Dialogue Summarization on SAMSum — From Chats to Fast, Faithful Notes

By Alejandro Silva

Problem & Business Value

Problem Analysts read long chats; it’s slow and inconsistent.

Goal: concise, faithful summaries for faster case handling and better records.

Value: fewer manual edits, shorter handle time, consistent notes at scale.

Modeling

Model Choice

BERT2BERT (encoder–decoder) vs GPT-2 (decoder-only)

Training model

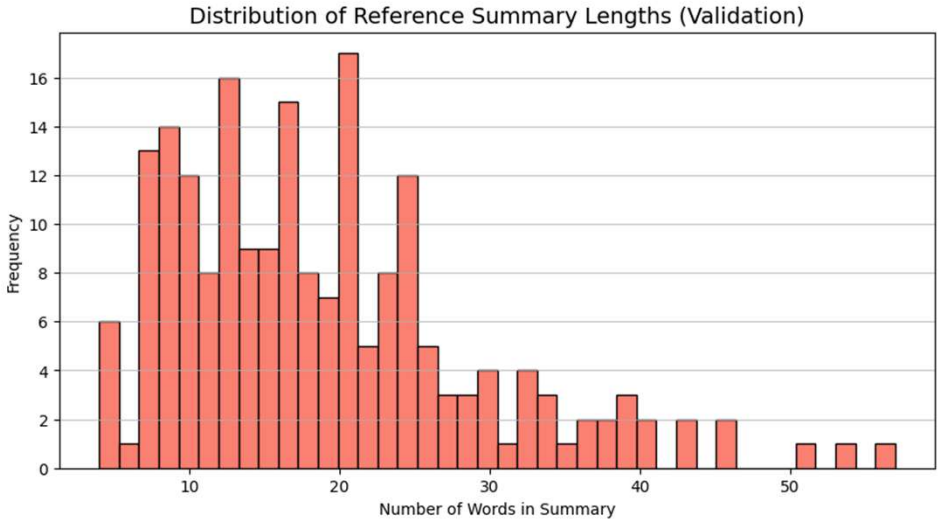
Optimizing Hyperparameters

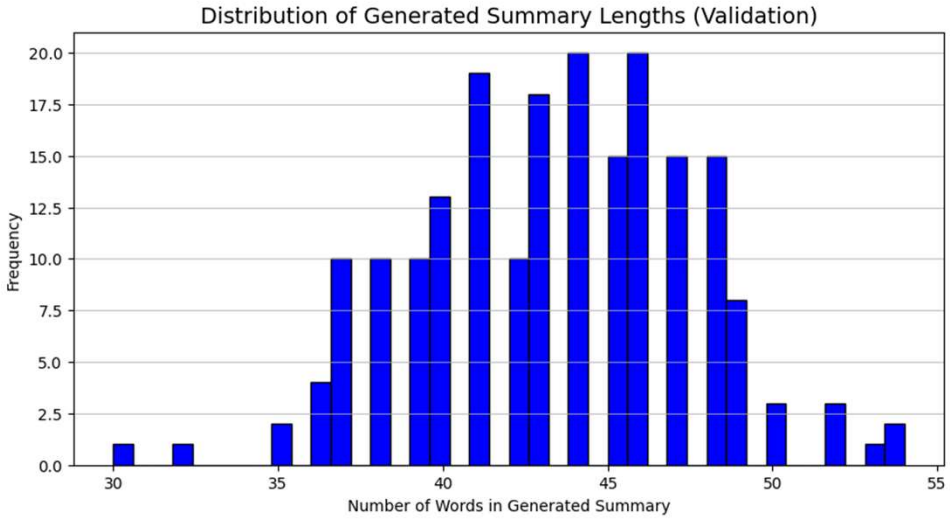
Decoding Optimization

Polish for production

Results

```
=== Aggregated stats (validation) ===
Avg words per dialogue: 91.51
Avg words per speaker turn: 8.17
Avg compression ratio (reference / dialogue): 0.280
Avg compression ratio (generated / dialogue): 0.809
Avg compression ratio (generated / reference): 3.063
```





Example

Dialogue	George: Fun fact time XD George: IQ decreases by 20% after a 2-week holiday Pete: lol Matt: haha wonder what happens after a gap year xD Pete :D
Reference Summary	IQ decreases by 20% after a 2-week holiday.
Model Summary	george's iq decreases by 20 % after a 2 - week holiday. he's looking forward to the fun fact that iq is 20 % faster after a two - week break. matt and matt want to know what happens after a gap year. they're happy about it.

Business Implications & Risk Mitigation

- Training was done with a subset
- The inference phase was done with full dataset
- Balance GPU / CPU
- Business users expect short readable notes

Conclusion, Recommendations & Future Projects

Model done successfully with potential improvements for the future

- More robust Multi-dataset
- Training with full set
- Verbosity & Length control
- Faithfulness
- Noisy references