Automating IETF
Insights Generation with AI

https://arxiv.org/pdf/2410.13301

Jaime Jiménez, IETF121 2024-11-08

Objectives

- Automate IETF WG report generation.
- Explore GenAI capabilities for summarization and insight extraction.
- Evaluate the application with different LLMs (Ollama local Models, OpenAI, Anthropic) for reliable reports.

Mitigating hallucinations

- GenAI often produces content not based on factual or accurate information or *hallucinations*
- It applies to all content types: text,images,video...
- Several strategies exist to mitigate this: output self-checks by the LLM, prompt manipulation, self-supervision with reasoning strategies and other.
- However, one of the simplest and most effective mechanisms is to require ground truth material...

IETF open records

Tool	Description	Example	Status on this work
IETF Mail Archive	Searchable email archive by date, author, and subject	-	✗
Datatracker API	RESTful API for accessing IETF documents, meetings and WG data	`GET https://datatracker.ietf.org/api/v1/doc/rfc1234`	✓
IETF GitHub	GitHub organization for IETF working groups and projects	-	–
rsync IETF Server	Backup server for IETF documents and materials	`$ rsync -av rsync.ietf.org::proceedings/119/ proceedings/ietf119`	✓
Youtube	The IETF Channel has 4500 video recordings of every meeting		✗

IETF specs work well with LLMs

- In the standard's context, IETF drafts are much easier to consume by LLMs than other documents.
  - plaintext
  - markdown with aasvg diagrams and utf-8 text.
  - tagged html.
- common document patterns in terms of interfaces, message examples, req/res interactions, IANA, Security considerations and references.
- RFCs and Internet protocols are already part of the LLM's training data

Workflow

1. **Data Retrieval**: `rsync` and crawl for the contents of the proceedings. Mainly meeting minutes, participant list, presented drafts and agendas.
2. **Data Preprocessing**: Normalize and organize the collected input. This involves removing duplicates, standardizing participant names, and structuring the data in a db for analysis.
3. **LLM Integration**: Apply Retrieval-Augmented Generation (RAG) on the docs for retrieval, together with other text API query for context. Customize prompts to retrieve structured output from the LLM. Use GraphRAG for globally queries affecting a larger corpus.
4. **Report Generation**: Generate summaries, format output in LaTeX or Markdown and perform post-processing error corrections.

We tested several models

Model Name	Parameters	Size Category	Type	Context Window [1]
GPT-4	1.76T	Very Large	API	8,192 tokens
Claude 3 Sonnet	175B	Very Large	API	100,000 tokens
GPT-4-Turbo	175B	Very Large	API	4,096 tokens
Command-R	35B	Large	Local	131,072 tokens
Mixtral	46.7B	Large	Local	32,768 tokens
Llama 3	8B	Large	Local	8,192 tokens
Gemma2	8.5B	Small	Local	8,192 tokens
Phi-3	3.8B	Small	Local	512 tokens

[1] We were limited by our hardware. Larger context window costs memory, which is limited on GPUs.

Limitations

- We do not use Chain of Thought (CoT) or other genai-reasoning solutions that could further reduce hallucinations.
- We do not use all the extend of the records:
  - Meeting transcriptions tend to produce worse ouput.
- GraphRAG allows to provide context about multiple drafts, but it has a higher cost and requires indexing first.
  - At this point, insights that related to different WGs are not used (e.g., how does protocol X in WG A affect protocols Y and Z on WG B).

Example

Mostly...

- Accurate depiction of the events during the meeting.
- Accurate participation affiliation.
- Accurate name and role (e., AD, chair...) retrieval of presenters.
- Concise summarization of discussion and action points.

But every now and then...

- the code still takes a creative detour and invents some of the above! Far less often than zero-shot queries)

More examples at: https://arxiv.org/pdf/2410.13301

37.1 Attendee Overview

37.1.1 Attendance Summary

The IOTOPS session at IETF 119 in Brisbane was attended by a diverse group of 38 participants, representing prominent companies and institutions such as Isode Limited, RISE Research Institutes of Sweden, High North Inc, Itron, ALAXALA Networks, Corp., Arm, Ericsson, and Google. The attendance reflected a wide interest in the operational aspects of IoT from various sectors of the industry.

The discussions during the session were focused on the operational challenges and the security aspects of IoT. Notably, the session included presentations on the terminology for constrained-node networks, a comparison of CoAP security protocols, and an update on the IoT operational security summary. These discussions were enriched by contributions from attendees, which provided insights into the practical implications of the topics at hand. For instance, the terminology presentation by Carsten Bormann (draft-bormann-iotops-ietf-lwig-7228bis-00) aimed to update the classification of constrained-node networks, reflecting the evolution of technology since the original RFC 7228.

Meeting materials are available via the direct link: IOTOPS Session Notes.

37.2 Meeting Discussions

37.2.1 Terminology for Constrained-Node Networks

Carsten Bormann presented the revised terminology for constrained-node networks, which is crucial for understanding and operating IoT systems. The updated document aims to reflect the technological advancements and provide a common language for IoT operations. The group discussed the potential need for a charter update to accommodate the terminology draft, with a consensus that the terms are indeed part of IoT operations.

37.2.2 Comparison of CoAP Security Protocols

John Mattsson presented a comparison of CoAP security protocols (draft-ietf-iotops-security-protocol-comparison-04), summarizing the recent changes and suggesting that the document is stable enough for publication. The group agreed to initiate a Working Group Last Call (WGLC) while parking the document for normative references.

37.2.3 IOTOPS Security Summary Update

Brendan Moran provided an update on the IoT operational security summary (draft-ietf-iotops-security-summary-01), which references baseline security documents and technologies relevant to IoT. The discussion highlighted the importance of considering regional regulatory requirements, such as the EU Cyber Resilience Act, and the potential need for additional authors to address these aspects.

37.2.4 IoT Operational Issues

Karsten Walther shared practical operational issues encountered in IoT deployments, emphasizing the need for standardization to address common problems. The presentation sparked a discussion on the role of browsers in IoT and the challenges of network configuration in virtualized environments. The group considered the value of documenting these experiences in an Internet-Draft to facilitate broader discussions and potential solutions within the IETF community.

The session concluded with a commitment to further explore the presented issues, with the potential for new work items to emerge from these discussions. The next steps include initiating a WGLC for the

Thanks!

https://arxiv.org/pdf/2410.13301