Capstone P1

MULTI-LEVEL SUMMARIZATION OF INDIVIDUAL RESEARCH PAPERS

Niloofar Karimi, Rimsha Kayastha, Yanzheng Liu

PROBLEM STATEMENT

Research papers are getting more and more complex. According to a study, 22% of abstracts published in 2015 had scores below 0 on the Flesch Reading Ease scale, meaning they exceeded college graduate reading level, compared to 14% in 1881. Research dissemination focused on peer-to-peer communication within academia rather than public outreach.

WHY IS SIMPLIFYING RESEARCH PAPERS IMPORTANT?

Awareness about new biomedical advancements for stakeholders

Fewer manual annotations and summarizations

Encouragement of interdisciplinary research

Inclusion opportunities for independent researchers

PROJECT GOAL

Ensuring:

- Reading level-wise simplification
- Domain-specific semantic simliarity

METHODOLOGY PLAN

Model & Approach

- Use BART (Bidirectional and Auto-Regressive Transformers) for abstractive multi-level summarization.
- Data collator for padding
- Fine-tune on general and medical texts to generate summaries for general and expert audiences.
- Control summary detail through prompt design, decoding strategies, and fine-tuning objectives.

Evaluation

- Automatic metrics: ROUGE, BERTScore for fidelity and semantic similarity.
- Readability metrics: Flesch Reading Ease, Flesch–Kincaid, SMOG, Dale–Chall ... for audience alignment.
- Statistical tests: ANOVA and regression to assess readability differences and quality relationships.

Implementation details

- Python pipeline using pandas, spaCy, scikit-learn, and Hugging Face Transformers.
- Learning rate: 3e-5

DATASETS AVAILABLE

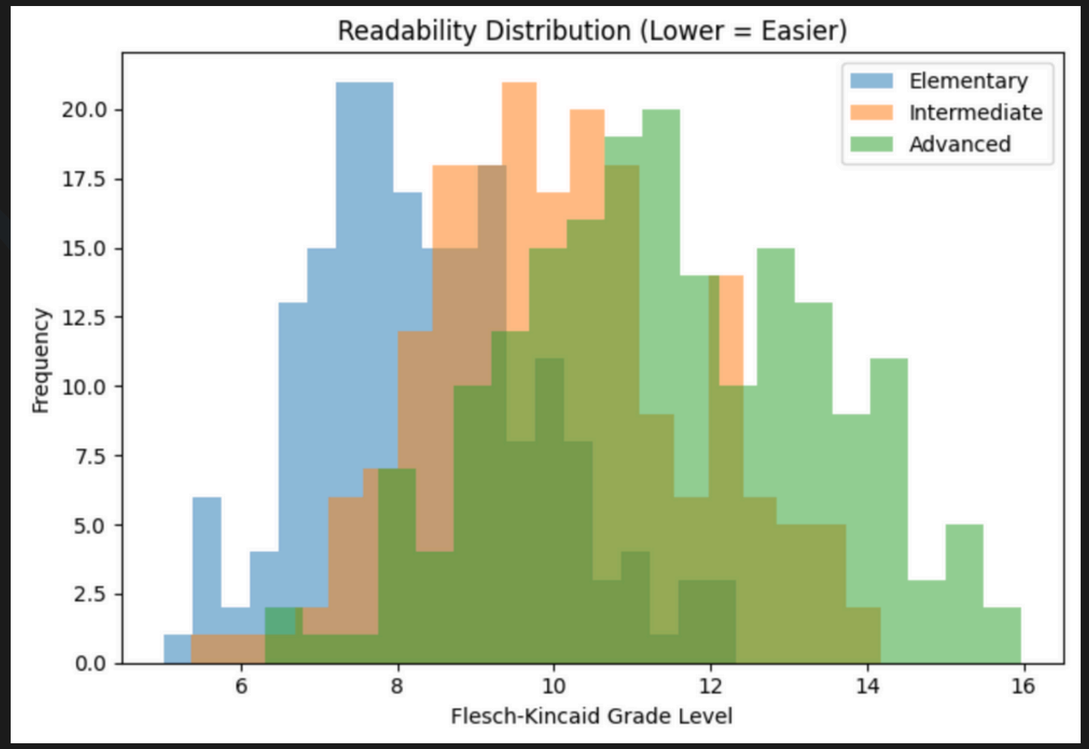
PREMINILARY RESULTS:

EDA

1. OneStopEnglish Dataset:

- Word count increases from **~534 to ~825**, unique words from **~279 to ~434**, and average sentence length from **~22 to ~29**, confirming complexity scaling.
- Elementary texts are easiest, grade levels rising progressively through intermediate and advanced. (Fig. 1).
- Readability metrics: **FKGL = 8.4 → 11.3, SMOG = 10.1 → 13.0**, showing rising linguistic difficulty.
- **Cosine similarity ≈ 0.7–0.8** across levels, meaning preserved despite higher complexity.

