Academic Paper Writing in Artificial Intelligence

Results - Tables



Course Website
Download slide here!

Instructor: Ying-Jia Lin
Octobor 29th, 2026

Outline

- Week 7 Assignment Summary (20 min)
- What a Results section is (30 min)
- Related Work in **3(4) steps** and examples (30 min)
- **Your Turn for revision and new assignment (70 min)**
  - Feel free to discuss with me !!

Week 7 Assignment Summary (1/3)

- Step 3 (Week 7) is not clearly written.
  - A lack of literature **synthesis**
- Study A proposed method X. Study B developed method Y. Study C investigated method Z. ⬅️ Just literature **summary**
- **Research-oriented** comparisons between your work and others are needed.

Week 7 Assignment Summary (2/3)

- No clear evidence from previous works support research gap (step 2-3).
- ❌ However, previous studies have not adequately investigated …
  - You should tell the readers **in which conditions** previous studies are not enough for your research topic.
    - E.g., data selection? career retention? optimizer comparisons?

Week 7 Assignment Summary (3/3)

- Repetitive emphasis for your key difference (step 3) for every paragraph
- Every paragraph should have a different summary.
  - Just put step 3 in the last paragraph closest to your main research.

Where we are

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

The Results section in illustration



From Metrics (Week 3) to Results (Week 8)

- In Week 3, we said, "Our model reaches **92.4**."
- Week 3 set up the comparison: 92.4 of what? Compared with what?
- Week 8 explains the result. Explaining has two parts:
  - What the comparison shows — always in Results, in every field
  - Why it came out that way — in Results for AI venues,
    but in Discussion (Week 10) for medical journals

Results (This page is from Week1)

- Results from experiments are described in this section.
- The most important part of your paper.

Task (metric)	System	Score (%)
Doc retrieval (Recall)	Our baseline	87.65
	BEVERS[a]	92.60
Sent retrieval (Recall)	Our baseline	76.65
	BEVERS[a]	86.60
RTE (Accuracy)	Our baseline	61.17
	BEVERS[a]	69.73
	GPT-3.5 (*zeroshot*)	43.17
	GPT-3.5 (*3-shot*)	44.20
	GPT-4 (*zeroshot*)	47.23
	GPT-4 (*3-shot*)	48.40
Full pipeline (FEVER Score)	Our baseline	52.47
	BEVERS[a]	64.80

Table 4: Results on the different stages. *a*: DeHaven and Scott (2023).

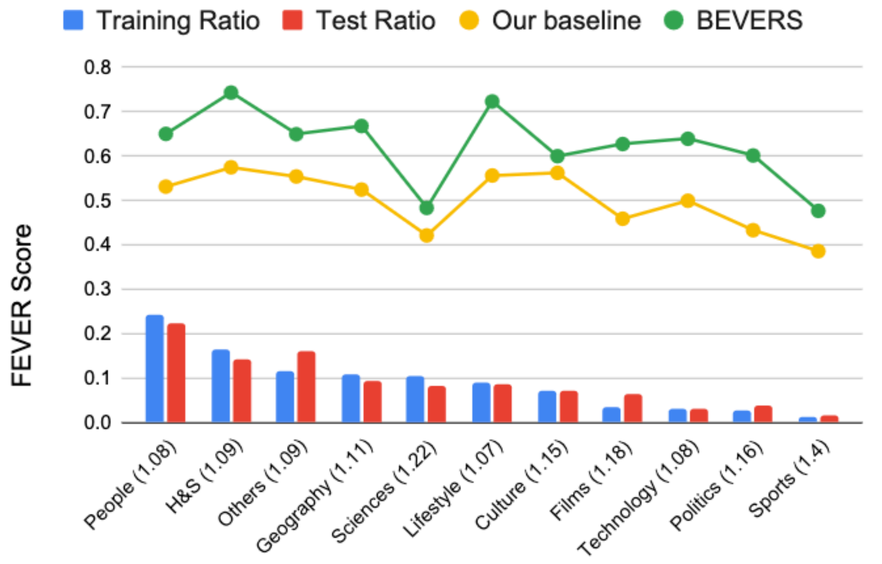


Figure 2: Performance comparisons for the claims of different domains in the full-pipeline setting. H&S refers to the domain of Humanities & Social Sciences. Values in the parentheses are the average number of evidence pages.

