we introduce AdaMatch-Cyclic, a bidirectional model that utilizes AdaMatch to identify keywords and "keypatches" as guides. Extensive testing on two public CXR datasets demonstrates AdaMatch's superior performance and effectiveness in generating detailed, interpretable reports. Source code will be made available.

Exercise 10: Evaluating an Introduction

Let's again look at the introduction of Vaswani et al. (2017) (next page):

- What exactly is the work?
- Why is the work important?
- What is needed to understand the work?
- How will the work be presented?

Which of these questions are present in this excerpt?

Exercise 10: Evaluating an Introduction

Recurrent neural networks, long short-term memory [13] and gated recurrent [7] neural networks in particular, have been firmly established as state of the art approaches in sequence modeling and transduction problems such as language modeling and machine translation [35, 2, 5]. Numerous efforts have since continued to push the boundaries of recurrent language models and encoder-decoder architectures [38, 24, 15].

Recurrent models typically factor computation along the symbol positions of the input and output sequences. Aligning the positions to steps in computation time, they generate a sequence of hidden states h_t, as a function of the previous hidden state h_{t-1} and the input for position t. This inherently sequential nature precludes parallelization within training examples, which becomes critical at longer sequence lengths, as memory constraints limit batching across examples.

In this work we propose the Transformer, a model architecture eschewing recurrence and instead relying entirely on an attention mechanism to draw global dependencies between input and output. The Transformer allows for significantly more parallelization and can reach a new state of the art in translation quality after being trained for as little as twelve hours on eight P100 GPUs.

Exercise 11: Evaluating Sectioning

Here's the sectioning of the Vaswani et al. (2017) paper:

1. Introduction
2. Background
3. Model Architecture
   - 3.1 Encoder and Decoder Stacks
   - 3.2 Attention
     - 3.2.1 Scaled Dot-Product Attention
     - 3.2.2 Multi-Head Attention
     - 3.2.3 Applications of Attention in our Model
   - 3.3 Position-wise Feed-Forward Networks
   - 3.4 Embeddings and Softmax
   - 3.5 Positional Encoding
4. Why Self-Attention

Exercise 11: Evaluating Sectioning

5. Training
   - 5.1 Training Data and Batching
   - 5.2 Hardware and Schedule
   - 5.3 Optimizer
   - 5.4 Regularization
6. Results
   - 6.1 Machine Translation
   - 6.2 Model Variation
   - 6.3 English Constituency Parsing
7. Conclusion

Which *strategy* do the authors use to structure the middle? Do they follow Alley's advice on *sectioning?* Are the headings *descriptive?* Which level of *depth* do they us?

Overview

Lesson Plan

Venues and 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 $\rightarrow$ ARR

Steps 4,5 $\rightarrow$ Individual venue

ARR

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

Lesson Plan

General Principles of Scientific Writing

Anatomy of an (NLP) Paper

Publishing in *CL

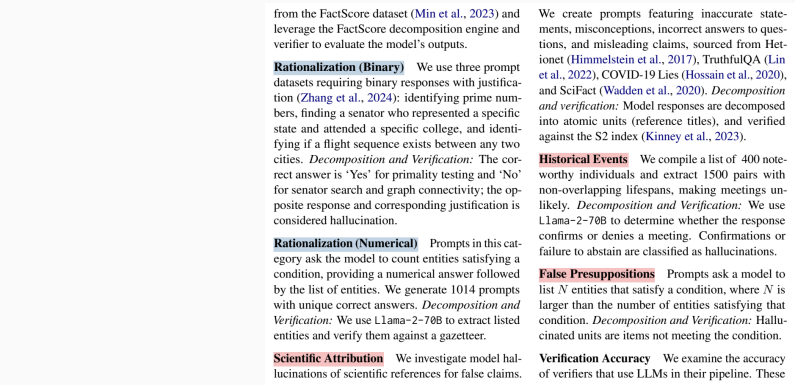
- Publication Process
- Formatting
- More on Content

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 Hetionet (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

