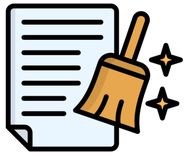
Academic Paper Writing in Artificial Intelligence

Datasets



Instructor: Ying-Jia Lin
September 17th, 2026

Couse Website
Download slide here!

Outline

- General guide (20 min)
- How to write the “Dataset” section? (30 min)
- How to cite using BibTeX (5 min)
- **Your Turn (70 min, including break)**
- How to build a table using LaTeX (10 min)
- **Your Turn (15 min)**

General Guide

For all sections, especially the Dataset section.

1. Voice and tense: we not I; present vs. past
2. One sentence, one idea.
3. Write the full name instead of abbreviations **at first mention**
    - Naming consistency throughout the paper: QLoRA, QLORA, Q-LoRA
4. Grammar traps (e.g., fewer noise, less examples)

1️⃣ Voice - Use “we”, not “I”

- Use we even if you are the only author. It is the convention in academic papers.

  ❌ In this paper, I propose a new training method.

  ✅ In this paper, we propose a new training method.

- 💡 This does not mean we need to always write in **the active voice**. Medical journals often prefer the **passive** in Methods.

  ✅ The statistics of biomedical NER datasets are listed in Table 3.

1 Tense and Examples

Past Tense

- What you did
- What others did

Present Tense

- What is always true
- What the paper does

1 Tense and Examples

Past Tense

- What you did
- What others did

Present Tense

- What is always true
- What the paper does

We removed 312 reports without ICD codes.

The model was initialized from the pre-trained GPT-2.

We split the data into 8:1:1.

Lin et al. proposed an unsupervised model using pseudo-labeling.

1 Tense and Examples

Past Tense

- What you did
- What others did

We removed 312 reports without ICD codes.

The model was initialized from the pre-trained GPT-2.

We split the data into 8:1:1.

Lin et al. proposed an unsupervised model using pseudo-labeling.

Present Tense

- What is always true
- What the paper does

“We propose …”; “Table 2 shows …”

The dataset contains 4,392 text reports.

Our model outperforms the baseline approaches.

2 One sentence, one idea.

Although our proposed method achieved better performance than the baseline methods on most evaluation metrics, the improvement was relatively small, which suggests that further investigation may still be necessary to determine whether the proposed method can consistently outperform existing approaches across different datasets.

2️⃣ One sentence, one idea.

Although our proposed method achieved better performance than the baseline methods on most evaluation metrics, the improvement was relatively small, which suggests that further investigation may still be necessary to determine whether the proposed method can consistently outperform existing approaches across different datasets.

⬇️

Our method outperformed the baselines on most metrics. However, the improvement was small. Further experiments are needed to determine whether the method generalizes to other datasets.

2️⃣ One sentence, one idea.

Although our proposed method achieved better performance than the baseline methods on most evaluation metrics, the improvement was relatively small, which suggests that further investigation may still be necessary to determine whether the proposed method can consistently outperform existing approaches across different datasets.

⬇️

Our method outperformed the baselines on most metrics. However, the improvement was small. Further experiments are needed to determine whether the method generalizes to other datasets.

One sentence, one main idea. Simple is the best.

3️⃣ Naming consistency

- Write the full name at the first mention, with the short form in brackets.

✅ “We fine-tune a large language model (LLM) with quantized low-rank adaptation (QLoRA). The LLM is …”

3️⃣ Naming consistency

- Write the full name at the first mention, with the short form in brackets.

✅ “We fine-tune a large language model (LLM) with quantized low-rank adaptation (QLoRA). The LLM is …”

- **Use the same short form everywhere else in your manuscript.**

3️⃣ Naming consistency

- Write the full name at the first mention, with the short form in brackets.

✅ “We fine-tune a large language model (LLM) with quantized low-rank adaptation (QLoRA). The LLM is …”

- **Use the same short form everywhere else in your manuscript.**

❌ QLoRA … QLORA … Q-LoRA … Qlora

✅ QLoRA … QLoRA … QLoRA … QLoRA

4️⃣ Grammar Traps

4️⃣ Grammar Traps

- ❌ We use BERT model.