Fig. 1 : Readability Distribution across Levels



DATASETS AVAILABLE

PREMINILARY RESULTS:

3. M3 Biomedical Dataset:

The M3 dataset includes biomedical papers with multi-level summaries for laypersons, students, and experts, serving as the basis for our second-phase domain-specific summarization experiments.

- ~4,000 biomedical articles with three-level human summaries for multi-level abstractive summarization.
- Captures readability shifts from lay (FKGL ≈ 8–10) to expert (≈ 14–16) audiences.
- Will fine-tune BART on scientific text and compare generalization with OneStopEnglish and CLEAR.

PREPROCESSING:

- **Data Cleaning & Alignment:** Verified **189** complete triplets across elementary, intermediate, and advanced levels.
- **Pair Construction**: Created **567** paired examples (adv→inter, adv→elem, inter→elem).
- **Control Tokens Added:** Added <TO_INTER> and <TO_ELEM> to guide target reading level.
- **Group-Aware Split:** **480** training pairs and **87** validation pairs; ensured no article overlap.
- **Tokenization Check:** Over **95%** of samples fit within the **1024-token** limit.
- **Readability Validation:** Clear separation among levels, confirming proper complexity scaling.

RESULTS

TRAINING:

Fig .2: Validation Loss: Dropped from 4.13 to 2.30, confirming consistent model improvement.

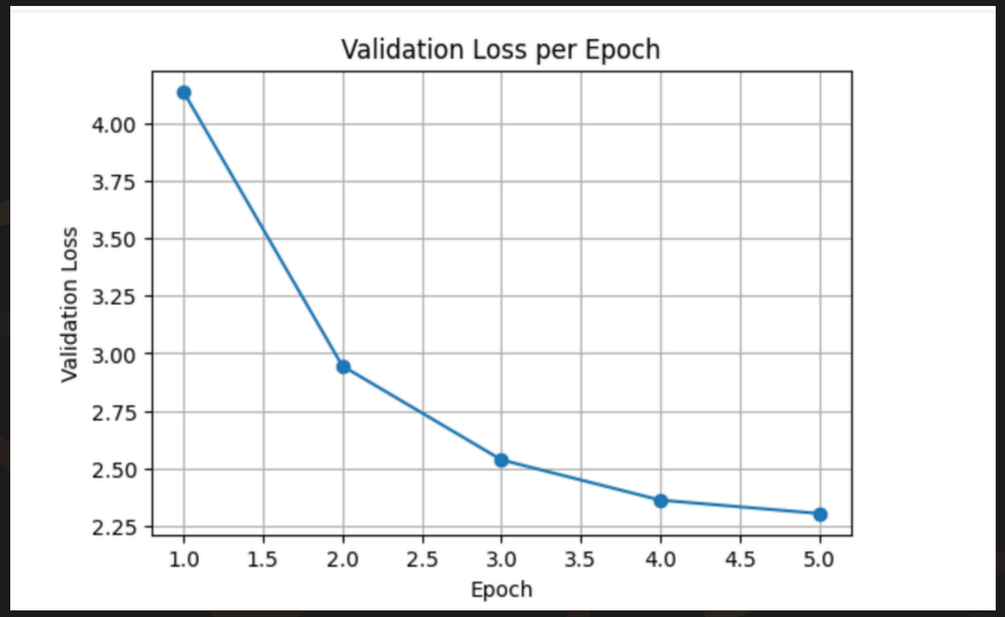
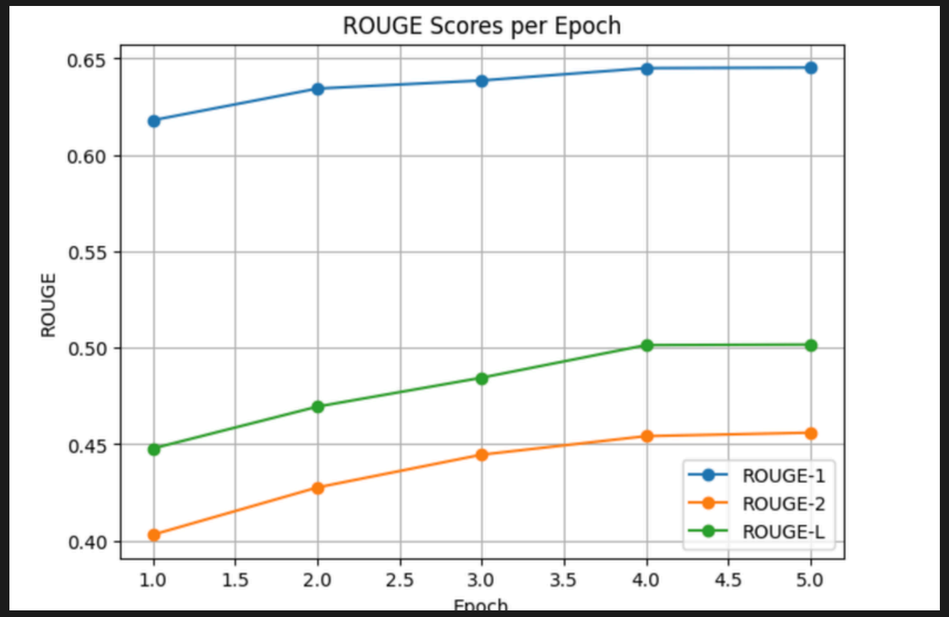


