Academic Paper Writing in Artificial Intelligence

Results - Figures (Week 9)

Course Website
Download slide here!

Instructor: Ying-Jia Lin
November 5th, 2026

Outline

- Week 8 Assignment Summary (15 min)
- What a Results section (with Figures) is (5 min)
- Results section in **3(4) steps** and examples (20 min)
- Captions (20 min)
- **Your Turn for revision and new assignment (90 min)**
  - Feel free to discuss with me !!

Week 8 Assignment Summary (1/3)

- **Every important claim in the Results section should be supported by the reported evidence.**
  - ❌ Our method consistently outperforms the baselines across all categories.
  - ✅ Our method achieves the highest Macro-F1 score of 0.89, outperforming the strongest baseline by 0.05.

Week 8 Assignment Summary (2/3)

- **The Results section should highlight key findings rather than simply repeat what is shown in tables.**
  - ❌ Baseline A achieved an accuracy of 85.2% and an F1 score of 0.81. Baseline B achieved 86.4% accuracy and 0.83 F1, while Baseline C achieved 87.1% accuracy and 0.84 F1. Our method achieved 90.3% accuracy and 0.89 F1.
  - ✅ Our method achieved the highest accuracy (90.3%) and F1 score (0.89), outperforming the strongest baseline by 3.2 percentage points and 0.05, respectively.

Week 8 Assignment Summary (3/3)

- **Do not overstate the evidence.**

Method	Macro-F1
Baseline	0.82
Ours (with attention)	0.89

- ❌ Our method achieved a higher macro-F1 because the attention mechanism successfully captured important features.
  - Problem: The results show a performance difference, but they do not establish why the improvement occurred.
- ✅ Our method (with attention) achieved a higher macro-F1 than the baseline (0.89 vs. 0.82), an improvement of 0.07.

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

Where we are (detailed)

Week	Topic	
7	Related Work	done
8	Results (I): Tables	done
9	Results (II): Figures and Captions	Today
10	Discussion	Next

[Recap] The Results section in illustration



Preparing a table? Or a figure?

- Main results are usually a table. Analysis results are often a figure.

Use a table when the reader needs …	Use a figure when the reader needs …
the exact numbers	the pattern: a trend, a gap, a crossing point
to compare a few methods on a few metrics	to see many points at once

Why sometimes using a figure is better

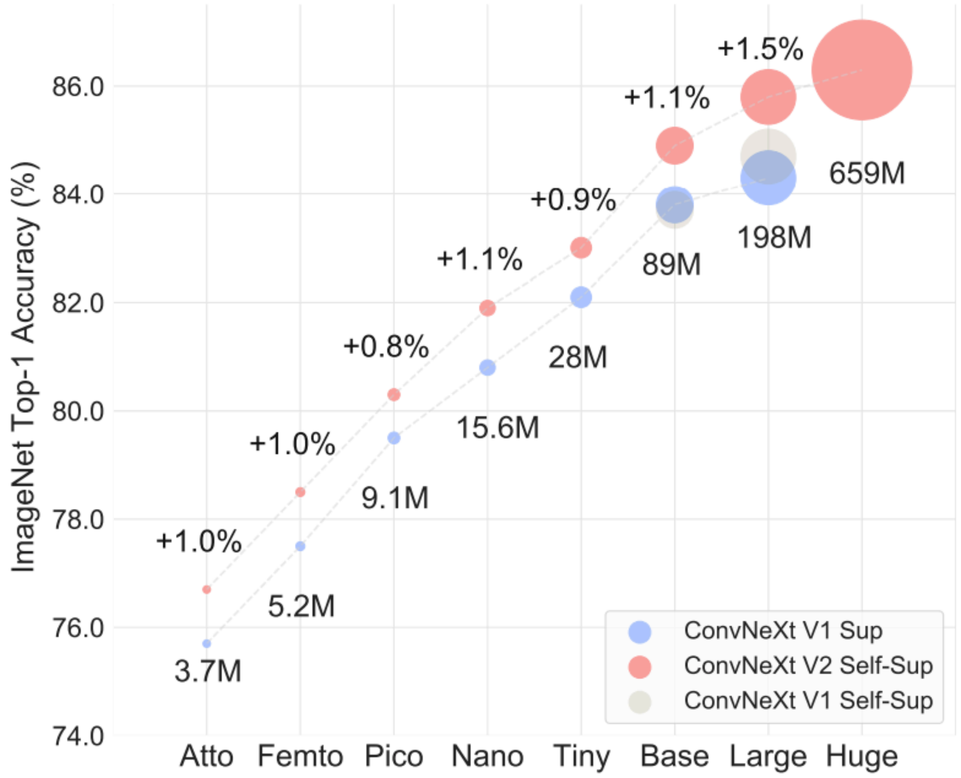


Figure 1. **ConvNeXt V2 model scaling**. The ConvNeXt V2 model, which has been pre-trained using our fully convolutional masked autoencoder framework, performs significantly better than the previous version across a wide range of model sizes.

- ConvNeXt V2 (CVPR 2023): eight model sizes. A table could hold these numbers, but only a figure shows at a glance that the gap stays at every size.

A Results section can have subsections

- 💡 Week 8: one subsection answers one question, with its own table or figure.

SCoRe (ICLR 2025)	*JMIR Med Inform (2026)*
6.1 Benchmark Results → Table 2	Model Accuracy Comparison → Fig. 2
6.2 Inference-Compute Scaling → Fig. 5	Performance in Different ISS Ranges → Fig. 3
6.3 Ablation Studies → Table 4	… with Tolerance Threshold → Fig. 4
	ISS Range Classification → Table 2

Three steps to write the Results section

1. Point — tell the reader which Figure (panel) to look at, and what it compares.

2. Main finding — state the pattern first, then give 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 Figure (panel) to look at, and what it compares.

