Talk to your web navigation agent!

WebLINX

Real-World Website Navigation with Multi-Turn Dialogue

Xing Han Lù* xing.han.lu@mail.mcgill.ca
Zdeněk Kasner* kasner@ufal.mff.cuni.cz
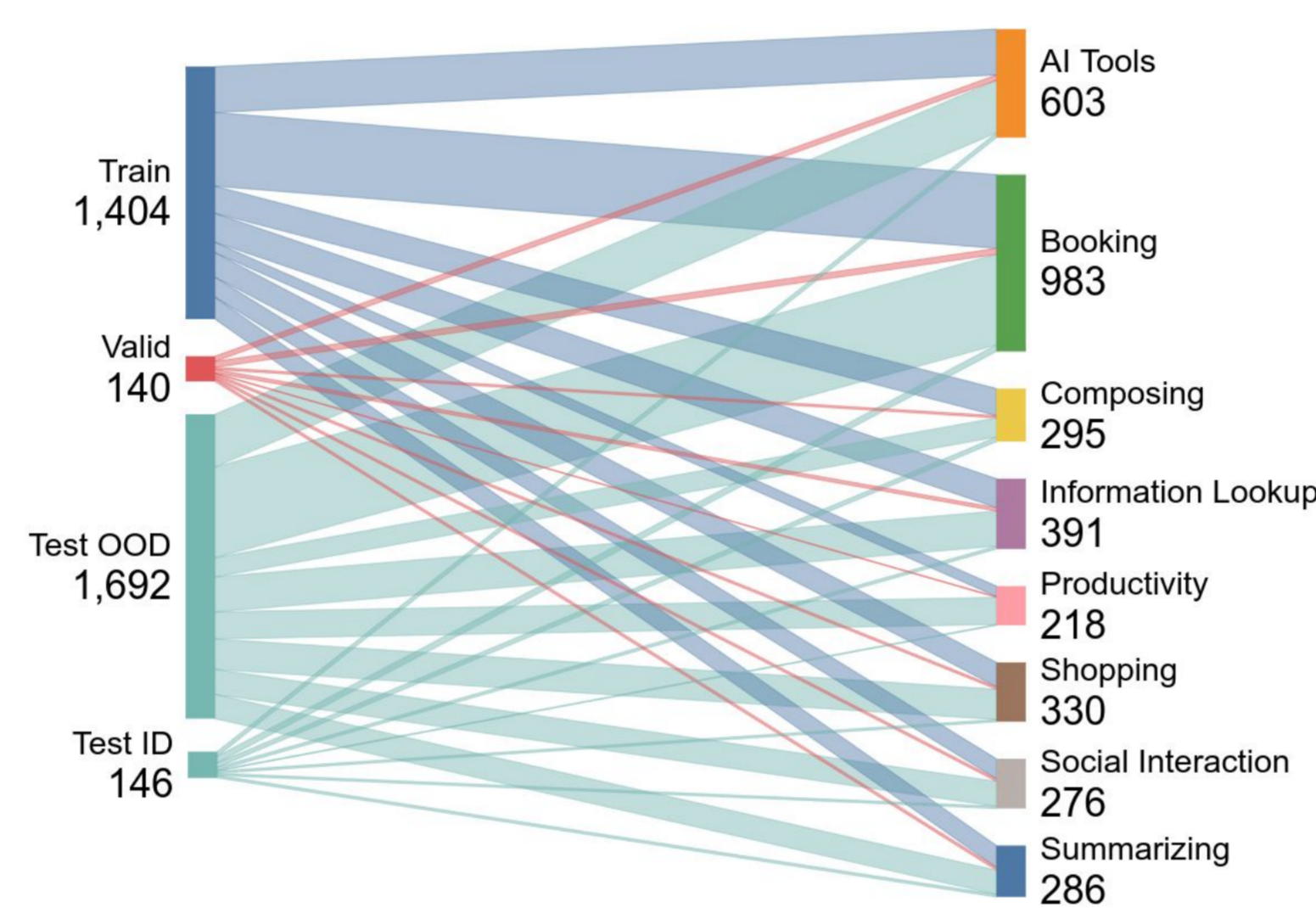**Siva Reddy** siva.reddy@mila.quebec