Fig .3: ROUGE Scores: Reached ROUGE-1 = 0.65, ROUGE-2 = 0.46, ROUGE-L = 0.50, showing strong text alignment.



RESULTS

Fig .4: Compression Ratio: Stabilized near 0.68, maintaining concise yet informative outputs.

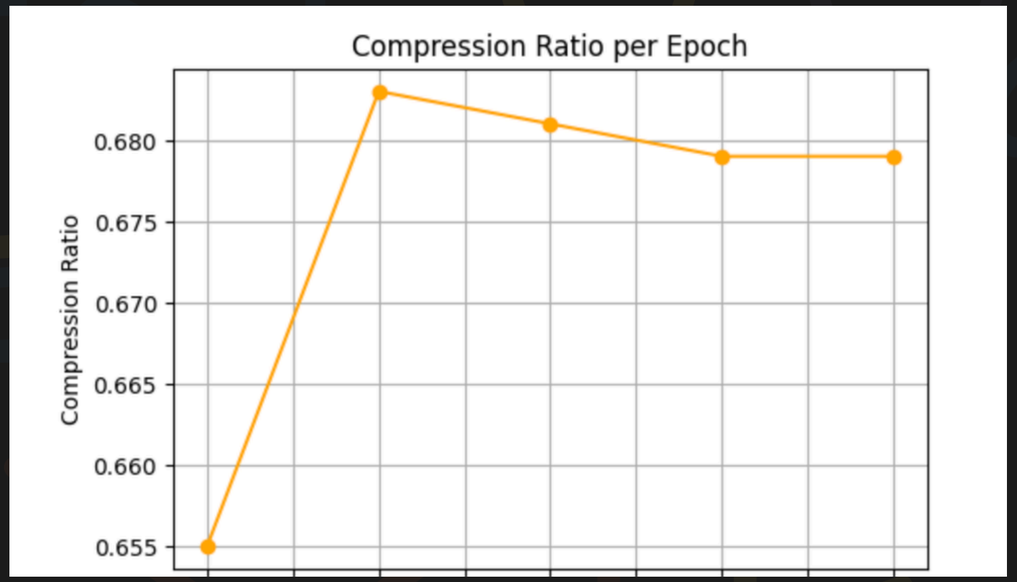


Fig .5: Overall: The model achieved higher fidelity and stability while maintaining concise summaries.

Metric	Initial	Final	Change
Eval Loss	3.768	2.253	40.2%↓
ROUGE-1	0.614	0.647	5.3%↑
ROUGE-2	0.399	0.459	15.2%↑
ROUGE-L	0.445	0.509	14.4%↑

RESULTS

SOURCE :
New research shows that underwater noise could disturb the animalsÕ communication and ability to nd prey. The low rumble of passing ships has, for a long time, been connected to the disturbance of large whales...