Results (This page is from Week1)

- Results from experiments are described in this section.
- The most important part of your paper.

This week ✅

Task (metric)	System	Score (%)
Doc retrieval (Recall)	Our baseline	87.65
	BEVERS[a]	92.60
Sent retrieval (Recall)	Our baseline	76.65
	BEVERS[a]	86.60
RTE (Accuracy)	Our baseline	61.17
	BEVERS[a]	69.73
	GPT-3.5 (*zeroshot*)	43.17
	GPT-3.5 (*3-shot*)	44.20
	GPT-4 (*zeroshot*)	47.23
	GPT-4 (*3-shot*)	48.40
Full pipeline (FEVER Score)	Our baseline	52.47
	BEVERS[a]	64.80

Table 4: Results on the different stages. a: DeHaven and Scott (2023).

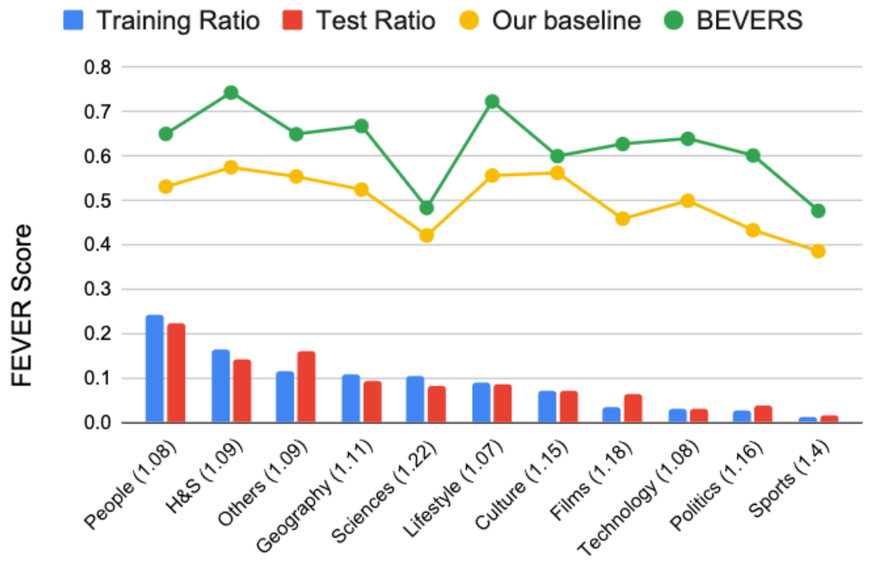


Figure 2: Performance comparisons for the claims of different domains in the full-pipeline setting. H&S refers to the domain of Humanities & Social Sciences. Values in the parentheses are the average number of evidence pages.

Results (This page is from Week1)

- Results from experiments are described in this section.
- The most important part of your paper.

This week ✅

Task (metric)	System	Score (%)
Doc retrieval (Recall)	Our baseline	87.65
	BEVERS[a]	92.60
Sent retrieval (Recall)	Our baseline	76.65
	BEVERS[a]	86.60
RTE (Accuracy)	Our baseline	61.17
	BEVERS[a]	69.73
	GPT-3.5 (*zeroshot*)	43.17
	GPT-3.5 (*3-shot*)	44.20
	GPT-4 (*zeroshot*)	47.23
	GPT-4 (*3-shot*)	48.40
Full pipeline (FEVER Score)	Our baseline	52.47
	BEVERS[a]	64.80

Table 4: Results on the different stages. *a*: DeHaven and Scott (2023).