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

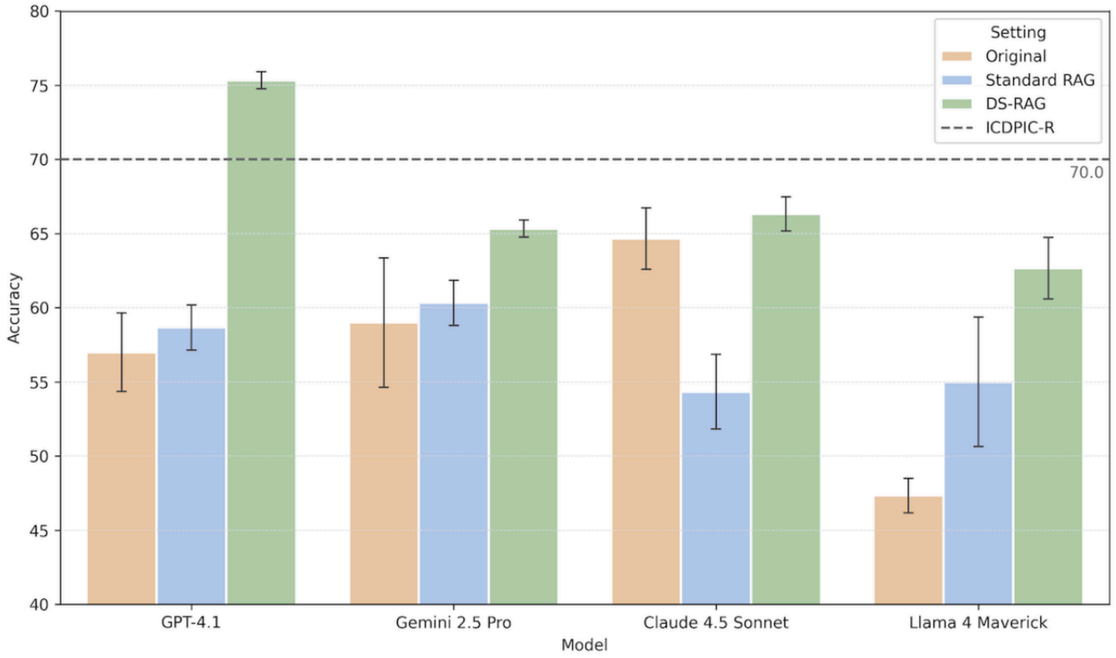
3. Exceptions — say where the pattern does not hold.

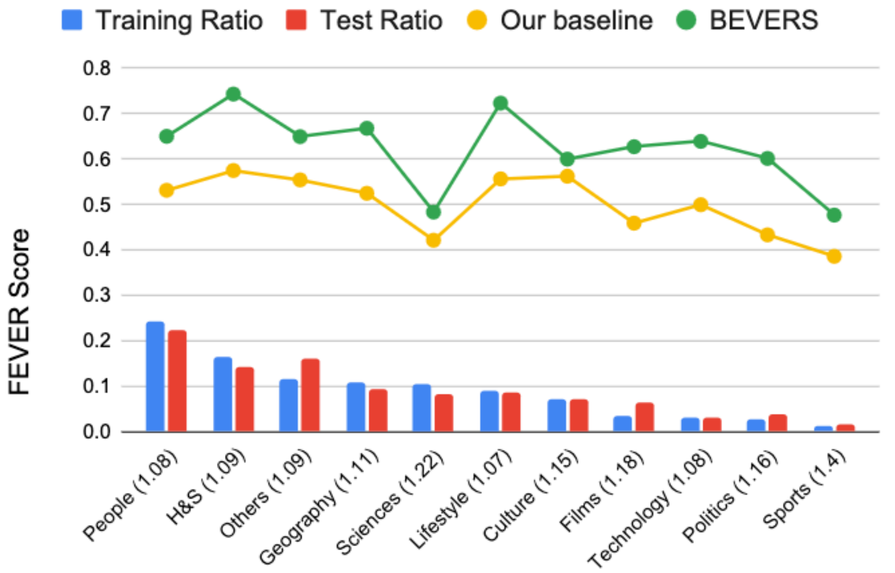
4. Explain — AI / CS venues only: one or two sentences on why.

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

What is a “pattern”?

- Pattern = what the reader sees at a glance:
  - which bar is highest; whether a line goes up.





Step 1 — Point

Many results figures have more than one panel.

Step 1 — Point

Many results figures have more than one panel.

❌ **(Figure 3 has four panels)** “As shown in Figure 3, …" — which of the four plots?

Step 1 — Point

Many results figures have more than one panel.

❌ **(Figure 3 has four panels)** “As shown in Figure 3, …" — which of the four plots?

✅ "Figure 3a shows accuracy on mild cases; Figure 3b shows severe cases."

Step 1 — Point

Many results figures have more than one panel.

❌ **(Figure 3 has four panels)** “As shown in Figure 3, …" — which of the four plots?

✅ "Figure 3a shows accuracy on mild cases; Figure 3b shows severe cases."

- Label panels (a), (b), (c) in the figure itself, not only in the caption.

Step 1 — Point

Many results figures have more than one panel.

❌ **(Figure 3 has four panels)** “As shown in Figure 3, …" — which of the four plots?

✅ "Figure 3a shows accuracy on mild cases; Figure 3b shows severe cases."

- Label panels (a), (b), (c) in the figure itself, not only in the caption.
- Refer to them in the same order as they appear.

Step 1 — Point

Many results figures have more than one panel.

❌ **(Figure 3 has four panels)** “As shown in Figure 3, …" — which of the four plots?

✅ "Figure 3a shows accuracy on mild cases; Figure 3b shows severe cases."

- Label panels (a), (b), (c) in the figure itself, not only in the caption.
- Refer to them in the same order as they appear.
- If you never mention a panel in the text, ask whether it should be there.

Step 2 — Main finding (pattern + numbers)

❌ Only the pattern:

"DS-RAG performs much better than the other settings."

❌ Only the numbers: a list of twelve values, one per bar.

✅ The pattern, then the numbers that support it:

"All models improved with DS-RAG. GPT-4.1 rose from 57.0% to 75.3%, an increase of 18.3 percentage points."

Example 1 (text)

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.

