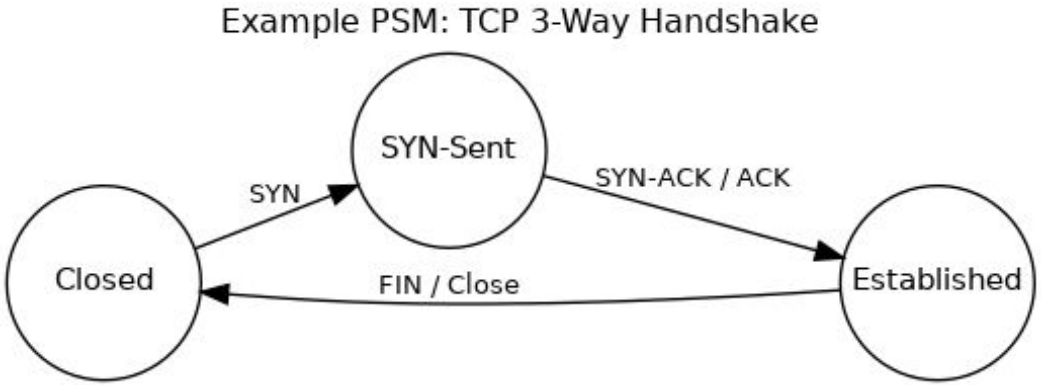
PURDUE UNIVERSITY

PSMBench: A Benchmark and Dataset for Evaluating LLMs on Extracting Protocol State Machines from RFCs

Zilin Shen, Xinyu Luo, Imtiaz Karim, Elisa Bertino

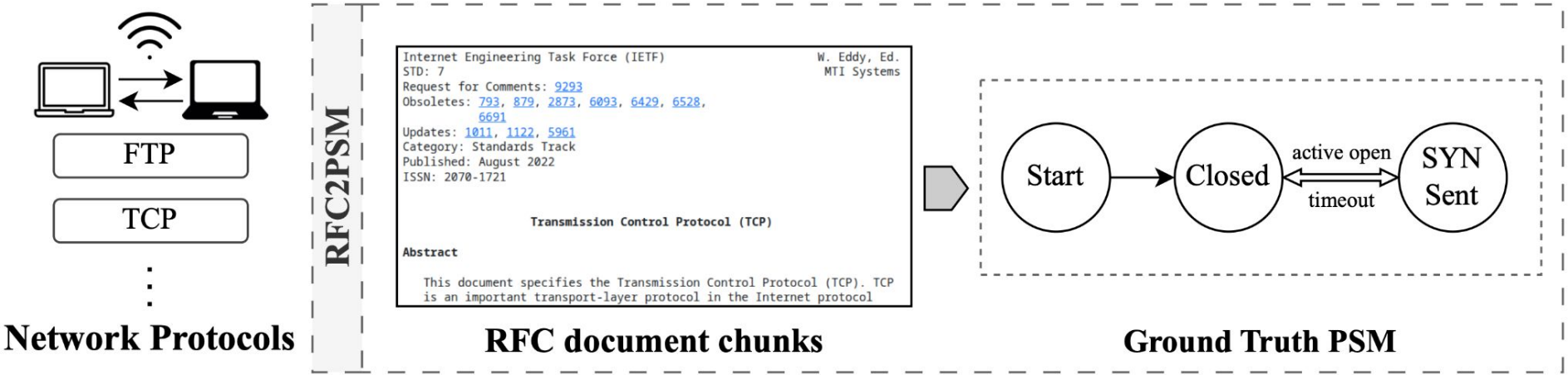
Background: Protocol State Machines and RFCs

- **Protocol State Machines (PSMs)** describe valid message flows between states in a network protocol.
- PSMs are essential for **security testing, fuzzing, and formal verification**.
- **RFCs (Request for Comments)** are long, **natural-language** documents specifying protocol behaviors.
- However, RFCs lack explicit state-machine structure, making PSM construction difficult.