4️⃣ Grammar Traps

- ❌ We use BERT model.
- ✅ We use **the** BERT model. ← articles

4️⃣ Grammar Traps

- ❌ We use BERT model.
- ✅ We use **the** BERT model. ← articles
- ❌ Previous **researches** show…

4️⃣ Grammar Traps

- ❌ We use BERT model.
- ✅ We use **the** BERT model. ← articles
- ❌ Previous **researches** show…
- ✅ Previous **studies** show… ← "research" has no plural

4️⃣ Grammar Traps

- ❌ We use BERT model.
- ✅ We use **the** BERT model. ← articles
- ❌ Previous **researches** show…
- ✅ Previous **studies** show… ← "research" has no plural
- ❌ Many **literatures** have…

4️⃣ Grammar Traps

- ❌ We use BERT model.
- ✅ We use **the** BERT model. ← articles
- ❌ Previous **researches** show…
- ✅ Previous **studies** show… ← "research" has no plural
- ❌ Many **literatures** have…
- ✅ **Much of the literature** has… ← neither does "literature"

4️⃣ Grammar Traps

- ❌ We use BERT model.
- ✅ We use **the** BERT model. ← articles
- ❌ Previous **researches** show…
- ✅ Previous **studies** show… ← "research" has no plural
- ❌ Many **literatures** have…
- ✅ **Much of the literature** has… ← neither does "literature"
- ❌ **Because** A, **so** B.

4️⃣ Grammar Traps

- ❌ We use BERT model.
- ✅ We use **the** BERT model. ← articles
- ❌ Previous **researches** show…
- ✅ Previous **studies** show… ← "research" has no plural
- ❌ Many **literatures** have…
- ✅ **Much of the literature** has… ← neither does "literature"
- ❌ **Because** A, **so** B.
- ✅ **Because** A, B. ← English uses one, not both

4️⃣ Grammar Traps

- ❌ We use BERT model.
- ✅ We use **the** BERT model. ← articles
- ❌ Previous **researches** show…
- ✅ Previous **studies** show… ← "research" has no plural
- ❌ Many **literatures** have…
- ✅ **Much of the literature** has… ← neither does "literature"
- ❌ **Because** A, **so** B.
- ✅ **Because** A, B. ← English uses one, not both

- ❌ Fewer noise; less examples

4️⃣ Grammar Traps

- ❌ We use BERT model.
- ✅ We use **the** BERT model. ← articles
- ❌ Previous **researches** show…
- ✅ Previous **studies** show… ← "research" has no plural
- ❌ Many **literatures** have…
- ✅ **Much of the literature** has… ← neither does "literature"
- ❌ **Because** A, **so** B.
- ✅ **Because** A, B. ← English uses one, not both

- ❌ Fewer noise; less examples
- ✅ Less noise; Fewer examples

4️⃣ Grammar Traps

- ❌ We use BERT model.
- ✅ We use **the** BERT model. ← articles
- ❌ Previous **researches** show…
- ✅ Previous **studies** show… ← "research" has no plural
- ❌ Many **literatures** have…
- ✅ **Much of the literature** has… ← neither does "literature"
- ❌ **Because** A, **so** B.
- ✅ **Because** A, B. ← English uses one, not both

- ❌ Fewer noise; less examples
- ✅ Less noise; Fewer examples

Do not make too many mistakes in your manuscript. Use Grammarly or DeepL. 💰

Datasets

Overall Structure and the Dataset Section

	Main Body (Your Main Paper)
	Abstract
	Introduction
Depends on where you submit	Related Work (Literature Review)
	Methods
	Results
Depends on where you submit	Discussion
	Conclusion