GENERATED:
New research shows that underwater noise could stop these animals communicating and make it more dif cult for them to nd food. It is well known that noise from ships disturbs large whales...

Word counts: Source=633, Generated=403

SOURCE :
<TO_ELEM> “It was mental in there. It was crazy. It was absolutely disgusting, disgusting.” Haggerty had ve...

GENERATED :
““It was mental in there. It was crazy. People were behaving like animals – it was horrible, disgusting.” Sh...

Word counts: Source=663, Generated=321

VALIDATION

Table 2. Evaluation Metrics on Validation Set (20%)

Metric	Source	Generated	Target (Reference)
ROUGE-1	0.7736	0.6827	—
ROUGE-2	0.6100	0.4653	—
ROUGE-L	0.6854	0.5117	—
BERT F1	0.4987	0.4545	—
FKGL (Readability)	11.03	10.19	8.94

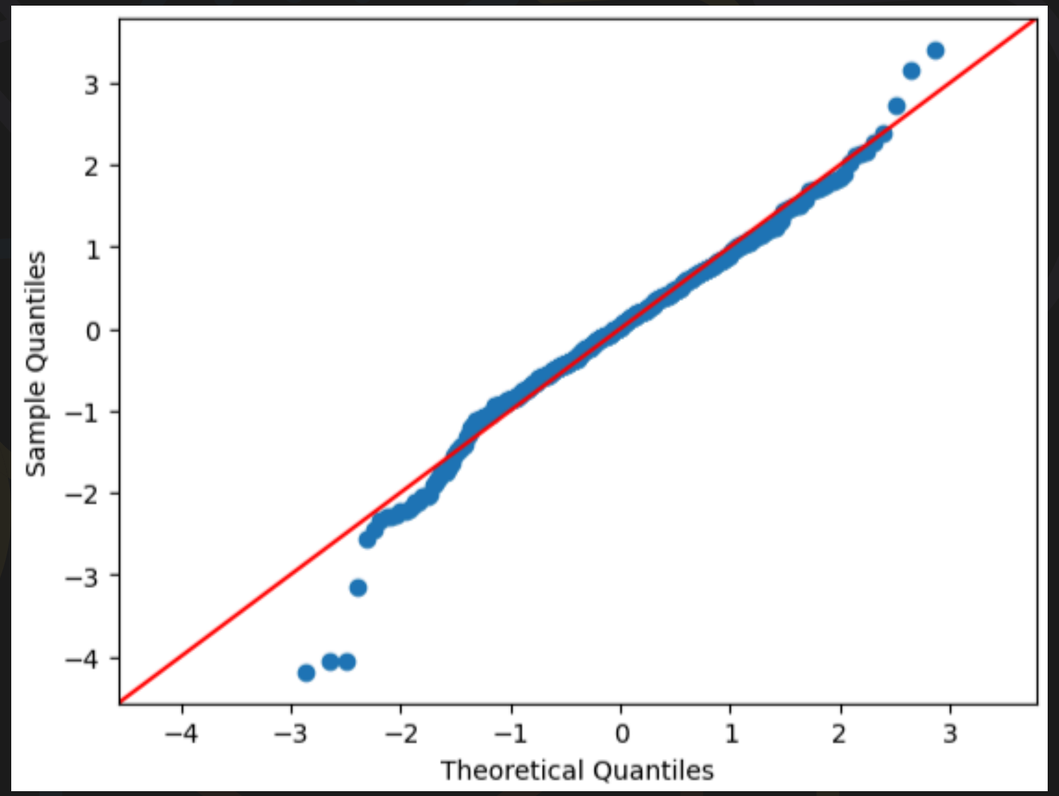
```
=== Regression with Direction Control ===
                            OLS Regression Results
==============================================================================
Dep. Variable:            target_fkgl   R-squared:                       0.767
Model:                            OLS   Adj. R-squared:                  0.765
Method:                 Least Squares   F-statistic:                     523.2
Date:                Fri, 24 Oct 2025   Prob (F-statistic):          2.15e-150
Time:                        15:21:40   Log-Likelihood:                -601.94
No. Observations:                 481   AIC:                             1212.
Df Residuals:                     477   BIC:                             1229.
Df Model:                           3
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                        0.5441      0.244      2.234      0.026       0.066       1.023
C(direction)[T.adv→inter]        1.6944      0.095     17.872      0.000       1.508       1.881
C(direction)[T.inter→elem]       0.9355      0.099      9.414      0.000       0.740       1.131
source_fkgl                      0.6844      0.020     33.431      0.000       0.644       0.725
==============================================================================
Omnibus:                       34.989   Durbin-Watson:                   1.991
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               68.999
Skew:                          -0.437   Prob(JB):                     1.04e-15
Kurtosis:                       4.637   Cond. No.                         72.4
==============================================================================
```

```
source_rouge1      0.773755
source_rouge2      0.610034
source_rougeL      0.685412
source_bert_f1     0.498662
generated_rouge1   0.682744
generated_rouge2   0.465317
generated_rougeL   0.511685
generated_bert_f1  0.454546

=== AVERAGE FKGL READABILITY ===
                 avg_fkgl
source_fkgl     11.026537
target_fkgl      8.935735
generated_fkgl  10.189682

 Model-generated summaries underperform the source in semantic similarity — may need more fine-tuning.
```

