Shared task: Lay summary

Hugging Face

Team: Bossy Beaver, UBC MDS-CL

Bossy_Beaver: Team members

- Hui Yin Lam
- Hayden Chiu
- Minh Nguyen
- Minzhi Huang
- Tushar Choudhary

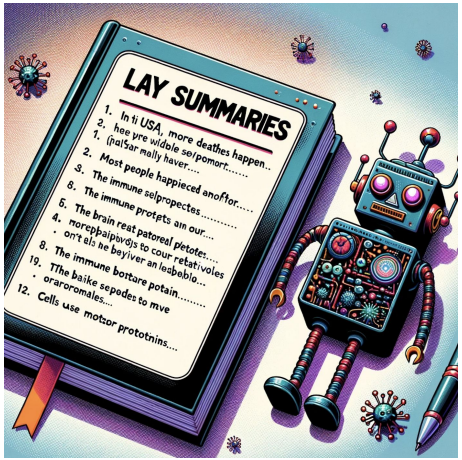
Supervisor / Mentor:

- Jian Zhu

Motivation of Lay summarization

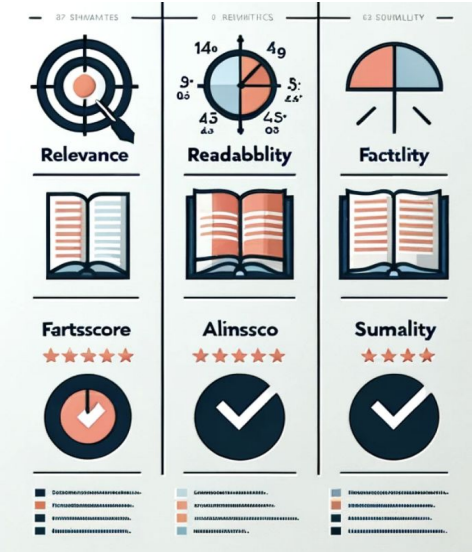
- To simplify and summarize medical research papers
- Make research works simple and easy to understand
- Should retain important facts
- General public can understand and appreciate scientific research results

Data



- 2 medical datasets: eLife and PLOS
- PLOS is much larger (about 24k samples)
- eLife is compact (about 4k samples)
- Contains paper text, headings, keywords and gold lay summary by human experts

Evaluation Metrics



Relevance

- ROUGE (1, 2, and L)
- BERTScore

Factuality

- AlignScore
- SummaC

Readability

- Flesch-Kincaid Grade Level (FKGL)
- Dale-Chall Readability Score (DCRS)
- Coleman-Liau Index (CLI)
- LENS

Methods and Result

- Character Limit Abstract Baseline: first 1000 characters
  - Very high factuality, low readability
- T5-based medical model: T5 transformer model
  - limited input length, bad quality

LLMs: Mistral variants, input length up to 32k tokens

- Mistral 7B: pretrained generative text instruct model
- BioMistral 7B: biomedical pretrained Mistral
- Mixtral 8x7B: combination of 8 models

Methods and Result

[INST]
Simplify and summarize in 200 to 300 words, put answer in 1 paragraph, keep important and factual details: "*article*"
[\INST]

LLMs - Zero-shot

- Mistral 7B : 3 second/sample using colab A100
- BioMistral 7B : 3 second/sample using colab A100
- Mixtral 8x7B : 90 second/sample using colab A100
- Mixtral 8x7B over API: fast but limited input length

