Poster: Automated Extraction of Protocol State Machines from 3GPP Specifications with Domain-Informed Prompts and LLM Ensembles

Miao Zhang[1], Runhan Feng[2,*], Hongbo Tang[1,*], Yu Zhao[1], Jie Yang[1], Hang Qiu[1], Qi Liu[2]
[1] Information Engineering University, China
[2] Purple Mountain Laboratories, China
1152461073@qq.com, fengrunhan@pmlabs.com.cn, tahobo@sina.com, itsyz@foxmail.com,
yj_csu@126.com, qiuhang_ndsc@163.com, liuqi@pmlabs.com.cn

***Abstract*—Mobile telecommunication networks are foundational to global infrastructure and increasingly support critical sectors such as manufacturing, transportation, and healthcare. The security and reliability of these networks are essential, yet depend heavily on accurate modeling of underlying protocols through state machines. While most prior work constructs such models manually from 3GPP specifications, this process is labor-intensive, error-prone, and difficult to maintain due to the complexity and frequent updates of the specifications. Recent efforts using natural language processing have shown promise, but remain limited in handling the scale and intricacy of cellular protocols. In this work, we propose SpecGPT, a novel framework that leverages large language models (LLMs) to automatically extract protocol state machines from 3GPP documents. SpecGPT segments specifications into semantically meaningful paragraphs and constructs domain-informed prompts to guide model interpretation. We evaluate SpecGPT on three representative 5G protocols (NAS, NGAP, and PFCP) using manually annotated ground truth, and show that it outperforms existing approaches, demonstrating the effectiveness of LLMs for protocol modeling at scale.**

***Index Terms*—Large Language Models, State Machine, 3GPP Standards**

I. Introduction

Mobile telecommunication networks serve as a core global communication infrastructure, and their security and reliability are critical given their integration into key industries [1]; protocol state machines are the foundational framework for network security evaluation, yet extracting these models from complex 3GPP specifications is highly challenging, as manual construction is inefficient and error-prone, while existing methods perform poorly on cellular protocols and LLM applications in this domain remain limited by hallucination and protocol complexity. To bridge these gaps, we propose SpecGPT, an LLM-driven framework for automated protocol FSM extraction from 3GPP specifications, integrating semantic segmentation, domain-specific prompting, RAG, feedback validation and model ensembling to ensure reliable extraction. Evaluated on three mainstream 5G core protocols with manual annotated benchmarks, our method achieves outstanding state transition extraction performance and surpasses existing state-of-the-art solutions.

II. Background

Cellular networks have evolved from 1G to 5G, shifting from voice-centric systems to high-performance data-driven architectures supporting high speeds, ultra-low latency and massive machine-type communications [2]; a standard cellular network includes RAN and CN, with network functions decoupled from dedicated hardware to form modular service-based architectures, the 5G CN consists of independent, role-specific network functions (NFs), connected via standardized interfaces and dedicated protocol stacks for control-plane and user-plane operations. 3GPP is a global alliance of telecom standards bodies, developing universal mobile specifications to ensure cross-vendor interoperability and industry standardization, with a full specification system covering RAN, CN and UE, and its lengthy, frequently updated documents make manual analysis inefficient, creating an urgent need for automated parsing solutions.

III. Design

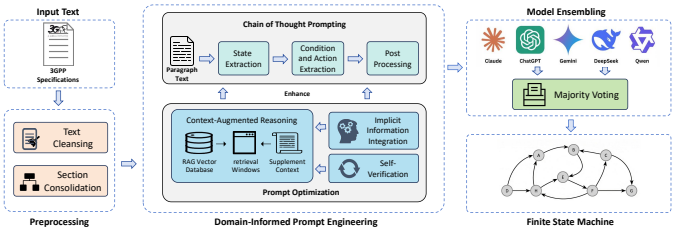

Fig. 1. Overview of SpecGPT

A. Overview

To address the non-trivial task of extracting 3GPP protocol state machines, we propose SpecGPT, a framework using LLMs, domain-informed prompts, and LLM ensembles for robustness. As illustrated in Figure 1, it comprises three core stages.

B. Preprocess

We first perform text cleaning on raw 3GPP Word documents using regular expressions to remove non-body content

(headers, footers, etc.). To mitigate LLM input truncation, we use a section-tree based merging strategy: parse section numbers to build a hierarchical tree, then apply a bottom-up approach to recursively merge leaf nodes under the same parent, generating semantically coherent chunks aligned with the original document structure.

C. Prompt Engineering

We integrate domain knowledge with Chain-of-Thought prompting to decompose the FSM extraction task into a sequence of logical steps. We first perform state extraction, where we classify protocols into state-oriented protocols such as NAS and procedure-oriented protocols such as PFCP, and handle state abbreviations such as *PLMN-SEARCH* (whose full qualified state is *5GMM-REGISTERED.PLMN-SEARCH*) through multi-level joint reasoning. We then conduct transition extraction, where we clarify the logical boundaries between conditions and actions in prompts. Finally, we carry out post-processing, where we validate the output JSON format and remove any pseudo-states or empty states to ensure reliability.