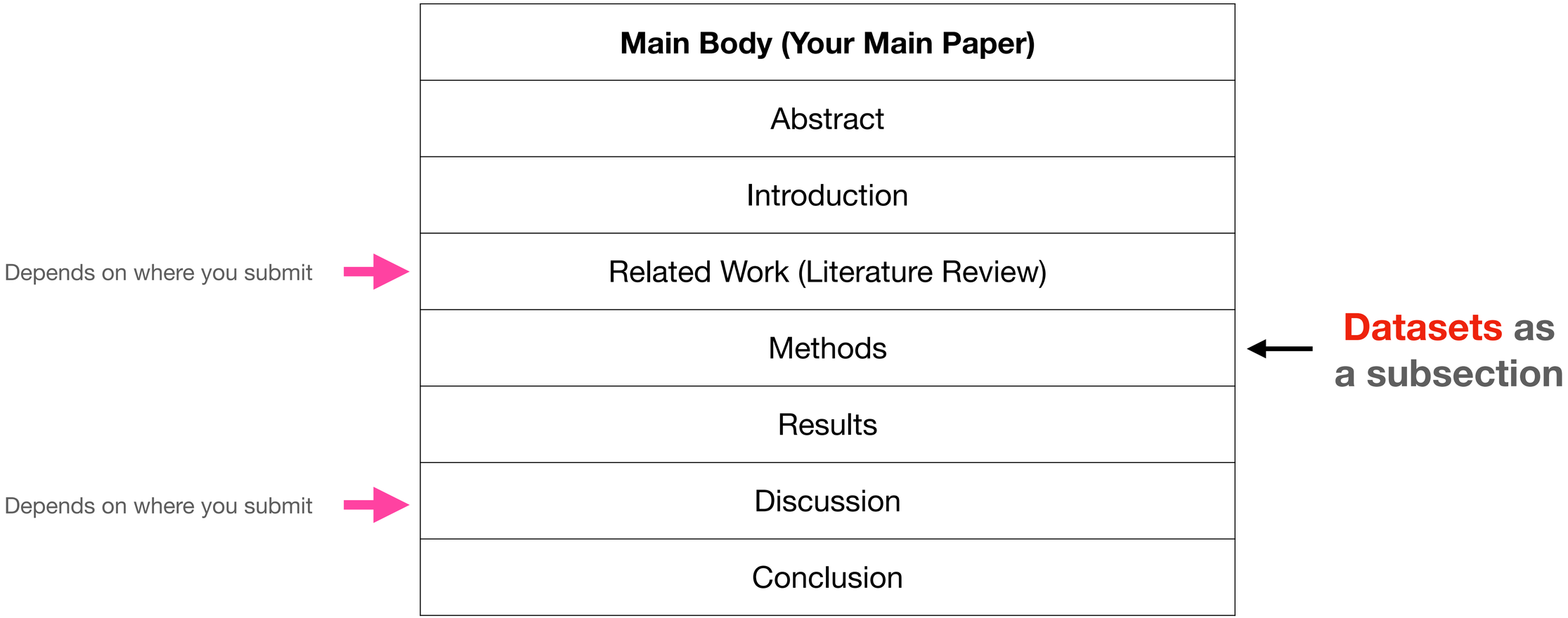
Overall Structure and the Dataset Section

Main Body (Your Main Paper)
Abstract
Introduction
Related Work (Literature Review)
Methods
Results
Discussion
Conclusion

Depends on where you submit

Depends on where you submit

Overall Structure and the Dataset Section



The Purpose of Dataset(s)

- In machine learning research, the results are **data-driven**.
- To test a model, we use a test set.
- To train a model, we use a training set.
- You should explicitly tell readers how you use your dataset(s).