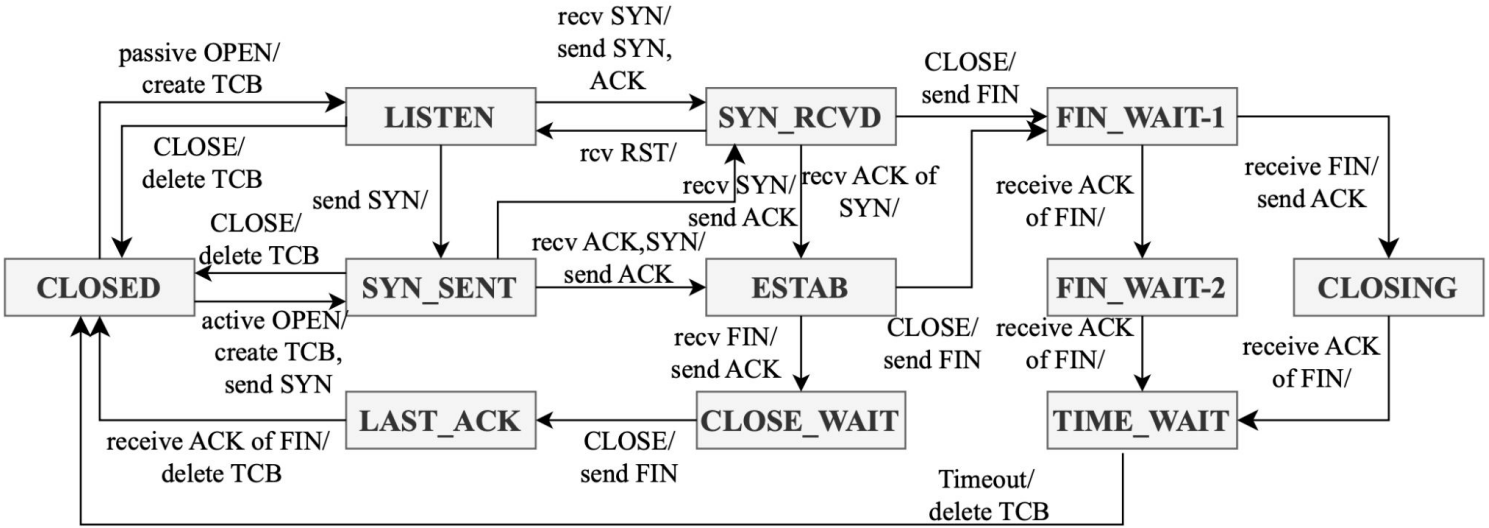
Why PSMBench?

- **RFC** documents are long, technical, and ambiguous; manual PSM extraction is costly and error-prone.
- Existing works are ***protocol-specific*** and lack reproducible, **cross-protocol** evaluation.
- Goal: enable automatic, quantitative, and structure-aware evaluation of **RFC-to-PSM** extraction.



A Multi-Protocol Ground-Truth Dataset

- Covers **14** network protocols across **L2–L7**.
- Includes **108 states, 297 transitions**, and ≈ **1,580 RFC pages**.
- Each PSM manually annotated and cross-verified **(κ = 0.82/0.78)**.
- Stored in structured JSON for visualization and benchmarking.

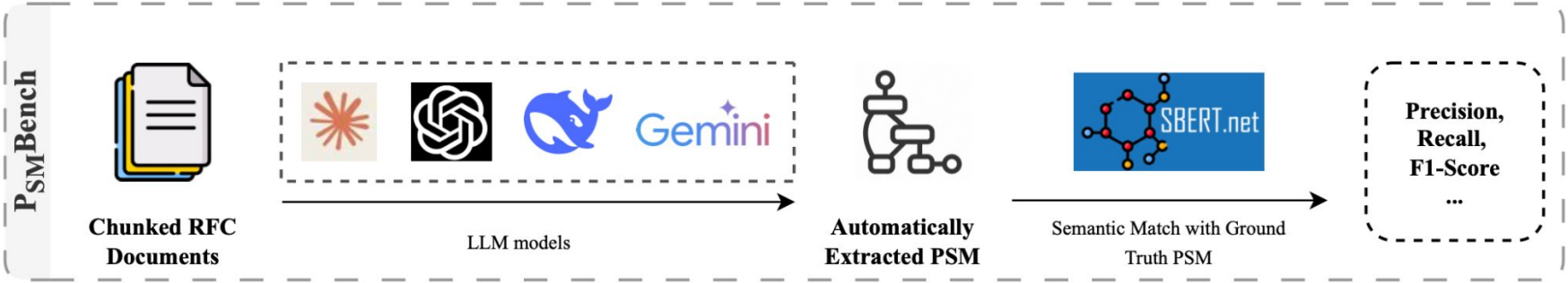


How PSMBench Works

1. **Segmentation:** Split RFCs by section (with metadata section number and name).
2. **Extraction:** LLM extracts partial PSMs (including partial states and partial transitions) from each segment.
3. **Merging & Scoring:** Combine into full PSM and evaluate via semantic structural similarity.