Next week ✅

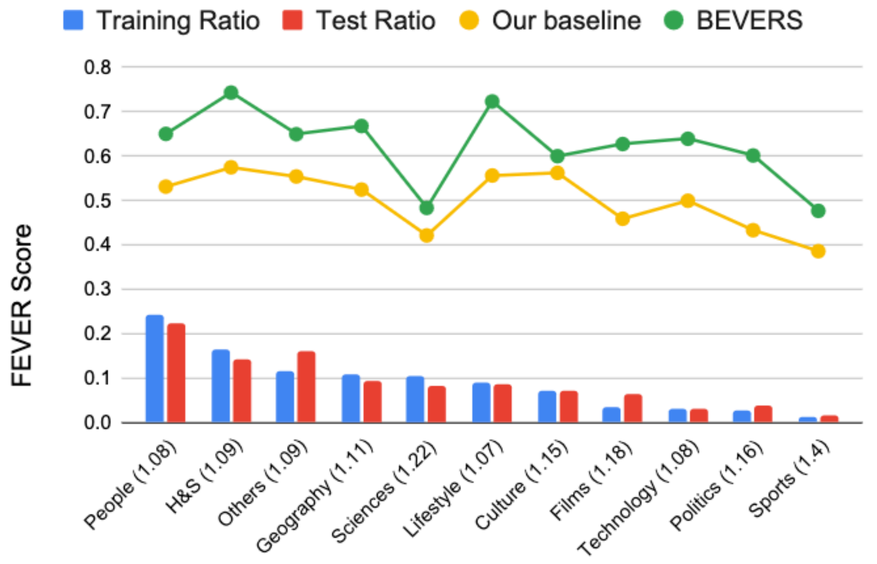


Figure 2: Performance comparisons for the claims of different domains in the full-pipeline setting. H&S refers to the domain of Humanities & Social Sciences. Values in the parentheses are the average number of evidence pages.

Why writing text if values’ve been in a table?

- **The table holds the evidence. The text makes the claim.**
- **A table gives the reader numbers. It does not tell them what to see.**
  - Never make the reader do the comparison themselves.
  - If your text only says "The results are shown in Table 2", every reader has to work out the answer alone — and some will find a different one.
  - Your paragraph tells them what the numbers show, and points to the numbers that prove it.

The test

- Cover the table. Ask a classmate to read only your Results paragraph. Then ask: "What did this paper find, and by how much?"
  - If they can answer both parts, the paragraph works.
  - If they say "your method is better" but cannot say better than what, or by how much, the paragraph is not finished.

How a Results section is organized

One question, one table, one subsection

Subsection	The question it answers	Usually shown in
Main Results	Does our method beat the baselines?	Table
Ablation Study	Which part of our method matters?	Table
Analysis	When does it work, and when does it fail?	Table or figure (Week 9)

Start each subsection with its question

❌ Starting with the table:

"Table 3 shows the ablation results."

✅ Starting with the question:

"To see which component contributes most, we remove each one in turn and report the results in Table 3."

The reader should know what to look for before they look.

(This is the same idea as Week 7: say what a group of papers has in common before you list the papers.)

AI venues vs medical journals

	AI conference	Medical journal
Separate Discussion section?	often none, or merged into the Conclusion	always, often the longest section
Summing up the pattern	in Results	in Results
Explaining why	in Results, right after the number	in Discussion
How findings are worded	"outperforms", "improves"	"was associated with", with 95% CI or p-value

AI conference — explain right away

Woo, Sanghyun, et al. "Convnext v2: Co-designing and scaling convnets with masked autoencoders." CVPR 2023.

> We present results in Table 2f where we either remove GRN from fine-tuning or add newly initialized GRN only at the time of fine-tuning. Either way, we observe a significant performance degradation, **suggesting that keeping GRN in both pre-training and fine-tuning is important.**

💡 What happened, and what it means, in the same sentence.