VALIDATION



Generated FKGL: 8.014 | Average Target FKGL: 8.923

OLS Regression Results

Dep. Variable:	target_fkgl	R-squared:	0.610
Model:	OLS	Adj. R-squared:	0.609
Method:	Least Squares	F-statistic:	750.1
Date:	Fri, 24 Oct 2025	Prob (F-statistic):	4.47e-100
Time:	15:21:40	Log-Likelihood:	-725.57
No. Observations:	481	AIC:	1455.
Df Residuals:	479	BIC:	1463.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.4745	0.277	5.332	0.000	0.931	2.018
source_fkgl	0.6794	0.025	27.388	0.000	0.631	0.728

Omnibus:	23.067	Durbin-Watson:	2.045
Prob(Omnibus):	0.000	Jarque-Bera (JB):	28.963
Skew:	-0.440	Prob(JB):	5.14e-07
Kurtosis:	3.819	Cond. No.	62.2

=== PER-DIRECTION METRICS ===

	direction	rouge1	rouge2	rougeL	bert_f1
0	adv→elem	0.628244	0.368734	0.442041	0.893914
1	adv→inter	0.734462	0.556212	0.582393	0.922590
2	inter→elem	0.684071	0.469796	0.509239	0.906836

PROJECT MILESTONES

Phase 1 – General-Domain Summarization (Oct 1–31)

- **Conduct exploratory data analysis on OneStopEnglish corpus**
- **Preprocess and clean text (tokenization, labeling, readability scoring)**
- **Fine-tune BART for 3 readability levels: elementary, intermediate, advanced**
- **Evaluate using ROUGE, BERTScore, and readability indices (FKGL, SMOG, Dale–Chall)**
- **Perform statistical testing (ANOVA, regression) for readability effects**
- Document results as baseline performance

Phase 2 – Biomedical Summarization (Nov 1–25)

- Fine-tune BART on the M3 dataset (expert, student, layperson summaries)
- Compare general vs. biomedical performance on readability and accuracy
- Perform statistical testing (ANOVA, regression) for readability effects
- Finalize model evaluation and visualizations
- Prepare final report and presentation by Nov 30

REFERENCES

- **OneStopEnglish** corpus Vajjala, Sowmya, and Ivana Lučić. “OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification.” Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications (BEA), 2018, pp. 297–304. ACL Anthology+1

- **CLEAR** (CommonLit Ease of Readability) corpus Crossley, Scott A., Aron Heintz, Joon Suh Choi, Jordan Batchelor, Mehrnoush Karimi, and Agnes Malatinszky. “A large-scaled corpus for assessing text readability.” Behavior Research Methods, 2022. SpringerLink+2pmc.ncbi.nlm.nih.gov+2

- **Cochrane Database** of Systematic Reviews (MSLR subset) DeYoung, Jay, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Lu Wang. "MS^2: Multi-Document Summarization of Medical Studies." arXiv preprint arXiv:2104.06486, 2021. https://arxiv.org/abs/2104.06486

- **BioLaySumm Corpus** Goldsack, Tomas, Zhihao Zhang, Chenghua Lin, and Carolina Scarton. "Making Science Simple: Corpora for the Lay Summarisation of Scientific Literature." Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022. https://github.com/TGoldsack1/Corpora_for_Lay_Summarisation

THANK YOU FOR LISTENING!