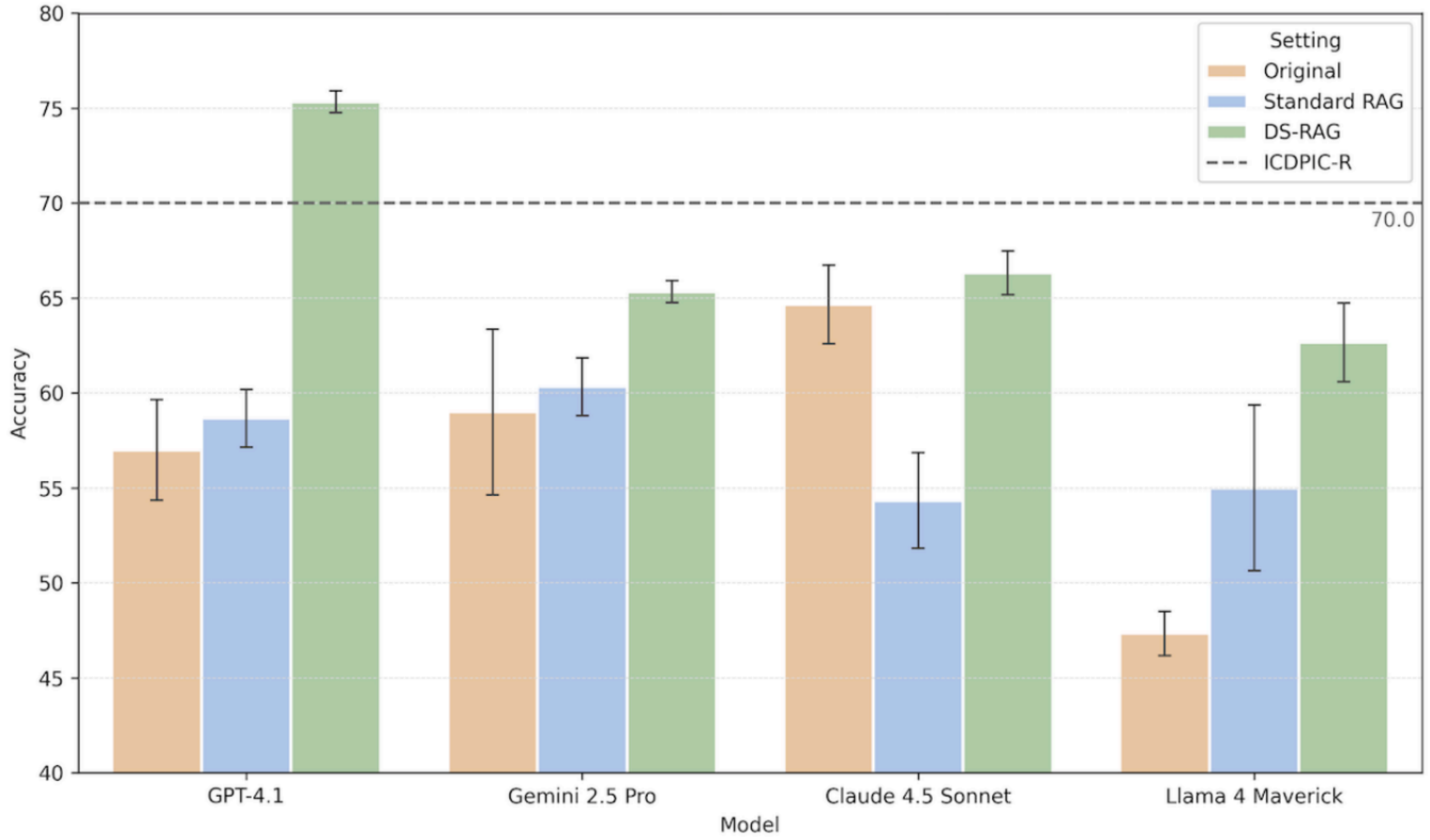
We present the results of the LLMs on the ISS calculation task in Figure 2. In the original setting (without RAG), models achieved modest EM accuracy. Among the LLMs, Claude 4.5 Sonnet performed the best (64.7%), followed by Gemini 2.5 Pro (59.0%) and GPT-4.1 (57.0%), with Llama 4 Maverick the lowest (47.3%).

Red: Pattern, Green: Numbers

Example 1 (figure)

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.

Figure 2. Exact-match accuracy of 4 large language models for automated Injury Severity Score calculation in a physician-verified test set of 100 trauma cases from Chang Gung Memorial Hospital, Linkou, Taiwan, in 2022. Exact match was defined as equality between the predicted ISS and physician-verified ISS. The 3 experimental settings were original (prompting only), standard retrieval-augmented generation (RAG) using the pretrained MedCPT retriever [18], and domain-specific RAG (DS-RAG) using a retriever fine-tuned on trauma-related International Classification of Diseases, Tenth Revision, Clinical Modification data (proposed). The horizontal dashed line indicates the International Classification of Diseases and R statistical software (ICDPIC-R; gem_max) baseline [24]. Bars represent mean accuracy, and error bars represent the SD over 3 runs.



Error bars (along with Step 2)

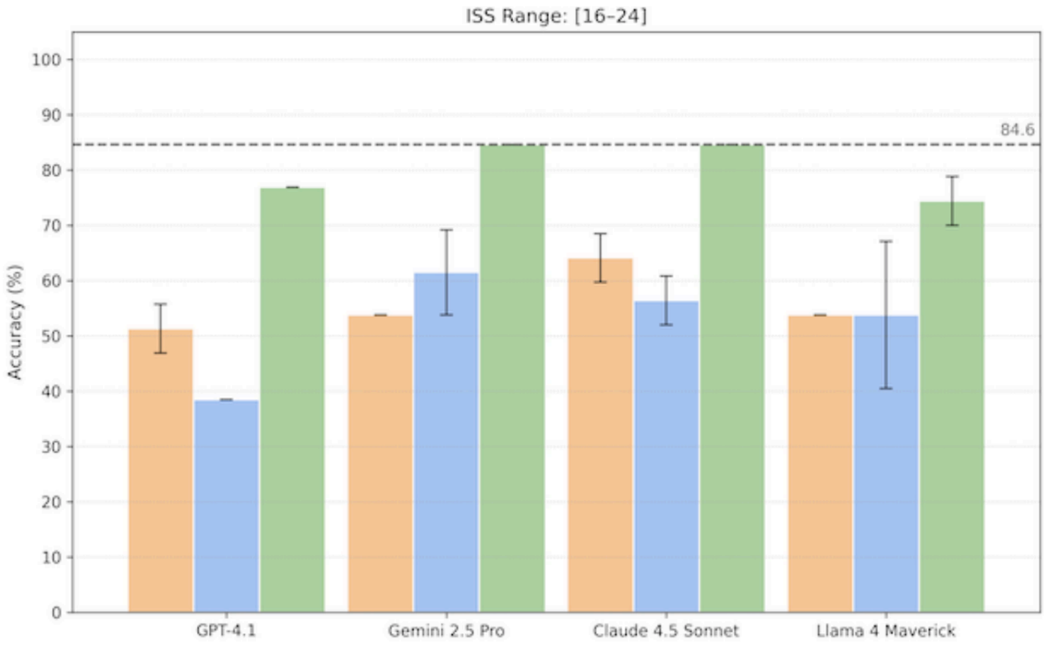
- Not a step, but error bars may limit your claims. ⚠️ Do claim too much!