AI conference — compare with other work right away

Kumar, Aviral, et al. "Training language models to self-correct via reinforcement learning." International Conference on Learning Representations. Vol. 2025. 2025.

> … replacing REINFORCE with STaR in Stage II results in significantly lower absolute
>
> performance …, which contrasts with the findings in Havrilla et al. (2024a)
>
> …
>
> **This suggests that** leveraging on-policy samples is especially critical in the self-correction setting …

💡 Result → comparison with another paper → explanation. All inside Results.

Medical journal — Results only reports

Tseng, LW., Lu, YC., Tseng, LC. et al. Model-dependent prompt effects challenge reliable deployment of medical LLMs in traditional Chinese medicine. npj Digit. Med. (2026). https://doi.org/10.1038/s41746-026-03397-5

> … within the DeepSeek family, explicit CoT prompting **was associated with** a 34.9-percentage-point increase in adjusted predicted accuracy relative to no CoT (95% CI, 30.6–39.1 percentage points) … No single prompt strategy yielded consistent benefits across all model families, indicating that prompt engineering effects are highly model-dependent.

- 💡 "associated with", not "caused" — medical writing is careful about cause
- 💡 one line summing up the pattern, but no why

Medical journal — the why waits for Discussion

Tseng, LW., Lu, YC., Tseng, LC. et al. Model-dependent prompt effects challenge reliable deployment of medical LLMs in traditional Chinese medicine. npj Digit. Med. (2026). https://doi.org/10.1038/s41746-026-03397-5

Same paper, **Discussion**:

> "… GPT-4o achieved 90.4% on USMLE-style questions despite scoring only 64.21% on the TCM examination in the present study under its best prompt condition. **This within-model disparity suggests that the gap is driven by domain-specific knowledge rather than general model capability.**"

💡 A comparison with another study, and an explanation. Both were held back until Discussion.

⚠️ In an AI paper, these two sentences could sit right next to the results table.

Both styles agree: start with the question

- SCoRe (ICLR 2025), before its ablation table:
  - "Concretely, we aim to answer the following questions: (1) the importance of multi-turn training … (2) the importance of multi-stage training …"
- Levin et al. (International Journal of Nursing Studies, 2025), start of Section 3.4:
  - "This subsection corresponds to our third aim—evaluating the internal logic and consistency of decision-making processes— …"

Three steps to write the Results section

1. Point — tell the reader which table to look at, and what it compares.

2. Main finding — state the pattern first, then give one or two numbers as proof.

3. Exceptions — say where the pattern does not hold.

4. Explain — AI / CS venues only: one or two sentences on why.

Three steps to write the Results section

1. Point — tell the reader which table to look at, and what it compares.

2. Main finding — state the pattern first, then give one or two numbers as proof.

3. Exceptions — say where the pattern does not hold.

4. Explain — AI / CS venues only: one or two sentences on why.

💡 Steps 1–3 are for everyone. Medical journals stop after Step 3.
💡 Target length: one paragraph per table, usually 4–8 sentences.

Step 1 — Point

💡 One sentence. Which table, what is compared, on what data.

Step 1 — Point

💡 One sentence. Which table, what is compared, on what data.

❌ Starting with a number, before the reader knows where to look: "Our method achieves 86.73 F1.”

Step 1 — Point

💡 One sentence. Which table, what is compared, on what data.

❌ Starting with a number, before the reader knows where to look: "Our method achieves 86.73 F1.”

✅ "Table 2 compares our method with the baselines on the test sets of Dataset A and Dataset B.”

Step 1 — Point

💡 One sentence. Which table, what is compared, on what data.

❌ Starting with a number, before the reader knows where to look: "Our method achieves 86.73 F1.”

✅ "Table 2 compares our method with the baselines on the test sets of Dataset A and Dataset B.”

⚠️ Do not repeat the setup here. The datasets, metrics, and baselines are already in your Experimental Setup (Weeks 2 and 3).

