ANRW2025

Applied Networking Research Workshop

EMPOWERING IETF COLLABORATION WITH NLP SEARCH INNOVATIONS AND LLM-ENHANCED RFC WRITING

Jie Bian, Michael Welzl

University of Oslo

{jiebi, michawe}@ifi.uio.no

Motivation I

- Discussions can become fragmented and disorganized as communications are scattered across a massive mail archive (together with other WGs' mails) or buried in personal mailboxes (together with other business or work-related mails).
- The vast and unstructured nature of the IETF email archive, coupled with its constant growth, creates a significant challenge for locating specific discussions or proposals, often requiring time-consuming manual sifting through irrelevant or outdated content.

Motivation II

- The journey from the initial draft to the final RFC can be long, often spanning several years
  - *The workload of the RFC*
  - *The number of participants, the locations of authors*
  - *The communication and collaboration (non-real-time)*
- Based on a rough estimate from 20 RFCs published in 2018 and analysed in RFC 8963
  - *The average RFC is approximately 35 pages long.*
  - *The average delay from initial draft submission to final publication is around 1182 days (3.24 years).*

Contribution

We want to streamline the workflows to alleviate the burden on human contributors and potentially accelerate the transition from initial drafts to finalized RFCs.

- TextFlow: A curated dataset which includes the WG discussions alongside the corresponding changes in I-Ds (before and after).
- TextFlow Search: An Information Retrieval (IR) based search system that links these comments or discussions to specific I-D details.
- TextFlow Edit: A generative task where we instructed the LLM to produce specific parts of the I-Ds based on the previous versions and the associated comments.

Dataset: TextFlow

- Source:
  - *IETF mail archive: no existing connections*
  - ***GitHub:** built-in PR-Issue relationship*
- Overview:
  - *200 WG repositories total*
  - 160 collected before Dec. 2024 → used for training, evaluation, and testing
  - 40 collected after Dec. 2024 → used as Out-of-Domain (OOD) set

TextFlow

- I-D

If no ACK_FREQUENCY frames have been received, this value defaults to 3, which is the recommended packet threshold for loss detection in (Section 18.2 of QUIC-RECOVERY).

- Comment (GitHub Issue)

It seems optimal to default it to 3, because that's the default packet threshold, but the principle of least surprise argues for 1 (ie: no behavior change until an ACK_FREQUENCY frame is received), so I lean toward 1.

TextFlow

- I-D pre-version

If no ACK_FREQUENCY frames have been received, this value defaults to 3, which is the recommended packet threshold for loss detection in (Section 18.2 of QUIC-RECOVERY).

- I-D post-version

If no ACK_FREQUENCY frames have been received, the endpoint immediately acknowledges any subsequent packets that are received out of order, as specified in Section 13.2 of QUIC-TRANSPORT, as such the default value is 1.

TextFlow Search

- Information Retrieval
  - *Sparse Retrieval: BM25*
  - *Dense Retrieval: LLM*
- Zero-shot Configuration
  - *Encoder-based model: BGE*
  - *Decoder-based model: Mistral*
- Supervised Fine-tuning
- Two Directions
  - Query: I-D → Target: Comment
  - Query: Comment → Target: I-D

Result

🔍 Query: Internet-Draft (I-D)

- The **targeted comment** typically ranks (**Mean Reciprocal Rank (MRR)**):
  - **1st–2nd** on the **test set**
  - **2nd–3rd** on the OOD **set**

💬 Query: Comment

- The targeted I-D achieves a similar ranking, with even higher metric scores.

TextFlow Edit

- Instruction Tuning
  - *A distinct type of supervised fine-tuning (SFT) featured by training the model with pairs of instructions and their corresponding outputs*
- Test Time Scaling (TTS)
  - *Allocating extra time and computational resources to LLMs during inference can improve results more effectively than adjusting their parameters*
- General Model
  - *Zero-shot configuration*
  - *Instruction-tuning with inputs and outputs*
- Reasoning Model
  - *w/o TTS*
  - *TTS I: allow extra time for the model to “think”, and explicitly output its thought process*
  - *TTS II: allow extra time to “think” and then “wait” to verify its generated result, and correct it accordingly if necessary*
  - *Insturction-tuning as above, without the thought process data*

Prompt for TTS

We give the model both the pre-version text and comment and let it output the revised text.

- You are an experienced IETF RFC author. Thoroughly review the original text snippet within an internet draft and feedback from the WG. Identify and highlight parts needing revision. Assess the quality of the feedback. If the feedback is unclear, low-quality, or irrelevant, use your expertise to make independent judgments. For complex feedback, take your time to carefully consider what it refers to and thoughtfully incorporate it into the original text. Finally provide well-considered revisions **only**.

Evaluation

Similarity comparison (Bert Score, BLEU, etc.) of I-D generated by LLM

We compare the generated I-D against:

✅ **New version:** Created by the Working Group (expected outcome)

📄 **Old version:** Used by the LLM as a reference

💬 **Comments or discussions:** Also used as reference material by the LLM

Results

The LLM generated I-D, generally:

- 📄 **Most similar** to the **old version**
- 🆕 **Fairly similar** to the **new version**, but slightly less than to the old
- 💬 **Least similar** to the **comment**

Comparisons of Models:

- SFT works well for generaly purpose model.
- TTS I works well for reasoning based model.
- The reasoning model often generates additional content, such as summaries and explanations, which reduces its similarity to the expected output.

Edit Result Demo

Edit by WG member

- If no ACK_FREQUENCY frames have been received, the endpoint immediately acknowledges any subsequent packets that are received out of order, as specified in Section 13.2 of QUIC-TRANSPORT, as such the default value is 1.

Edit by Reasoning Model (Qwen-32B) with TTS I

- If no ACK_FREQUENCY frames have been received, this value defaults to 1, which aligns with the principle of least surprise, maintaining the status quo until an ACK_FREQUENCY frame is received.

Future Work

- We hope it will also work efficiently with the IETF email archive data.
- Now the two systems run independently, but we can make them work together as a Retrieval Augmented Generation (RAG) system.

Thank you!

Q & A