Error bar	What it tells you
Standard deviation (SD)	how much the runs differ from each other
95% confidence interval (CI)	a range that very likely contains the true value (Week 8)

Example 2 (a large error bar)

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.

We note that Llama 4 Maverick showed a relatively large standard deviation across the three runs in the severe subgroup. This variability may reflect run-to-run differences in model predictions amplified by the small subgroup size (n=13).



What you can and cannot claim with error bars

- Overlapping error bars do not prove "no difference”. Separate error bars do not prove a real difference. *They are a warning sign, not a test.*
- To claim a difference, run a significance test and report the result.
- The p-value is the chance of seeing a difference this large if there were really no difference. Small p (e.g. < 0.05) → the difference is unlikely to be luck.

[Recap] 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.

[Recap] 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.

[Recap] 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.

[Recap] 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 explanations in Discussion (depending on your target journals.)

Sentence Bank (1/2)

Pointing to the figure (both are common in good papers; pick either)

- "As shown in Figure 2, …" / "Figure 2 shows that …"
- "… (Figure 2)." / "… (Figure 2b, green bars)."

Describing what the reader sees

- "The gap widens as …" / "The curve flattens after 10k steps."
- "The dashed line marks …; only … stays above it."
- "As k increases, recall rises from … to …, while precision falls from … to …"

Panels (figure)

- "Figure 3a shows …; Figure 3b shows …"

Sentence Bank (2/2)

Statistics and tests

- "The improvement over [baseline] is statistically significant (McNemar's test, p = …).”
- "This difference did not reach statistical significance (p = …).”
- "[Our method] reached …% (95% CI …–…%), compared with …% (95% CI …–…%) for [baseline].”
- "[Model] shows a large standard deviation across runs (± …), so this result should be read with caution."

Caption

[Week4] Three steps to write a (good) caption

1. What it is — a noun phrase, not a sentence.

> The proposed training framework for unsupervised Chinese word segmentation.

2. How to read it — name the things that readers have to work out.

> Green boxes are pre-trained models; orange boxes are the steps we introduce. A closed lock means the parameters are frozen.

3. What to notice — tell the hidden things or key takeaways (optional).

> The classifier predicts a 0/1 boundary label for every character.

Example 1 — IEEE journal

Liu, Chenxi, et al. "St-llm+: Graph enhanced spatio-temporal large language models for traffic prediction." IEEE Transactions on Knowledge and Data Engineering 37.8 (2025): 4846-4859.

TABLE III
MODEL COMPARISON ON TRAFFIC DATASETS IN TERMS OF MAE, RMSE, AND WAPE (%)

Models	NYCTaxi Pick-up			NYCTaxi Drop-off			CHBike Pick-up			CHBike Drop-off		
	MAE	RMSE	WAPE	MAE	RMSE	WAPE	MAE	RMSE	WAPE	MAE	RMSE	WAPE
DCRNN	5.40	9.71	20.43%	5.19	9.63	19.82%	2.09	3.30	42.26%	1.96	2.94	39.61%
STGCN	5.71	10.22	21.62%	5.38	9.60	20.55%	2.08	3.31	42.08%	2.01	3.07	40.62%
ASTGCN	7.43	13.84	28.04%	6.98	14.70	26.60%	2.76	4.45	55.71%	2.79	4.20	56.49%
GWN	5.43	<u>9.39</u>	20.55%	<u>5.03</u>	<u>8.78</u>	19.21%	2.04	3.20	40.95%	1.95	2.98	39.43%
AGCRN	5.79	10.11	21.93%	5.45	9.56	20.81%	2.16	3.46	43.69%	2.06	3.19	41.78%
GMAN	5.43	9.47	20.42%	5.09	8.95	19.33%	2.20	3.35	44.06%	2.09	3.00	42.00%
STSGCN	6.19	11.14	25.37%	5.62	10.21	22.59%	2.36	3.73	50.09%	2.73	4.50	54.10%
ASTGNN	5.90	10.71	22.32%	6.28	12.00	23.97%	2.37	3.67	47.81%	2.24	3.35	45.27%
STG-NCDE	6.24	11.25	23.46%	5.38	9.74	21.37%	2.15	3.97	61.38%	2.28	3.42	46.06%
DGCRN	5.44	9.82	20.58%	5.14	9.39	19.64%	2.06	3.21	41.51%	1.96	2.93	39.70%
OFA	5.82	10.42	22.00%	5.60	10.14	21.36%	2.06	3.21	41.70%	1.96	2.97	39.68%
GATGPT	5.92	10.55	22.39%	5.66	10.39	21.60%	2.07	3.23	41.70%	1.95	2.94	39.43%
GPTGAT	5.47	9.74	20.77%	5.58	10.23	20.85%	2.05	3.20	41.69%	1.94	2.92	39.37%
GCNGPT	6.58	12.23	24.88%	6.64	12.24	25.32%	2.37	3.80	47.66%	2.24	3.48	45.37%
LLaMA-2	5.35	9.48	20.27%	5.66	10.74	21.63%	2.10	3.37	42.49%	1.99	3.03	40.28%
ST-LLM	<u>5.29</u>	9.42	<u>20.03%</u>	5.07	9.07	<u>19.18%</u>	<u>1.99</u>	<u>3.08</u>	<u>40.19%</u>	<u>1.89</u>	<u>2.81</u>	<u>38.27%</u>
ST-LLM+	**5.18**	**8.98**	**19.60%**	**4.94**	**8.68**	**18.86%**	**1.98**	**3.05**	**40.11%**	**1.88**	**2.79**	**38.20%**

Results are averaged from all prediction time steps. The bold and underlined font shows the best and the second-best results, respectively.

Example 1 — IEEE journal

Liu, Chenxi, et al. "St-llm+: Graph enhanced spatio-temporal large language models for traffic prediction." IEEE Transactions on Knowledge and Data Engineering 37.8 (2025): 4846-4859.