Results

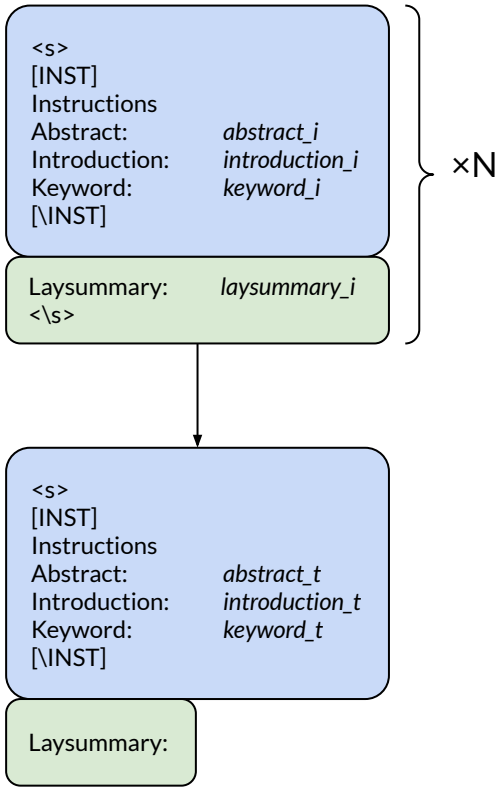
Better ----------> *Worse*

Relevance: Mixtral 8x7B > Mistral 7B > BioMistral 7B

Readability: BioMistral 7B < Mistral 7B < Mixtral 8x7B

Factuality: Mixtral 8x7B > Mistral 7B > BioMistral 7B

Methods and Result



LLMs - Few-shot

- 3 shots: 2.3 second/sample using colab A100
- 5 shots: 3.2 second/sample using colab A100

Results

Better ----------> *Worse*

Relevance:	BioMistral 7B > Mistral 7B 5shot > 3shot
Readability:	BioMistral 7B < Mistral 7B 5shot < 3shot
Factuality:	Mistral 7B > BioMistral 7B 3 shot > 5 shot

Discussion and Achievements

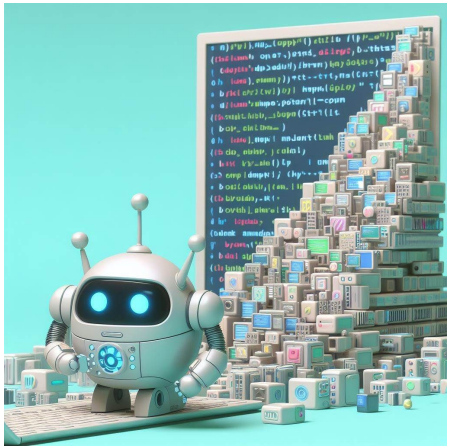
Best model (efficiency AND quality):

BioMistral 7B with 5shot prompting

Technical achievements:

- Experimented with various models and prompt strategies
- Worked with LLMs in Colab, cloud APIs
- Built self-contained solutions if required

Challenges



- Medical terms and knowledge is difficult to assess
- Article text is long (~ 30k tokens), far exceeding most transformers limit (512/1024)
- The GPU power required is HUGE.
- A100 GPU can take up to 90 sec / article
- Hard to fine-tune due to the metric used

TL;DR - Summary of our project

- Summarize and simplify medical papers for normal readers
- Data source: eLife (~4k rows) and PLOS (~24k rows)
- **Very challenging** in terms of domain technicality and GPU computational resources
- The best model is **BioMistral 7B** with few-shot
- Beat the baseline on several scores
- Very rewarding experience

Future

- Try llama 3 and mixtral 8x22b
- Fine tuning the LLM with train dataset

Thank you!

Thank you!

Thank you!

Thank you!

Thank you!

We finished!

We finished!

We finished!

We finished!

(AI-generated arts)

Appendix

Data

	lay_summary	article	headings	keywords	id
0	In the USA , more deaths happen in the winter ...	In temperate climates , winter deaths exceed s...	[Abstract, Introduction, Results, Discussion, ...	[epidemiology and global health]	elife-35500-v1
1	Most people have likely experienced the discom...	Whether complement dysregulation directly cont...	[Abstract, Introduction, Results, Discussion, ...	[microbiology and infectious disease, immunolo...	elife-48378-v2
2	The immune system protects an individual from ...	Variation in the presentation of hereditary im...	[Abstract, Introduction, Results, Discussion, ...	[microbiology and infectious disease, immunolo...	elife-04494-v1
3	The brain adapts to control our behavior in di...	Rapid and flexible interpretation of conflicti...	[Abstract, Introduction, Results, Discussion, ...	[neuroscience]	elife-12352-v2