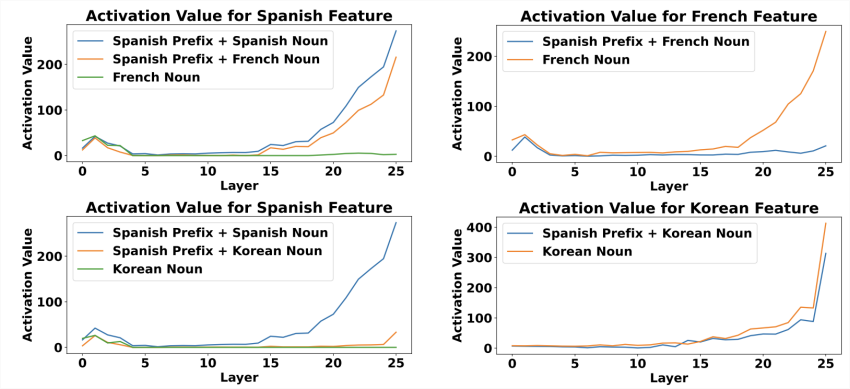
Table 1: Sample item from each task in our evaluation. All items are originally curated by Floyd et al. (In prep).

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.

From Hu et al. (2023).

Be aware of accessibility

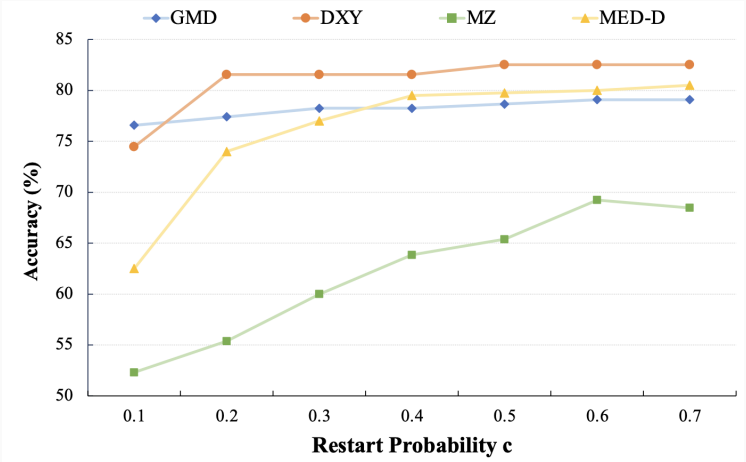
× *Unmarked lines* may be indistinguishable



From Deng et al. (2025).

Be aware of accessibility

✓ Use symbols to differentiate lines



From Yan et al. (2025).

Lesson Plan

General Principles of Scientific Writing

Anatomy of an (NLP) Paper

Publishing in *CL

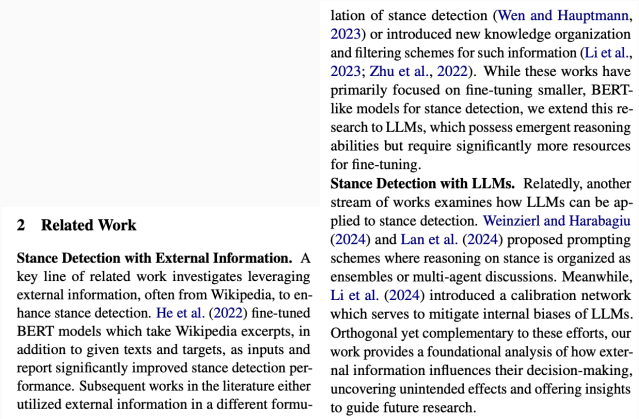
- Publication Process
- Formatting
- More on Content

Common types of NLP papers

- Methods
  - Ex: Method for specific task – Jin et al. (2025)
  - Ex: General method – Dai et al. (2025)
- Dataset
  - Ex: Data for existing task – Zhu et al. (2025a)
  - Ex: New *task* & data – Zhao and Caragea (2025)
- Review/Survey
  - Ex: Summarize literature on a topic – Zhu et al. (2025b)
- Position/Perspective
  - Ex: "The Impossibility of Fair LLMs" – Anthis et al. (2025)