Difference between Data and Datasets

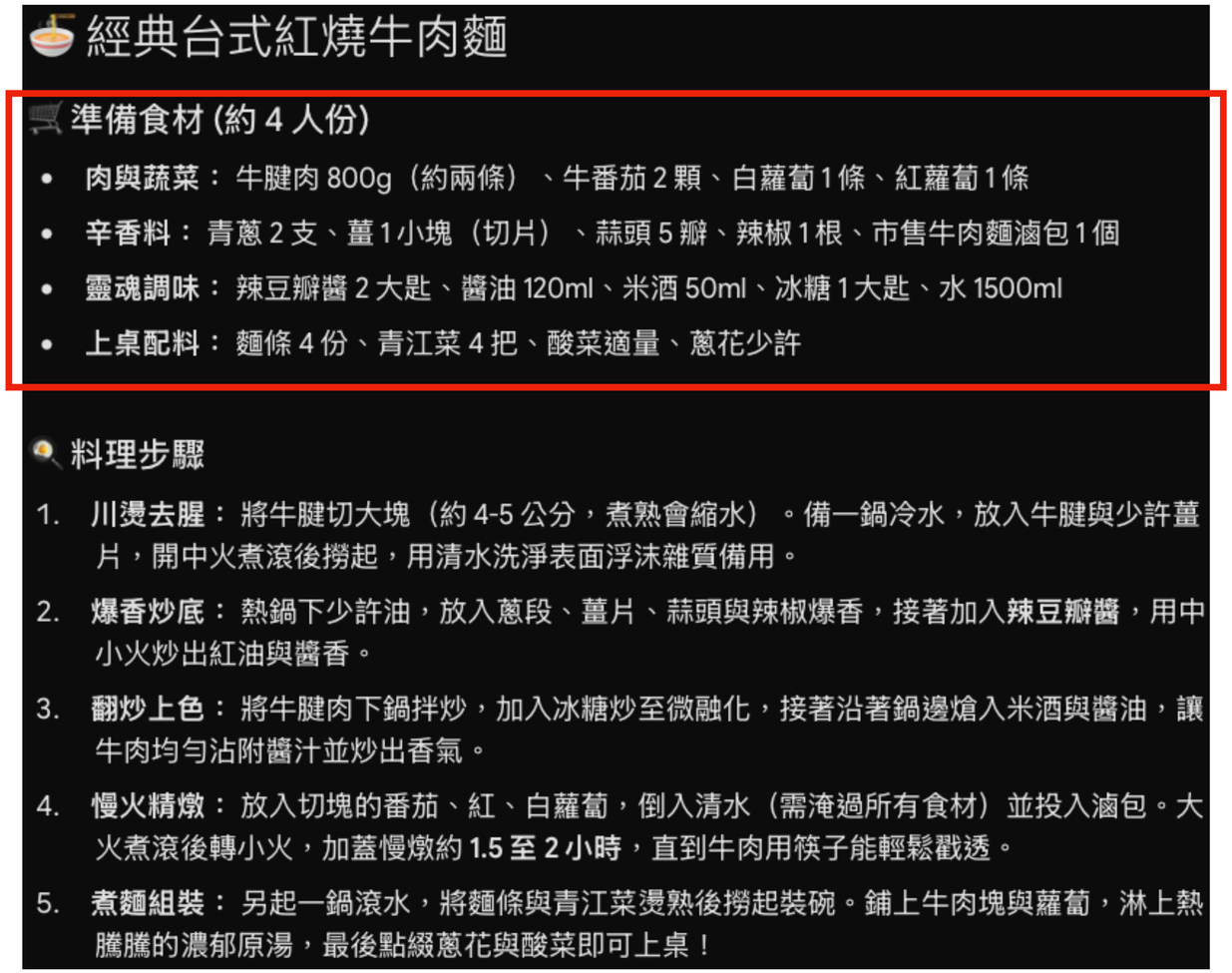
- In an academic paper, you don't just use data; you construct or utilize a dataset.
- A dataset represents a rigorous, reproducible foundation for your experiments.

	Data	**Dataset**
State	Raw, unorganized, noisy	Curated, filtered, refined
Structure	Unstructured	Structured
Annotation	Typically unlabeled	Usually labeled

Checklist

- **Source:** where does the dataset you use come from?
- **Statistics:** number of instances, class distribution, average length (NLP), image sizes (CV), and so on.
- **Preprocessing & Filtering:** Did you remove some weird data or perform normalization? Tell readers all.
- **Splits:** how to split your dataset into training, validation and test sets?
- **Ethical Considerations:** De-identification, IRB number (for medical papers).

Don’t be afraid !!