TABLE III
MODEL COMPARISON ON TRAFFIC DATASETS IN TERMS OF MAE, RMSE, AND WAPE (%)

Models	NYCTaxi Pick-up			NYCTaxi Drop-off			CHBike Pick-up			CHBike Drop-off		
	MAE	RMSE	WAPE	MAE	RMSE	WAPE	MAE	RMSE	WAPE	MAE	RMSE	WAPE
DCRNN	5.40	9.71	20.43%	5.19	9.63	19.82%	2.09	3.30	42.26%	1.96	2.94	39.61%
STGCN	5.71	10.22	21.62%	5.38	9.60	20.55%	2.08	3.31	42.08%	2.01	3.07	40.62%
ASTGCN	7.43	13.84	28.04%	6.98	14.70	26.60%	2.76	4.45	55.71%	2.79	4.20	56.49%
GWN	5.43	9.39	20.55%	5.03	8.78	19.21%	2.04	3.20	40.95%	1.95	2.98	39.43%
AGCRN	5.79	10.11	21.93%	5.45	9.56	20.81%	2.16	3.46	43.69%	2.06	3.19	41.78%
GMAN	5.43	9.47	20.42%	5.09	8.95	19.33%	2.20	3.35	44.06%	2.09	3.00	42.00%
STSGCN	6.19	11.14	25.37%	5.62	10.21	22.59%	2.36	3.73	50.09%	2.73	4.50	54.10%
ASTGNN	5.90	10.71	22.32%	6.28	12.00	23.97%	2.37	3.67	47.81%	2.24	3.35	45.27%
STG-NCDE	6.24	11.25	23.46%	5.38	9.74	21.37%	2.15	3.97	61.38%	2.28	3.42	46.06%
DGCRN	5.44	9.82	20.58%	5.14	9.39	19.64%	2.06	3.21	41.51%	1.96	2.93	39.70%
OFA	5.82	10.42	22.00%	5.60	10.14	21.36%	2.06	3.21	41.70%	1.96	2.97	39.68%
GATGPT	5.92	10.55	22.39%	5.66	10.39	21.60%	2.07	3.23	41.70%	1.95	2.94	39.43%
GPTGAT	5.47	9.74	20.77%	5.58	10.23	20.85%	2.05	3.20	41.69%	1.94	2.92	39.37%
GCNGPT	6.58	12.23	24.88%	6.64	12.24	25.32%	2.37	3.80	47.66%	2.24	3.48	45.37%
LLaMA-2	5.35	9.48	20.27%	5.66	10.74	21.63%	2.10	3.37	42.49%	1.99	3.03	40.28%
ST-LLM	5.29	9.42	20.03%	5.07	9.07	19.18%	1.99	3.08	40.19%	1.89	2.81	38.27%
ST-LLM+	**5.18**	**8.98**	**19.60%**	**4.94**	**8.68**	**18.86%**	**1.98**	**3.05**	**40.11%**	**1.88**	**2.79**	**38.20%**

Results are averaged from all prediction time steps. The bold and underlined font shows the best and the second-best results, respectively.

Example 1 — IEEE journal

Liu, Chenxi, et al. "St-llm+: Graph enhanced spatio-temporal large language models for traffic prediction." IEEE Transactions on Knowledge and Data Engineering 37.8 (2025): 4846-4859.

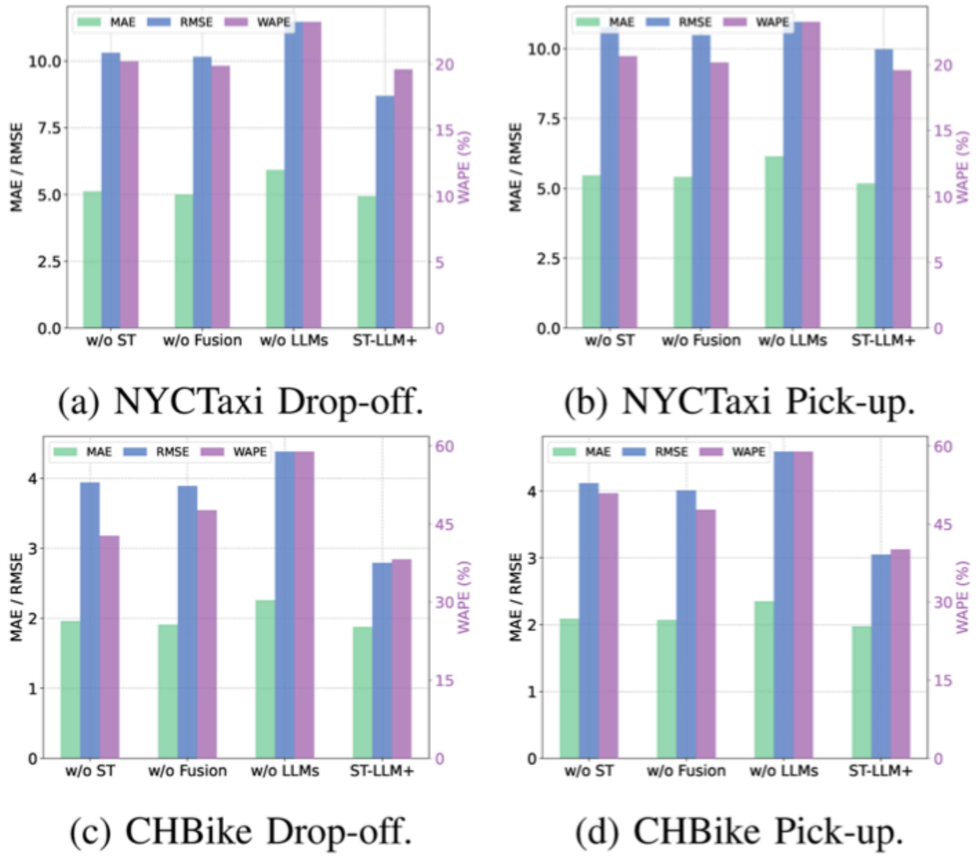


Fig. 2. Ablation study of ST-LLM+ key components.

Example 1 — IEEE journal

Liu, Chenxi, et al. "St-llm+: Graph enhanced spatio-temporal large language models for traffic prediction." IEEE Transactions on Knowledge and Data Engineering 37.8 (2025): 4846-4859.