Related Work

× *Beware:* Citing only recent work



2 Related Work

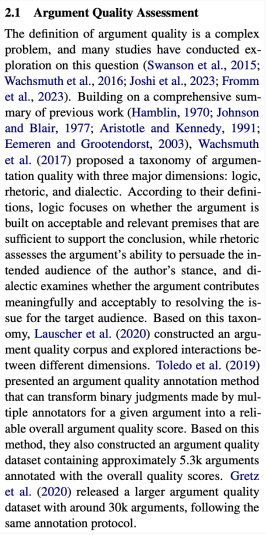
Stance Detection with External Information. A key line of related work investigates leveraging external information, often from Wikipedia, to enhance stance detection. He et al. (2022) fine-tuned BERT models which take Wikipedia excerpts, in addition to given texts and targets, as inputs and report significantly improved stance detection performance. Subsequent works in the literature either utilized external information in a different formulation of stance detection (Wen and Hauptmann, 2023) or introduced new knowledge organization and filtering schemes for such information (Li et al., 2023; Zhu et al., 2022). While these works have primarily focused on fine-tuning smaller, BERT-like models for stance detection, we extend this research to LLMs, which possess emergent reasoning abilities but require significantly more resources for fine-tuning.

Stance Detection with LLMs. Relatedly, another stream of works examines how LLMs can be applied to stance detection. Weinzierl and Harabagiu (2024) and Lan et al. (2024) proposed prompting schemes where reasoning on stance is organized as ensembles or multi-agent discussions. Meanwhile, Li et al. (2024) introduced a calibration network which serves to mitigate internal biases of LLMs. Orthogonal yet complementary to these efforts, our work provides a foundational analysis of how external information influences their decision-making, uncovering unintended effects and offering insights to guide future research.

From Nguyen and Kim (2025).

Related Work

× *Beware:* Too descriptive



2.1 Argument Quality Assessment

The definition of argument quality is a complex problem, and many studies have conducted exploration on this question (Swanson et al., 2015; Wachsmuth et al., 2016; Joshi et al., 2023; Fromm et al., 2023). Building on a comprehensive summary of previous work (Hamblin, 1970; Johnson and Blair, 1977; Aristotle and Kennedy, 1991; Eemeren and Grootendorst, 2003), Wachsmuth et al. (2017) proposed a taxonomy of argumentation quality with three major dimensions: logic, rhetoric, and dialectic. According to their definitions, logic focuses on whether the argument is built on acceptable and relevant premises that are sufficient to support the conclusion, while rhetoric assesses the argument's ability to persuade the intended audience of the author's stance, and dialectic examines whether the argument contributes meaningfully and acceptably to resolving the issue for the target audience. Based on this taxonomy, Lauscher et al. (2020) constructed an argument quality corpus and explored interactions between different dimensions. Toledo et al. (2019) presented an argument quality annotation method that can transform binary judgments made by multiple annotators for a given argument into a reliable overall argument quality score. Based on this method, they also constructed an argument quality dataset containing approximately 5.3k arguments annotated with the overall quality scores. Gretz et al. (2020) released a larger argument quality dataset with around 30k arguments, following the same annotation protocol.

Following these data collection efforts, the computational methods for argument quality assessment have evolved significantly. Marro et al. (2022) used argument structure information derived from graph embeddings to enhance the performance of argument quality assessment. Wang et al. (2023b) leveraged contrastive learning to distinguish arguments of different quality more effectively. Bao

From Jin et al. (2025).