Step 2 — Main finding: pattern first

	Dataset A		Dataset B	
Method	Acc ↑	F1 ↑	Acc ↑	F1 ↑
BERT	85.30	82.10	78.40	74.95
RoBERTa	86.92	84.05	81.52	76.02
Ours	88.40	86.73	81.22	78.60
Method	Acc ↑	F1 ↑	Acc ↑	F1 ↑

❌ Reading the table aloud:

"BERT gets 82.10 F1, RoBERTa gets 84.05, and our method gets 86.73. On Dataset B, BERT gets 74.95, RoBERTa gets 76.02, and ours gets 78.60."

✅ The pattern, then the proof:

"Our method achieves the best F1 on both datasets, improving over the strongest baseline, RoBERTa, by 2.68 points on Dataset A and 2.58 points on Dataset B."

Step 2 — Main finding: pattern first

	Dataset A		Dataset B	
Method	Acc ↑	F1 ↑	Acc ↑	F1 ↑
BERT	85.30	82.10	78.40	74.95
RoBERTa	86.92	84.05	81.52	76.02
Ours	88.40	86.73	81.22	78.60
Method	Acc ↑	F1 ↑	Acc ↑	F1 ↑

❌ Reading the table aloud:

"BERT gets 82.10 F1, RoBERTa gets 84.05, and our method gets 86.73. On Dataset B, BERT gets 74.95, RoBERTa gets 76.02, and ours gets 78.60."

✅ The pattern, then the proof:

"Our method achieves the best F1 on both datasets, improving over the strongest baseline, RoBERTa, by 2.68 points on Dataset A and 2.58 points on Dataset B."

⚠️ The reader can always look up the other numbers. **Your job is to choose which numbers matter.**

Step 2 — How to report a difference

- 💡 Compare with the strongest baseline, not the weakest. Beating BERT by 4.63 points sounds better, but reviewers will notice which row you picked.
- 💡 Say which baseline you are comparing with. "Improves by 2.68" — *over what?*
- ⚠️ No "significantly" unless you ran a significance test.
- ⚠️ Avoid vague words like "huge" or "dramatic". The number already says how big it is.

Step 3 — Exceptions

- Say where your method **does not work**. Just state it — no reason yet.

> The one exception is accuracy on Dataset B, where RoBERTa is 0.30 points higher.

Step 3 — Exceptions

- Say where your method **does not work**. Just state it — no reason yet.

> The one exception is accuracy on Dataset B, where RoBERTa is 0.30 points higher.

💡 Never skip the exception. One honest sentence makes the rest of your paragraph easier to trust.

Step 4 — Explain (AI venues only)

Woo, Sanghyun, et al. "Convnext v2: Co-designing and scaling convnets with masked autoencoders." CVPR 2023.

- After the results, one or two sentences on why they came out that way.

> Either way, we observe a significant performance degradation, suggesting that keeping GRN in both pre-training and fine-tuning is important.

Step 4 — Explain (AI venues only)

Woo, Sanghyun, et al. "Convnext v2: Co-designing and scaling convnets with masked autoencoders." CVPR 2023.

- After the results, one or two sentences on why they came out that way.

> Either way, we observe a significant performance degradation, suggesting that keeping GRN in both pre-training and fine-tuning is important.

💡 Write your findings in Discussion (depending on your target journals.)

More examples

Example 1 — a medical journal

Lin, Ying-Jia, et al. "Automating Injury Severity Score Calculation From Free-Text Discharge Diagnoses Using Domain-Specific Retrieval-Augmented Large Language Models: Retrospective Validation Study." JMIR Medical Informatics 14.1 (2026): e93922.

As shown in Table 2, DS-RAG yielded consistent performance gains for all models in both accuracy and F1-score. GPT-4.1 and Llama 4 achieved the largest F1-score improvements (+16.71 and +12.13), and GPT-4.1 remained the most accurate model (84.00% accuracy). … Standard RAG also improved accuracy and F1-score for GPT-4.1, Gemini 2.5 Pro, and Llama 4 Maverick, **but reduced both for Claude 4.5 Sonnet** …

Example 2 — an AI conference paper

Lin, Ying-Jia, et al. "Cfever: A chinese fact extraction and verification dataset." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 38. No. 17. 2024.