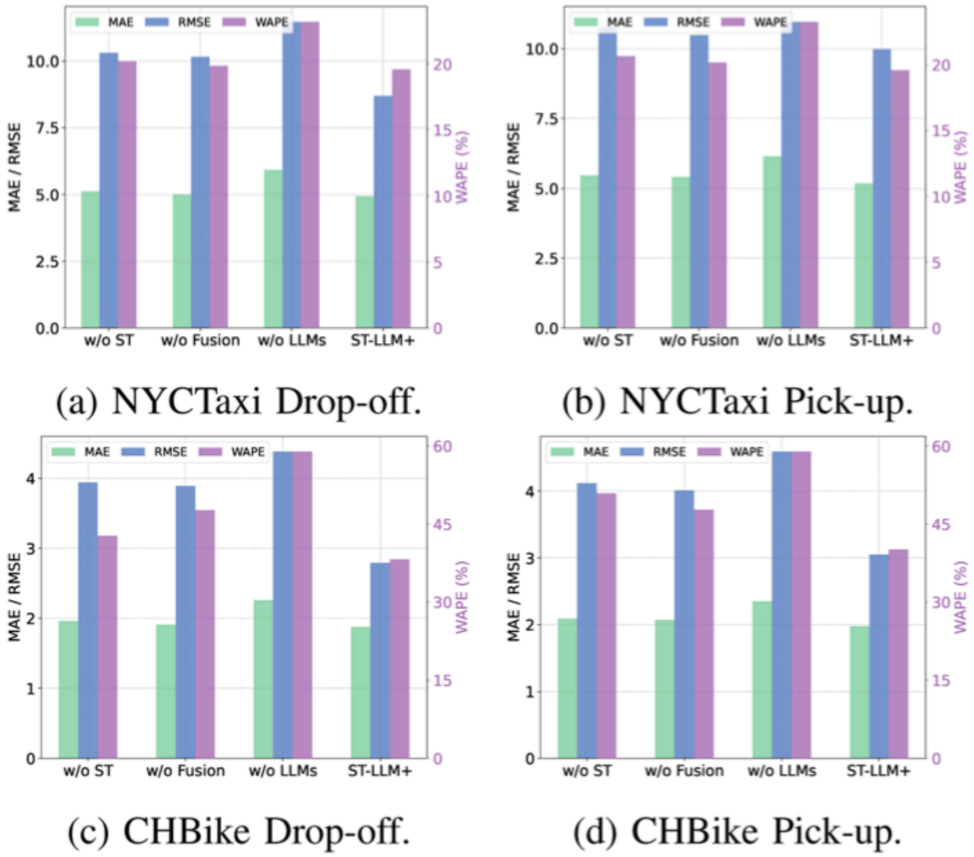
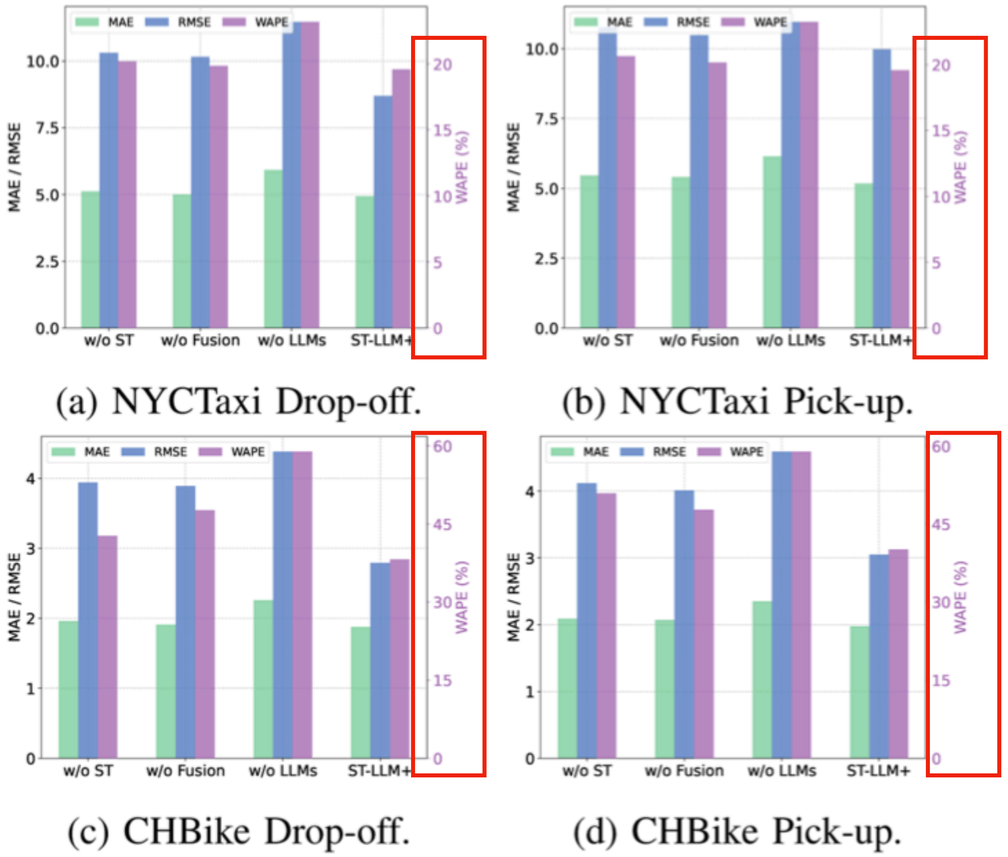


Fig. 2. Ablation study of ST-LLM+ key components.

Short caption (sometimes)

Example 1 — IEEE journal

Liu, Chenxi, et al. "St-llm+: Graph enhanced spatio-temporal large language models for traffic prediction." IEEE Transactions on Knowledge and Data Engineering 37.8 (2025): 4846-4859.