D. Ensembling

To mitigate LLM hallucinations, we use multiple mainstream LLMs with identical prompts. We align the outputs of these models through a consensus mechanism: we first check the consistency of the initial and next states of the extracted transitions, then verify the text overlap of the corresponding actions and conditions, and finally adopt a majority voting strategy to obtain the final finite state machine.

IV. Evaluation

We accessed five advanced LLMs (GPT 5.2, DeepSeek 3.2, Qwen 3 max, Claude opus 4.5, Gemini 3 Pro) via APIs. To address the lack of authoritative ground truth datasets for 3GPP protocols, we manually constructed a state machine ground truth dataset for 5G core network protocols (NAS, NGAP, PFCP) based on 3GPP Release 17, with rigorous validation by domain experts to ensure coverage and compliance with specifications.

We evaluated SpecGPT on extracting state machines from Release 17 NAS, NGAP, and PFCP protocols. A state machine transition was considered correct only when the states matched exactly and the overlap between condition and action spans reached at least 75%. All five large language models achieved a 100% F1-score for NAS state extraction. For NAS transition extraction, as shown in Table I, the models demonstrated high recall rates ranging from 79.89% to 94.97% but relatively low precision rates from 62.45% to 80.81% due to the hallucination issue, and the ensemble strategy boosted the F1-score by 6.27% to 22.68%. Further testing on PFCP and NGAP protocols shows that SpecGPT delivers solid overall performance on PFCP with a high F1-score, while the more structurally and logically complex NGAP protocol brings greater extraction difficulty, resulting in a relatively lower overall F1-score.

We further compared SpecGPT with Hermes [3], the state-of-the-art tool for 3GPP protocol FSM extraction, focusing on 5G NAS (the only protocol with reported Hermes results). Unlike Hermes, which only extracts explicit transitions and misses implicit logic, SpecGPT captures implicit transitions and achieves higher accuracy (86.41% for actions, 92.94% for conditions) than Hermes (81.39% and 86.40% respectively).

TABLE I
Comparison of the three indicators of various LLMs

Model	Precision (%)	Recall (%)	F1-score (%)
Claude opus 4.5	80.81	89.39	84.88
DeepSeek 3.2	69.37	86.03	76.81
Gemini 3 Pro	70.39	91.62	79.61
GPT 5.2	79.44	94.97	86.51
Qwen 3 max	62.45	79.89	70.10
Ensemble	92.27	93.30	92.78

V. Discussion

Despite SpecGPT reaching 92.78% F1-score and surpassing state-of-the-art 3GPP protocol extraction tools, it still has false positives and negatives caused by LLM hallucinations [4]. This issue can be eased via upgraded foundation models and self-consistency methods [5], and our extensible framework supports future LLM upgrades for better precision. Limited to state machine extraction only, we will mine other key 3GPP specification data in follow-up research. Practically, SpecGPT automates labor-heavy manual modeling, cuts manual workload, keeps pace with frequent spec updates, and provides structured models for downstream protocol verification and testing.

VI. Conclusion

This work proposes SpecGPT, an LLM-powered framework for automated protocol state machine extraction from 3GPP specifications. It integrates domain prompts, RAG, feedback validation and ensemble learning to handle complex cellular protocol documents, and outperforms existing methods on NAS, NGAP and PFCP protocols. The extracted structured state machines facilitate downstream protocol testing and formal verification, enhancing communication protocol robustness and security.

References

[1] R. Khan, P. Kumar, D. N. K. Jayakody, and M. Liyanage, "A survey on security and privacy of 5g technologies: Potential solutions, recent advancements, and future directions," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 1, pp. 196–248, 2019.

[2] S. R. Hussain and e. a. Echeverria, "5greasoner: A property-directed security and privacy analysis framework for 5g cellular network protocol," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security (CCS '19)*, 2019, pp. 669–684.

[3] A. Al Ishtiaq, S. S. S. Das, S. M. M. Rashid, A. Ranjbar, K. Tu, T. Wu, Z. Song, W. Wang, M. Akon, R. Zhang *et al.*, "Hermes: unlocking security analysis of cellular network protocols by synthesizing finite state machines from natural language specifications," in *33rd USENIX Security Symposium (USENIX Security 24)*, 2024, pp. 4445–4462.

[4] Z. Xu, S. Jain, and M. Kankanhalli, "Hallucination is inevitable: An innate limitation of large language models," *arXiv preprint arXiv:2401.11817*, 2024.

[5] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou, "Self-consistency improves chain of thought reasoning in language models," *arXiv preprint arXiv:2203.11171*, 2022.