🍜 經典台式紅燒牛肉麵

🛒 準備食材 (約 4 人份)

- **肉與蔬菜：** 牛腱肉 800g（約兩條）、牛番茄 2 顆、白蘿蔔 1 條、紅蘿蔔 1 條
- **辛香料：** 青蔥 2 支、薑 1 小塊（切片）、蒜頭 5 瓣、辣椒 1 根、市售牛肉麵滷包 1 個
- **靈魂調味：** 辣豆瓣醬 2 大匙、醬油 120ml、米酒 50ml、冰糖 1 大匙、水 1500ml
- **上桌配料：** 麵條 4 份、青江菜 4 把、酸菜適量、蔥花少許

🍳 料理步驟

1. **川燙去腥：** 將牛腱切大塊（約 4-5 公分，煮熟會縮水）。備一鍋冷水，放入牛腱與少許薑片，開中火煮滾後撈起，用清水洗淨表面浮沫雜質備用。
2. **爆香炒底：** 熱鍋下少許油，放入蔥段、薑片、蒜頭與辣椒爆香，接著加入**辣豆瓣醬**，用中小火炒出紅油與醬香。
3. **翻炒上色：** 將牛腱肉下鍋拌炒，加入冰糖炒至微融化，接著沿著鍋邊炝入米酒與醬油，讓牛肉均勻沾附醬汁並炒出香氣。
4. **慢火精燉：** 放入切塊的番茄、紅、白蘿蔔，倒入清水（需淹過所有食材）並投入滷包。大火煮滾後轉小火，加蓋慢燉約 **1.5 至 2 小時**，直到牛肉用筷子能輕鬆戳透。
5. **煮麵組裝：** 另起一鍋滾水，將麵條與青江菜燙熟後撈起裝碗。鋪上牛肉塊與蘿蔔，淋上熱騰騰的濃郁原湯，最後點綴蔥花與酸菜即可上桌！

- As we mentioned earlier, **Datasets** is a subsection in **Methods**.
- Writing content in Methods is just like writing the ingredient part of a recipe.
- ➡️ Imagine you are writing something to help others cook (reproduce) a bowl of beef noodles (your method)

❌ some beef
✅ 800g beef shank (two pieces)

❌ We use most of the qualified questionnaire responses.
✅ We collected 412 responses and excluded 37 that were incomplete, leaving 375 for the analysis.

Three Kinds of Dataset Section

Before you write, answer one question: **can your reader get this data?**

	Public dataset	**Private dataset**	**You built it**
Example	MIMIC, CheXpert	hospital records, your questionnaire	a corpus you annotated
Length	100–200 words	a full subsection	often a whole section
Must include	citation, statistics, split	cohort, inclusion/exclusion, IRB	how it was made, by whom, how good it is
Do not	re-explain the original paper	skip the annotators	

If you are the middle case

Hospital records and your own questionnaire — **nobody can rerun your work**.

So the writing has to do the job that reproducibility normally does.

Your reader asks three questions:

1. **Where did it come from?** — source, data size, content
2. **How was it selected, and by whom?** — inclusion, exclusion, who decided
3. **How do I know it is any good?** — agreement, representativeness, ethics

M1: Where did it come from?

We obtained the de-identified 2022 electronic medical records from the trauma department of Chang Gung Memorial Hospital, a Level I trauma center in Taiwan. The raw dataset comprised 4,392 cases, where each record included free-text discharge diagnoses, ICD-10-CM codes, AIS scores (1998 version, the current standard in Taiwan's National Health Insurance System), and an ISS score.

← **Source**

← **Dataset size**

← **The content**

M2-1: How was it selected?

Ninety cases were excluded because they could not be reliably annotated from the discharge diagnosis text alone: **the text lacked sufficient detail, could not be interpreted without reference to other diagnostic fields, or was inconsistent with the recorded diagnosis codes.**

← **Reasons**
← **for the data**
← **removal.**

M2-2: Who selected the data?

Who helped the data selection / creation.

Why the people are qualified.

What did they do to select the data

"We recruited nine writers and annotators… all native Chinese speakers … trained by the authors until they could annotate correctly and independently."

"A physician reviewed cases from this pool of 300 cases, assessing whether each discharge diagnosis contained sufficient clinical detail…"

M3: How do I know it is any good?