Related Work

✓ Do: Explicitly contrast with prior work

Language Models Pretrained language models (PLMs) such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) have been extensively used in stance detection (Glandt et al., 2021; Allaway and McKeown, 2020; Li et al., 2021). More recently, large language models (LLMs) have been developed, offering the advantage of handling downstream tasks directly through prompting techniques (Le Scao et al., 2023; Touvron et al., 2023; Team et al., 2023; Naveed et al., 2023). While some research on stance detection has utilized LLMs (Gatto et al., 2023; Li et al., 2023a; Fraile-Hernandez and Peñas, 2024), these applications are limited to monolingual scenarios. Existing multilingual stance detection works solely employ PLMs (Vamvas and Sennrich, 2020). In contrast, our study leverages both PLMs and LLMs in the multilingual context, which enables us to explore ZSSD in a more comprehensive setting, enhancing our understanding and capabilities in this area.

From Zhao and Caragea (2025).

Extras

- Limitations*
- Ethical considerations
- Acknowledgments
- Appendices

Limitations

× *Don't:* Surface level

Limitations

Currently, most assessments of CoT distillation focus primarily on accuracy (Magister et al., 2023; Ho et al., 2023; Shridhar et al., 2023; Wang et al., 2023c), which is insufficient because safe LLMs rely heavily on trustworthy CoTs. We hope the community to develop standards for evaluating the quality of CoTs, rather than relying solely on automatic assessments by GPT-4.

From Dai et al. (2025).

Limitations

× *Don't:* Focus on future work

Limitations

Our study focuses on three key domains where LLMs explicitly struggle—Math, Intention, and Time—building on insights from existing literature. However, LLMs also face challenges in areas such as long-tail knowledge and domain-specific expertise, where external resources are essential. Expanding SMART-ER to these domains could further refine model self-awareness and improve calibration in knowledge boundary, complementing the strong OOD performance that SMARTAgent has already demonstrated. Additionally, while we evaluate our approach on two major model families, extending our analysis to a broader range of architectures, including Qwen, DeepSeek, and varying model sizes, could further validate and enhance the generalizability of our findings.

From Qian et al. (2025).

Limitations

✓ Do: Describe scope of study (future work OK in this context)

6 Limitations

Hyperparameter Search The current ExPO adopts the simplest form of uniform extrapolation and requires manual hyperparameter search for α. Future work could explore how to determine the optimal α automatically and adaptively (i.e., using different α values for different model modules). For example, the information from optimizer states and parameter gradients during the later phase of alignment training could be useful for this purpose.

From Zheng et al. (2025).

Limitations

✓ Do: Discuss assumptions

Limitation

Dependency on Backbone Models The effectiveness of our GLPN-LLM framework is closely tied to the performance of the underlying backbone models, namely FCN. While these models provide strong feature representations, any limitations in their ability to capture comprehensive semantic relationships can directly impact the label propagation process. Consequently, the overall detection accuracy is highly dependent on the quality and robustness of these backbone models. Additionally, the reliance on specific backbones may limit the adaptability of our framework to other feature extraction architectures that might offer different advantages.

Reliance on High-Confidence Pseudo Labels Our approach relies on the generation of high-confidence pseudo labels by the LLM to enhance label propagation. However, the accuracy of these pseudo labels is contingent upon the LLM's ability to produce reliable predictions. Inaccurate or biased pseudo labels can introduce noise into the label propagation process, potentially degrading the model's performance. Ensuring the reliability of pseudo labels is crucial, and future work may explore more robust methods for pseudo label verification and refinement to mitigate this limitation.

From Hu et al. (2025).

Exercise 12: Evaluating limitations

Look at the limitations on the next page. Discuss the following questions:

1. Can you identify a the scope of the study?
2. Is it clear in which situations the results may not be applicable??
3. Are there points you would expand on?
4. Given the abstract, are there any potential limitations not mentioned you notice?

These limitations are from papers that appeared at ACL 2025.

Exercise 12: Evaluating limitations

Limitations

Sensitive to Prompts. As with other LLM prompting studies(Zhang et al., 2024b; Huang et al., 2024b; Zhang et al., 2024a; Chen et al., 2024), our results may be sensitive to prompt. While our prompts underwent rigorous review and testing, and our main experiments report averages across over 8,000 problems, optimizing prompts for this specific task remains a significant challenge and area for future research.