Double Y-axes

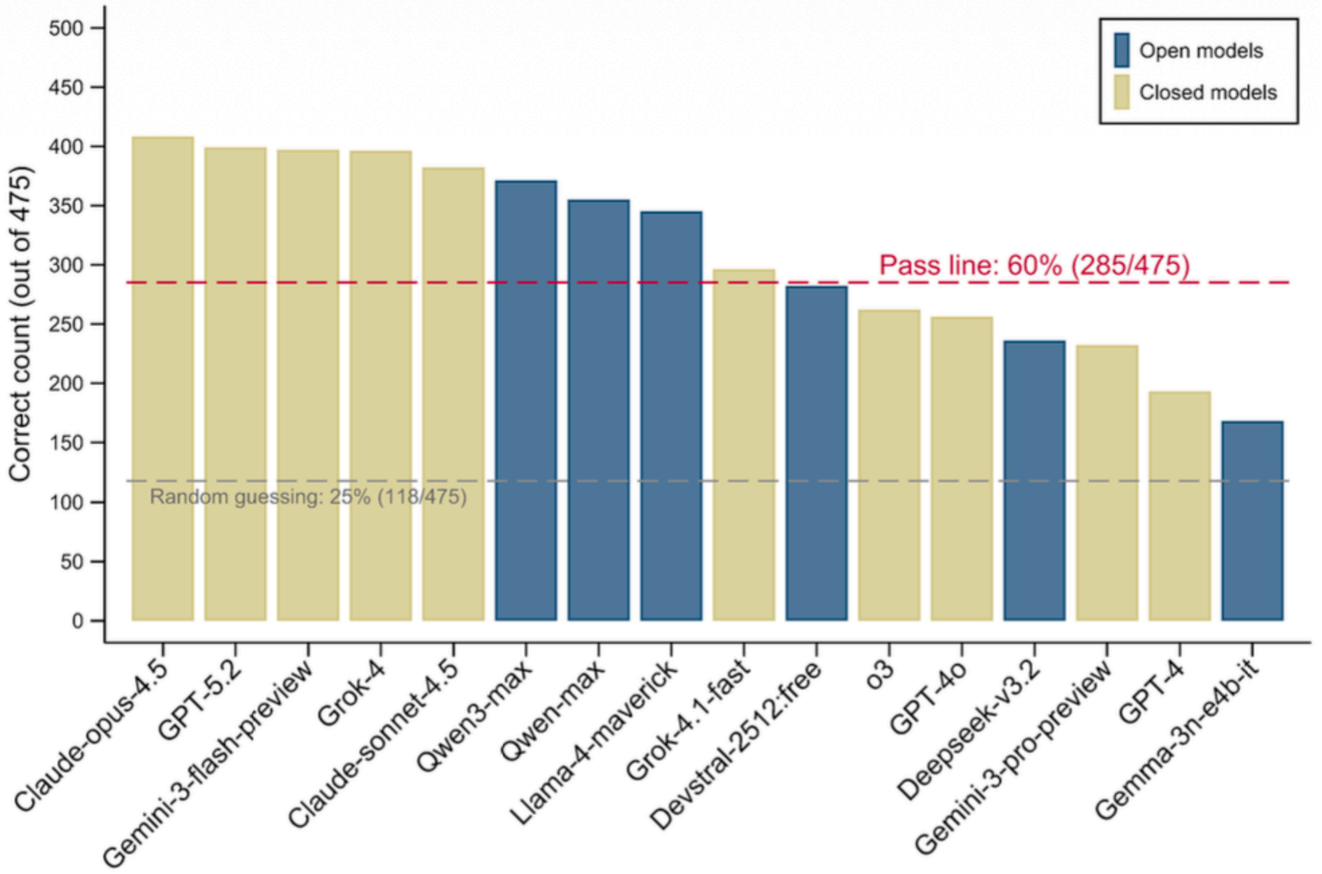
Fig. 2. Ablation study of ST-LLM+ key components.

Short caption (sometimes)

Example 2 - Medical Journal

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

Fig. 2 | Accuracy of 16 large language models under the pure zero-shot condition. The grey dashed line indicates chance-level accuracy for four-option multiple-choice questions (25%), and the red dashed line indicates the official passing threshold (60%; ≥285 of 475 correct responses). Blue bars indicate open models and yellow bars indicate closed models. Models are ranked from highest to lowest accuracy. Numerical data are provided in Supplementary Table 8a.



Example 3 - Medical Journal

Lee, Changsun, et al. "Read like a radiologist: Efficient vision-language model for 3d medical imaging interpretation." Medical Image Analysis (2026): 104077.

Table 1
Comparison of NLG metrics and CA metrics on the generated reports between MS-VLM and baseline methods.

Method	NLG metrics			CA metrics		
	BLEU-4	ROUGE-L	METEOR	Precision	Recall	F1 score
3D-CT-GPT	0.133	0.125	0.148	<u>0.239</u>	0.195	0.190
Merlin	**0.279**	0.431	**0.437**	0.202	0.223	0.214
CT2Rep (3D encoder)	<u>0.252</u>	<u>0.432</u>	0.328	**0.355**	0.132	0.175
MS-VLM (3D encoder)	0.184	0.364	0.331	0.193	0.231	0.207
MS-VLM (w/o VQA)	0.189	0.374	0.313	0.192	<u>0.298</u>	<u>0.232</u>
MS-VLM (proposed)	0.232	**0.438**	<u>0.396</u>	0.222	**0.329**	**0.261**

The best performance is highlighted in bold, and the second-best is underlined. CA metrics are averaged across 18 chest-related abnormalities.

Example 4 - 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.

Table 2. Accuracy and macroaveraged F1-score for classification of Injury Severity Score (ISS) ranges in our physician-verified test set of 100 trauma cases from Chang Gung Memorial Hospital, Linkou, Taiwan, in 2022. Cases were classified as mild (ISS 1-8), moderate (ISS 9-15), severe (ISS 16-24), or profound (ISS 25-75). Original denotes prompting without retrieval-augmented generation (RAG), standard RAG uses the pretrained MedCPT retriever [18], and domain-specific RAG (DS-RAG) uses a retriever fine-tuned on trauma-related ICD-10-CM (International Classification of Diseases, Tenth Revision, Clinical Modification) data. Values indicate the mean accuracy over 3 runs. International Classification of Diseases and R statistical software (ICDPIC-R; gem_max) [24] is a deterministic ICD-10-CM–based tool and is reported as a single reference value without SD.

Metric and method	GPT-4.1, mean (SD)	Gemini 2.5 Pro, mean (SD)	Claude 4.5 Sonnet, mean (SD)	Llama 4 Maverick, mean (SD)	ICDPIC-R
Accuracy (%)					80.00
Original	66.33 (1.15)	69.00 (4.36)	74.33 (1.53)	59.00 (1.73)	
Standard RAG	70.67 (1.53)	73.67 (2.52)	67.00 (1.00)	70.33 (2.52)	
DS-RAG	84.00 (1.00)	76.00 (0.00)	76.00 (2.65)	74.33 (1.53)	
F_1-score (%)					73.32
Original	67.37 (1.58)	67.84 (3.68)	73.39 (1.81)	62.72 (2.06)	
Standard RAG	68.31 (1.73)	75.52 (3.50)	67.29 (1.25)	69.76 (4.09)	
DS-RAG	84.08 (1.03)	78.39 (0.03)	79.41 (2.71)	74.85 (1.12)	

Example 5 - AI Conference

Kumar, Aviral, et al. "Training language models to self-correct via reinforcement learning." ICLR 2025.