***Fleiss Kappa is an example. You need to find a way to show the quality of your data.**

"Our dataset shows **0.7934** in Fleiss κ, higher than 0.6841 in FEVER and 0.74 in CHEF, measured from 4% and 3% of those datasets." ✅✅ (Evidence and comparisons written.)

"Three physicians independently annotated the data, achieving high inter-rater agreement (κ = **0.801**)." ✅ (Evidence written.)

"The annotations from the three physicians showed **a high level of agreement**." ☹️ (No evidence and no comparisons.)

Ethical Considerations

You cannot invent this section. This section usually follows a template.
Read your target journal's instructions, then two recent papers from it.

General case:

1. Who approved it? → "approved by the institutional research ethics board (approval no. 202500438B0)"

2. What about consent? → "informed consent was waived… because the study involved retrospective analysis of existing medical records"

3. How were people protected? → "de-identified before analysis… no identifiable participant images or personal identifiers"

Ethical Considerations (example)

1. Who approved it?

2. What about consent?

3. How were people protected?

This retrospective study was approved by the Institutional Review Board of [Institution Name] (IRB No. XXXXX). The requirement for informed consent was waived because the study involved retrospective analysis of de-identified clinical data. All patient information was anonymized prior to analysis, and no personally identifiable information was accessed or reported.

Example ✅

Luo, Ling, et al. "An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition." Bioinformatics 34.8 (2018): 1381-1388.

- In our experiments, two corpora (http://www.biocreative.org/resources/) released by the BioCreative challenge were used: the CHEMDNER corpus (Krallinger et al., 2015) and the CDR task corpus (Li et al., 2016). **Source**
- Overall statistics for each dataset are provided in Supplementary Material: BioCreative Corpora. **Statistics**
- Like many teams in the challenge, the original training set and development set were used as the training set. **Preprocessing & Filtering**
- Then we randomly selected 10% of the training set as the validation set to tune the hyper-parameters. **Splits**
- The chemical NER performance was measured with an F-score which attributes equal importance to precision and recall (F1 score) on the test set.

A useful order - but not a rule

Order	Category	Example
1	Source	In our experiments, two corpora (http://www.biocreative.org/resources/) released by the BioCreative challenge were used: the CHEMDNER corpus (Krallinger et al., 2015) and the CDR task corpus (Li et al., 2016).
2	Statistics	Overall statistics for each dataset are provided in Supplementary Material: BioCreative Corpora.
3	Preprocessing & Filtering	Like many teams in the challenge, the original training set and development set were used as the training set.
4	Splits	Then we randomly selected 10% of the training set as the validation set to tune the hyper-parameters.
5	Ethical Considerations	This retrospective study was approved by the Institutional Review Board of [Institution Name] (IRB No. XXXXX), with informed consent waived due to the use of de-identified clinical data. All data were anonymized before analysis, and no personally identifiable information was accessed or reported.

Example: what would you fix?

Priyadharshini, S., et al. "A successive framework for brain tumor interpretation using Yolo variants." *Scientific Reports* 15.1 (2025): 27973.

> The preprocessing pipeline accomplished three essential tasks by performing data augmentations while standardizing pixel intensity values to the [0, 1] interval and adjusting image dimensions to 512 × 512 pixels. The automated contour refining procedure enables the mask alignment approach to solve this problem.

Example: what would you fix? (Ans.)

Priyadharshini, S., et al. "A successive framework for brain tumor interpretation using Yolo variants." *Scientific Reports* 15.1 (2025): 27973.

The preprocessing pipeline accomplished three essential tasks. We performed data augmentations, standardized pixel intensity values to the [0, 1] interval, and adjusted image dimensions to 512 × 512 pixels. The automated contour refining procedure enables the mask alignment approach to solve this problem.

Example: the part you cannot fix

⚠️ The last sentence cannot be fixed without the authors since “this problem” was never stated.

Pre-processing

The preprocessing pipeline accomplished three essential tasks by performing data augmentations while standardising pixel intensity values to the [0, 1] interval and adjusting image dimensions to 512 × 512 pixels. The automated contour refining procedure enables the mask alignment approach to solve this problem.