For sentence retrieval, there's a huge performance gap between our simple baseline and BEVERS. **This may result from the training with an additional ternary classification task and the employment of the re-retrieval technique in BEVERS**, whose score is also far ahead of the other baselines close to our sentence selection method on FEVER. However, BEVERS scored 94.41% in recall reported in their paper (DeHaven and Scott 2023) for FEVER, **showing that our dataset remains a challenge for BEVERS.**

Writing Results is like taking IELTS Task 1

Academic Writing Sample Task – 1B

WRITING TASK 1

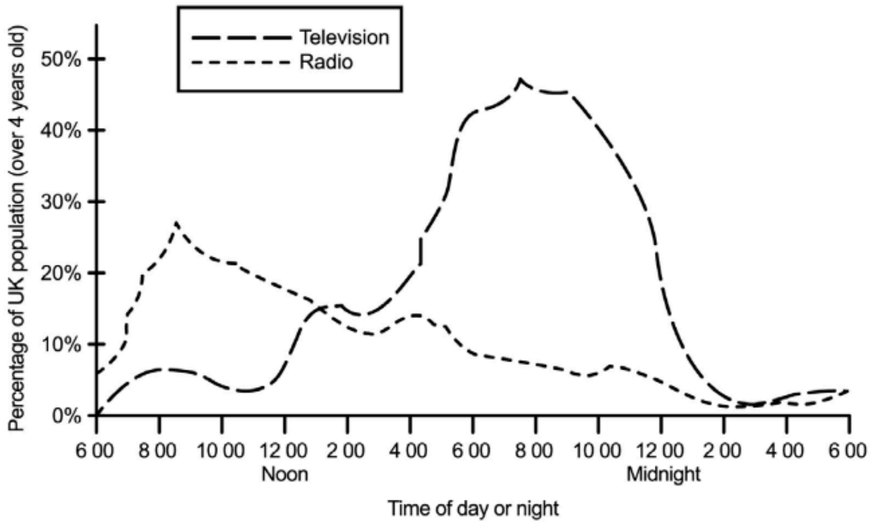
You should spend about 20 minutes on this task.

The graph below shows radio and television audiences throughout the day in 1992.

Summarise the information by selecting and reporting the main features, and make comparisons where relevant.

Write at least 150 words.

Radio and television audiences in UK, October – December 1992



https://ielts.org/cdn/Sample-tests/ielts-academic-writing-sample-tasks-2023.pdf

Common mistakes

1. Reading the table aloud — every number, no pattern
2. Comparing with the weakest baseline to make the gain look bigger
3. "Significantly" without a significance test
4. Percent instead of points for a difference between two scores
5. Skipping the rows where you lose
6. Numbers in the text that do not match the table — check every one
7. Explaining why for ten sentences — that is Discussion

Sentence Bank of Results

- "Table 2 compares our method with … on …"
- "Our method achieves the best … on all …, improving over the strongest baseline, …, by … points."
- "The one exception is …, where … is … points higher."
- (AI venues) "This suggests that …" / "We attribute this to …"
- "To see which component contributes most, we remove each one in turn (Table 3)."
- (Ablation study) "Removing … lowers … from … to …, showing that …"
- "… improves … despite at the cost of a small decrease in …"

Your Turn

NEXT WEEK ATTENTION!!

- Next week: **Results** section: Figures
  - Paragraph for a table
  - Caption for a table
- Please bring your preliminary results.

Results are not ready?

- If your experiments are still running, bring the table with empty cells and the row and column headings filled in. That is enough to work on.
- If you don’t have your results ready, please run small experiments this week.

Thank you for your participation 📚