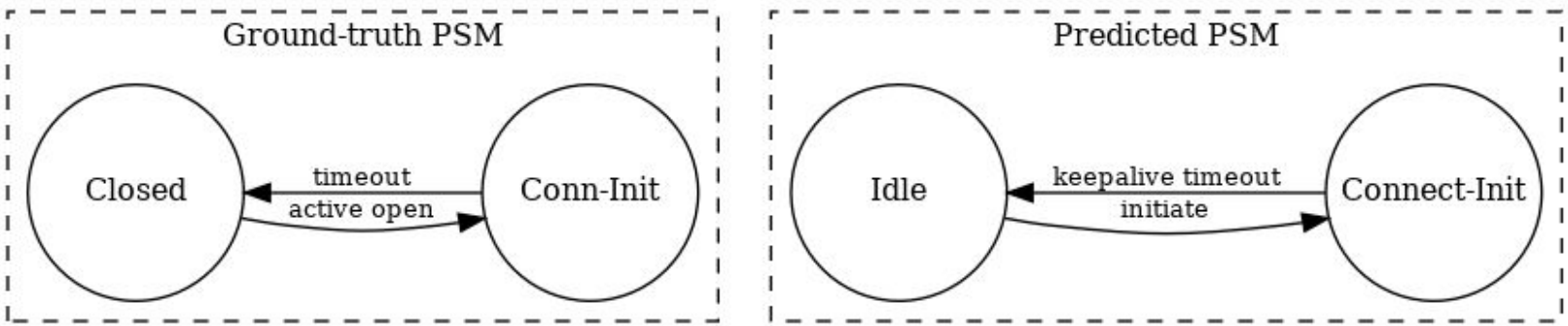
Highlights:

- **Windowing ablation:** sliding vs. section-based (+0.05 macro F1).



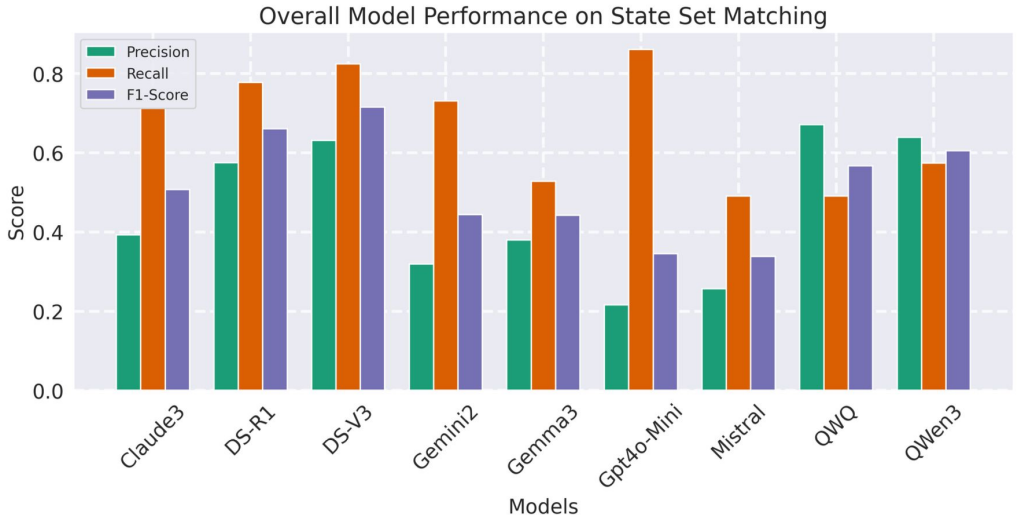
Semantic-Aware Evaluation

- Embedding-based **semantic matching** using Sentence-BERT (cos ≥ 0.5).
- Two-level evaluation:
  - **State-set** matching
  - **Transition** matching
- Report **Precision**, **Recall**, and **F1-score** for all models.



Key Experimental Results

- **State extraction > transition extraction:** structural reasoning remains difficult.
- **DeepSeek-V3** achieves the best state-set **F1 (0.715)**.
- Larger LLMs show higher recall but over-generation; smaller models are more precise.
- Gap between state and transition accuracy motivates future research.



Takeaways and Future Work

- Introduced **PSMBench**, the first benchmark for RFC-to-PSM extraction.
- Proposed **semantic structural metrics** for robust evaluation.
- Benchmarked **9 LLMs** across state and transition accuracy.
- Future: extend to **longer standards (e.g., Wi-Fi)**, adaptive merging, and **multimodal (figure + text)** PSM extraction.