Using Histogram equalization technology enhances MRI brain scans by improving their contrast which makes tumor detection regions easier to observe shown in Figs. 6 and 7. The initial preprocessing operation makes it easier for the YOLO model to detect and segment brain tumors correctly.

Example: the part you cannot fix

⚠️ The last sentence cannot be fixed without the authors since “this problem” was never stated.

Pre-processing

The preprocessing pipeline accomplished three essential tasks by performing data augmentations while standardising pixel intensity values to the [0, 1] interval and adjusting image dimensions to 512 × 512 pixels. The automated contour refining procedure enables the mask alignment approach to solve this problem.

Using Histogram equalization technology enhances MRI brain scans by improving their contrast which makes tumor detection regions easier to observe shown in Figs. 6 and 7. The initial preprocessing operation makes it easier for the YOLO model to detect and segment brain tumors correctly.

☑ **The this check**:

Search your draft for “this” and “these”

Every one must point back at a noun the readers have already met in your paper.

Example: what would you fix?

Priyadharshini, S., et al. "A successive framework for brain tumor interpretation using Yolo variants." *Scientific Reports* 15.1 (2025): 27973.

Using Histogram equalization technology enhances MRI brain scans by improving their contrast which makes tumor detection regions easier to observe shown in Figs. 6 and 7. The initial preprocessing operation makes it easier for the YOLO model to detect and segment brain tumors correctly.

Example: what would you fix? (small fixes)

Priyadharshini, S., et al. "A successive framework for brain tumor interpretation using Yolo variants." *Scientific Reports* 15.1 (2025): 27973.

Using histogram equalization technology enhances MRI brain scans by improving their contrast, which makes tumor detection regions easier to observe, shown in Figs. 6 and 7. The initial preprocessing operation makes it easier for the YOLO model to detect and segment brain tumors correctly.

Example: what would you fix? (Ans.)

Priyadharshini, S., et al. "A successive framework for brain tumor interpretation using Yolo variants." *Scientific Reports* 15.1 (2025): 27973.

❌ Using histogram equalization technology enhances MRI brain scans by improving their contrast, which makes tumor detection regions easier to observe, shown in Figs. 6 and 7. The initial preprocessing operation makes it easier for the YOLO model to detect and segment brain tumors correctly.

✅ Histogram equalization technology enhances MRI brain scans by improving their contrast, which makes tumor detection regions easier to observe (Figs. 6 and 7). The initial preprocessing operation also helps the YOLO model detect and segment brain tumors correctly.

Your Turn

How to cite using BibTeX

Outline

Show full outline

ELSEVIER

International Journal of Medical Informatics

Volume 220, 1 November 2026, 106592

Original Research Article

MedNLP-Hub: A knowledgebase platform for biomedical NLP tool discovery in clinical informatics

Shumin Ren [a 1], Yuxin Zhang [b 1], Jinyi Zhang [a c], Hongmei Lang [a], Hui Zong [b], Bairong Shen [b]

Show more

https://doi.org/10.1016/j.ijmedinf.2026.106592 ↗ Get rights and content ↗

Abstract

Objectives

Biomedical natural language processing (NLP) tools are used to extract

https://www.sciencedirect.com/science/article/pii/S1386505626003321

LaTeX Table

```
\begin{tabular}{ccc}
A & B & C \\
1 & 2 & 3 \\
4 & 5 & 6 \\
\end{tabular}
```

l = left-aligned
c = center-aligned
r = right-aligned

Summary

1. Use “we”. Describe the dataset in the present tense, what you did in the past.
2. One sentence, one idea.
3. Never change the naming consistency.
4. Ask yourself: can your reader get the data? The answer decides how much you write.
5. If your readers cannot get the data, you have to clearly describe where it came from, how it was selected and by whom, and how you know the data is good.

NEXT WEEK ATTENTION!!

- Next week we will write the **Evaluation Metrics and Baselines** subsections.
- Please prepare:
  - Metrics you use and the details (e.g., custom metric or tool details.)
  - Baseline methods in your work
  - Experimental settings
    - Implementation details
    - Settings for an ablation study
    - …

Thank you for your participation 📚