Split	No. of Samples	Mean Length	Max Length	Min Length
eLife_train (article)	4346	10159	28308	322
eLife_train (summary)	4346	382	686	177
eLife_val (article)	241	9989	23050	3393
eLife_val (summary)	241	389	672	234
eLife_test (article)	142	8911	16684	2496
PLOS_train (article)	24773	6750	26647	750
PLOS_train (summary)	24773	194	511	4
PLOS_val (article)	1376	6738	20394	755
PLOS_val (summary)	1376	194	284	55
PLOS_test (article)	142	6943	18481	1590

Table 1: Descriptive Statistics of the Official Dataset

Mini Val Set vs Full Val Set

- A mini val set of 10% of the full val set was extract for evaluation
- Comparing using the Mixtral8x7b model with zero-shot, they showed similar results (maximum ±4%)

Model (set)	R1	R2	RL	BScore	FKGL	DCRS	CLI	LENS	AScore	SummaC
Mixtral, mini dev set	0.4238	0.1175	0.3922	0.8426	16.0048	10.8478	17.0260	54.0225	0.8333	0.6557
Mixtral, full dev set	0.4130	0.1172	0.3784	0.8439	15.6589	11.1420	16.9323	56.0474	0.8212	0.6399

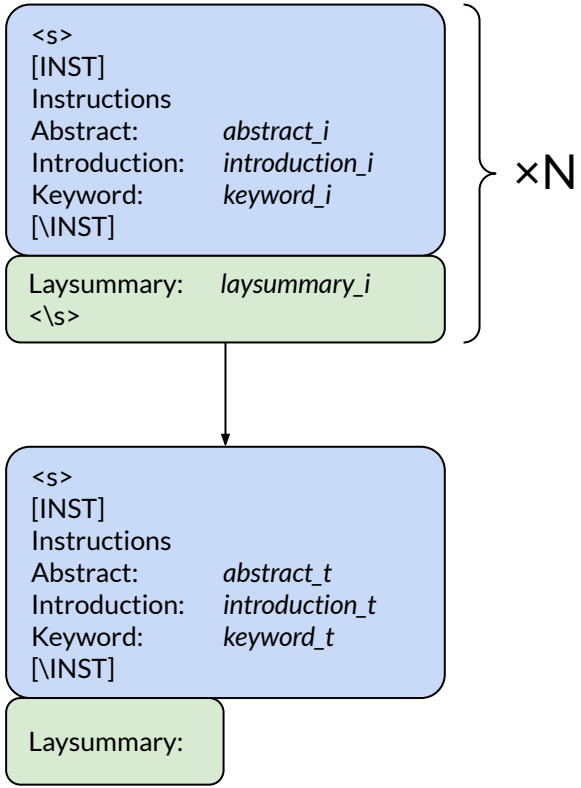
Table 4: Mixtral Model, mini dev vs. full dev set

Zero-shot on Mistral7B, BioMistral7b, Mixtral8x7B

```
sampling_params = SamplingParams(temperature=0.8, top_p=0.05, max_tokens=1024)
prompt = f"[INST] Simplify and summarize in 200 to 300 words, put answer in 1 paragraph:  {data_lst[0]} [/INST]"
```

Model (set)	R1	R2	RL	BScore	FKGL	DCRS	CLI	LENS	AScore	SummaC
Mistral 7B mini dev	0.395	0.106	0.361	0.838	15.766	10.909	16.943	54.659	0.751	0.577
Mistral 7B test	0.397	0.112	0.361	0.841	15.615	10.860	17.285	55.210	0.781	0.560
BioMistral mini dev	0.218	0.042	0.206	0.802	13.37	7.591	11.73	22.73	0.594	0.524
BioMistral test	0.307	0.072	0.282	0.826	13.32	8.593	14.24	55.83	0.6186	0.440
Mixtral 8x7B test	0.420	0.119	0.386	0.845	15.752	10.967	17.185	57.222	0.8445	0.662
Mixtral, mini dev set	0.4238	0.1175	0.3922	0.8426	16.0048	10.8478	17.0260	54.0225	0.8333	0.6557
Mixtral, full dev set	0.4130	0.1172	0.3784	0.8439	15.6589	11.1420	16.9323	56.0474	0.8212	0.6399