Generalizability to Other Programming Tasks. In accordance with scientific rigor, this study defines its scope as Human-LLM collaboration within competitive programming, a domain chosen to examine the capabilities and limitations of both LLMs and human performance. While acknowledging the potential relevance to broader programming tasks, we limit our evaluations and analyses to this specific context and defer extending the representativeness of our results to general software development or other programming domains. Despite this focus, elements of our work offer valuable insights applicable to diverse programming scenarios. The problem-solving process shares fundamental similarities across programming contexts, and our proposed human feedback taxonomy and methods for improving problem comprehension in LLMs may readily translate. Developers, for example, can leverage clear and detailed feedback on specifications, as demonstrated in our benchmark, to guide LLMs towards a better understanding of software requirements. We believe this highlights pathways for broader applicability and welcome further discussion.

Exercise 12: Evaluating limitations

Abstract

While recent research increasingly emphasizes the value of human-LLM collaboration in competitive programming and proposes numerous empirical methods, a comprehensive understanding remains elusive due to the fragmented nature of existing studies and their use of diverse, application-specific human feedback. Thus, our work serves a three-fold purpose: First, we present the first taxonomy of human feedback consolidating the entire programming process, which promotes fine-grained evaluation. Second, we introduce ELABORATIONSET, a novel programming dataset specifically designed for human-LLM collaboration, meticulously annotated to enable large-scale simulated human feedback and facilitate cost-effective real human interaction studies. Third, we introduce ELABORATION, a novel benchmark to facilitate a thorough assessment of human-LLM competitive programming. With ELABORATION, we pinpoint strengthes and weaknesses of existing methods, thereby setting the foundation for future improvement.

Exercise 12: Evaluating limitations

Limitations

Although this work demonstrates the potential of RLKGF, several issues need to be addressed. The quality of feedback derived from knowledge graphs depends heavily on the completeness and accuracy of the graph itself, particularly in open domains. Our experiments are limited to disease diagnosis tasks without exploring RLKGF's generalization to other tasks and domains. Additionally, due to data limitations, we do not conduct experiments across a broader medical framework.

The current task format is single-turn Q&A, and future work should explore multi-turn dialogues to better leverage the potential advantages of knowledge graph structure and semantics in multi-step reasoning. Moreover, RLKGF currently focuses primarily on entity-level feedback for model responses, with limited focus on overall response fluency. Furthermore, experimental comparisons show that although RLKGF improves consistency between model responses and knowledge, there is still significant room for enhancement. Designing appropriate reward ranges and investigating the impact of different methods on model parameter adjustments are crucial for continuous knowledge learning.

Appendices

Bonus content - main paper should be self-contained

✓ **DOs**

- Put examples
- Hyperparameter details
- Extra analyses
- Annotation instructions
- Follow formatting guidelines

× **DON'Ts**

- Put *all* your examples
- All experiment details
- Main results
- All annotation information
- Include content that spills into margins

Often used for details in the *Reproducibility Checklist* (we'll discuss in a few weeks)

Miscellaneous resources

- 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/

References

Michael Alley. 2018. *The Craft of Scientific Writing*, 4 edition. Springer, New York, NY.

Jacy Reese Anthis, Kristian Lum, Michael Ekstrand, Avi Feller, and Chenhao Tan. 2025. The impossibility of fair LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 105–120, Vienna, Austria. Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Wenting Chen, Linlin Shen, Jingyang Lin, Jiebo Luo, Xiang Li, and Yixuan Yuan. 2024. Fine-grained image-text alignment in medical imaging enables explainable cyclic image-report generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9494–9509, Bangkok, Thailand. Association for Computational Linguistics.

Chengwei Dai, Kun Li, Wei Zhou, and Songlin Hu. 2025. Capture the key in reasoning to enhance CoT distillation generalization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 441–465, Vienna, Austria. Association for Computational Linguistics.

Boyi Deng, Yu Wan, Baosong Yang, Yidan Zhang, and Fuli Feng. 2025. Unveiling language-specific features in large language models via sparse autoencoders. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4563–4608, Vienna, Austria. Association for Computational Linguistics.