💡 Two panels, one claim each.

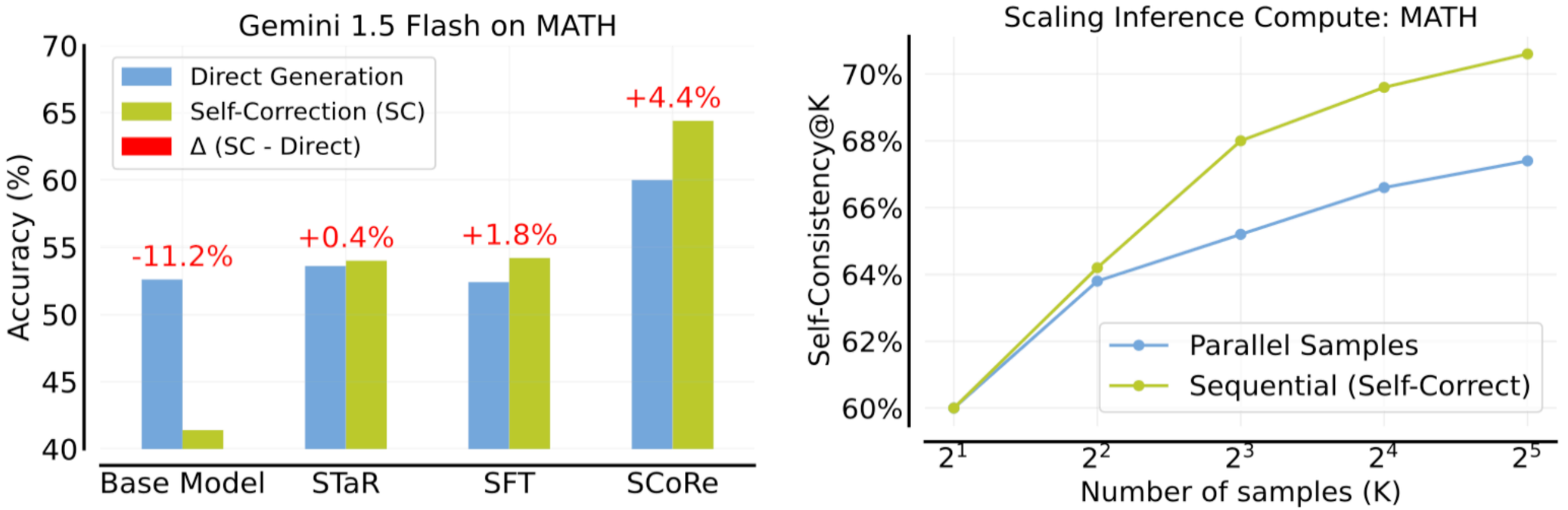


Figure 1: **Left:** ***SCoRe*** achieves state-of-the-art self-correction performance on MATH; **Right:** ***SCoRe*** inference-time scaling: spending samples on *sequential* self-correction becomes more effective than only on *parallel* direct samples (Section 6.2).

Example 6 - AI Conference

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

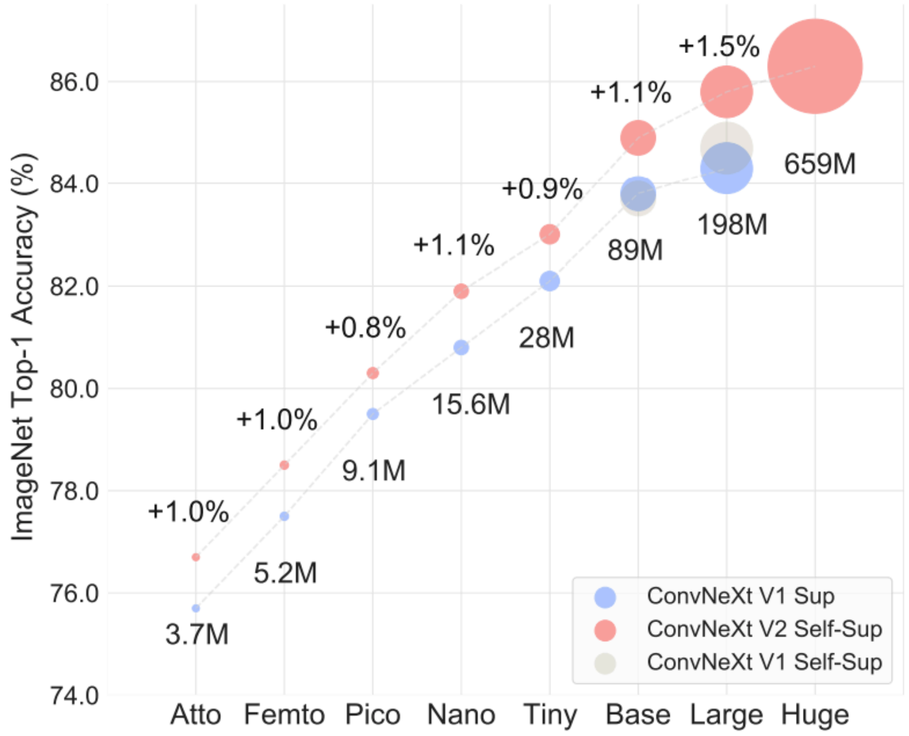


Figure 1. **ConvNeXt V2 model scaling**. The ConvNeXt V2 model, which has been pre-trained using our fully convolutional masked autoencoder framework, performs significantly better than the previous version across a wide range of model sizes.

⚠️ Look back at the ConvNeXt V2 figure from earlier today. The caption never says:

- what the size of each circle means, or what "3.7M … 659M" are
- what the "+1.0%" labels compare, and whether they are points or percent
- which test supports "significantly"

Common mistakes

*Omit this item if your journal's captions are short

[Captions]

1. **"Results of our method."** — of what, on what data, measured how?
2. Marks with no explanation — bold, stars, dashed lines, circle sizes*
3. Error bars without saying SD, SE, or CI, or how many runs
4. "Significantly" without a significance test
5. The caption repeats the paragraph — the caption says **how to read**; the text **says what it shows**

Your Turn

NEXT WEEK ATTENTION!!

- Next week: **Discussion** section
- Please bring:
  - your Results paragraphs from **Weeks 8 and 9**
  - every "why" you wanted to write but held back — especially if you are aiming at a medical journal
  - two or three papers that report results similar to or different from yours

Thank you for your participation 📚