Few-shot on Mistral7B, BioMistral7b



Model (shots)	R1	R2	RL	BScore	FKGL	DCRS	CLI	LENS	AScore	Summa(
Mistral (3shot) with-out tag	0.3520	0.0813	0.3207	0.8278	15.3336	11.0748	16.8215	60.4612	0.5725	0.5011
Mistral (5shot) with-out tag	0.3265	0.0726	0.2980	0.8225	15.0964	11.0395	16.8835	59.5380	0.5618	0.4964
Biomistral (3shot) with-out tag	0.3019	0.0696	0.2802	0.8260	14.4768	9.7798	14.7836	49.7806	0.6334	0.5879
Biomistral (5shot) with-out tag	0.3249	0.0790	0.3020	0.8270	14.2461	9.8423	14.7869	54.9962	0.6017	0.5740
Biomistral (3shot) with tag	0.4159	0.1256	0.3827	0.8500	14.1066	10.7583	15.6107	62.3642	0.8134	0.7483
Biomistral (5shot) with tag	0.4230	0.1299	0.3914	0.8509	13.7772	10.6360	15.1731	62.9318	0.7813	0.7281

Table 7: Comparison Between Number of shots

Bad generated examples (Biomistral, 5shot, no tag)

Repeated texts

We recently demonstrated that HIV infection causes significant bone loss and osteorarylarysis in the transgenic rat model of HIV infection . We show that the HIV transgenic rat exhibits a significant increase in B cell expression of the anti-osteocrarygerraphy factor (RANKH) , a key osteoclarygerraphy regulator , and a significant decrease in expression of the pro-osteoclarygerraphy factor (OPG) , a key osteproteghry regulator . These data suggest that the HIV transgenic rat exhibits a significant increase in osteoclarygerraphy and a significant decrease in osteproteghry. We also show that the HIV transgenic rat exhibits a significant increase in osteoclarygenesis , a process that is mediated by the RANKH/OPG ratio. These data suggest that the HIV transgenic rat exhibits a significant increase in osteoclarygenesis. We also show that the HIV transgenic rat exhibits a significant increase in osteoclarygenesis , a process that is mediated by the RANKH/OPG ratio. These data suggest that the HIV transgenic rat exhibits a significant increase in osteoclarygenesis. We also show that the HIV transgenic rat exhibits a significant increase in osteoclarygenesis , a process that is mediated by the RANKH/OPG ratio. These data suggest that the HIV transgenic rat exhibits a significant increase in osteoclarygenesis. We also show that the HIV transgenic rat exhibits a significant increase in osteoclarygenesis , a process that is mediated by the RANKH/OPG ratio. These data suggest that the HIV transgenic rat exhibits a significant increase in osteoclarygenesis. We also show that the HIV transgenic rat exhibits a significant increase in osteoclarygenesis , a process that is mediated by the RANKH/OPG ratio. These data suggest that the HIV transgenic rat exhibits a significant increase in osteoclarygenesis. ……

Bad generated examples (Biomistral, 5shot, no tag)

Continuing numberings

..... Rta is implicated in KS pathogenesis , and its expression is regulated by a chromatin insulator protein CTCF [15] , [16] , [17] , [18] , [19] . CTCF is an 11-Zn finger DNA binding protein that binds to most of the characterized chromatin insulator elements in vertebrates [17] , [18] , [19] . CTCF can repress or activate transcription [20] , [21] , [22] , prevent the spread of DNA methylation [21] and histone modifications [23] , and block the interactions between transcriptional enhancers and promoters [26] , [27] . Genome-wide mapping studies have found that CTCF colocalizes with certain histone modifications (e . g H3K4me3) and histone variants (e . g H2AZ) , as well as with cohesins at a high-percentage of binding sites [30] , [31] , [32] , [33] , [34] . Cohesins have a well-established role in mediating sister-chromatid cohesion [35] , [36] , and have also been implicated in developmental gene regulation [37] . Heritable mutations in human SMC1 and SMC3 result in a spectrum of developmental disorders collectively referred to as cohesinopathies , which include Cornelia de Langue syndrome [38] , [39] .....