*These authors contributed equally to this work.
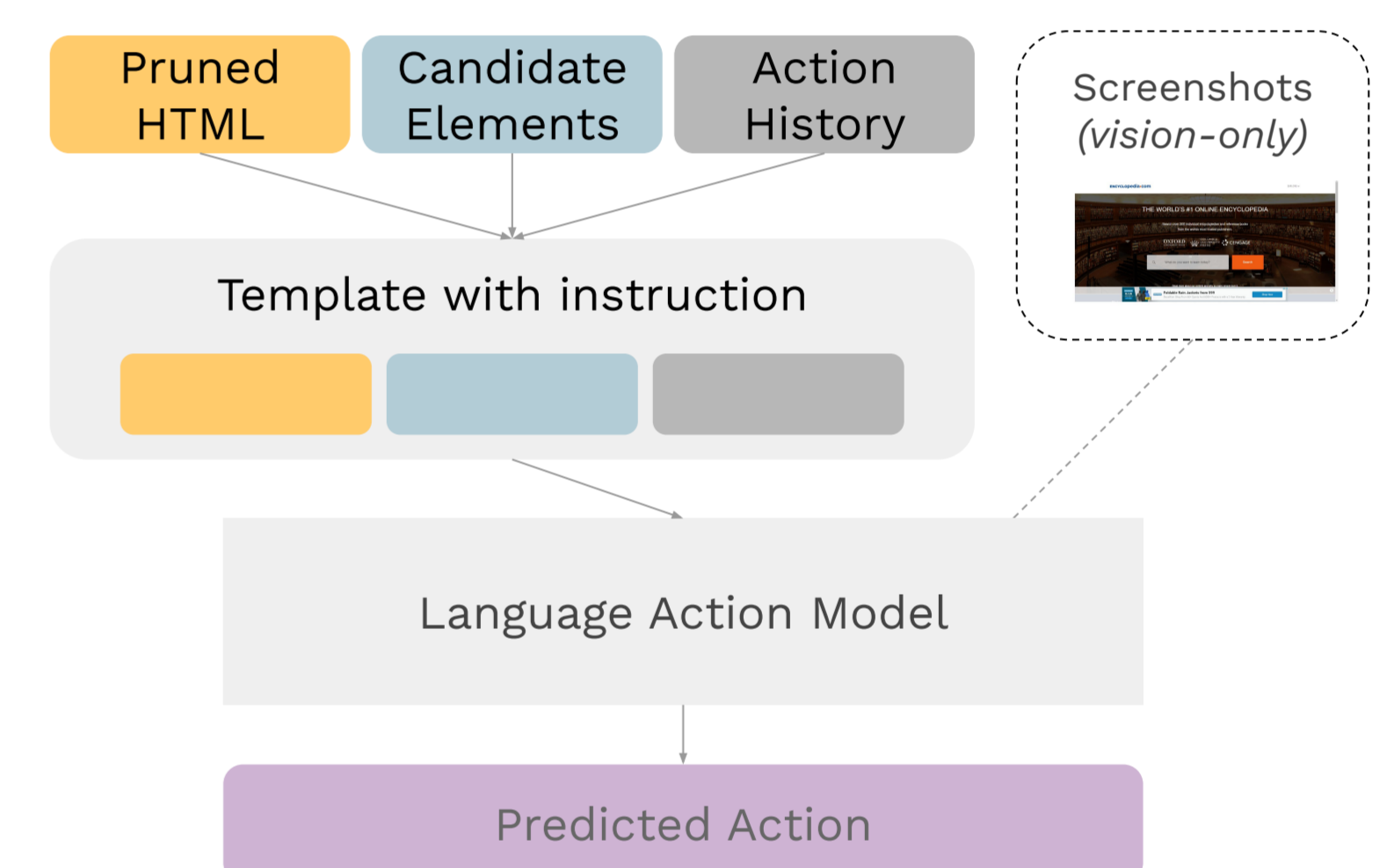
Mila • CHARLES UNIVERSITY

We propose **conversational web navigation:** the task of instruction-based real-world website navigation interleaved with **user-agent interactions**.
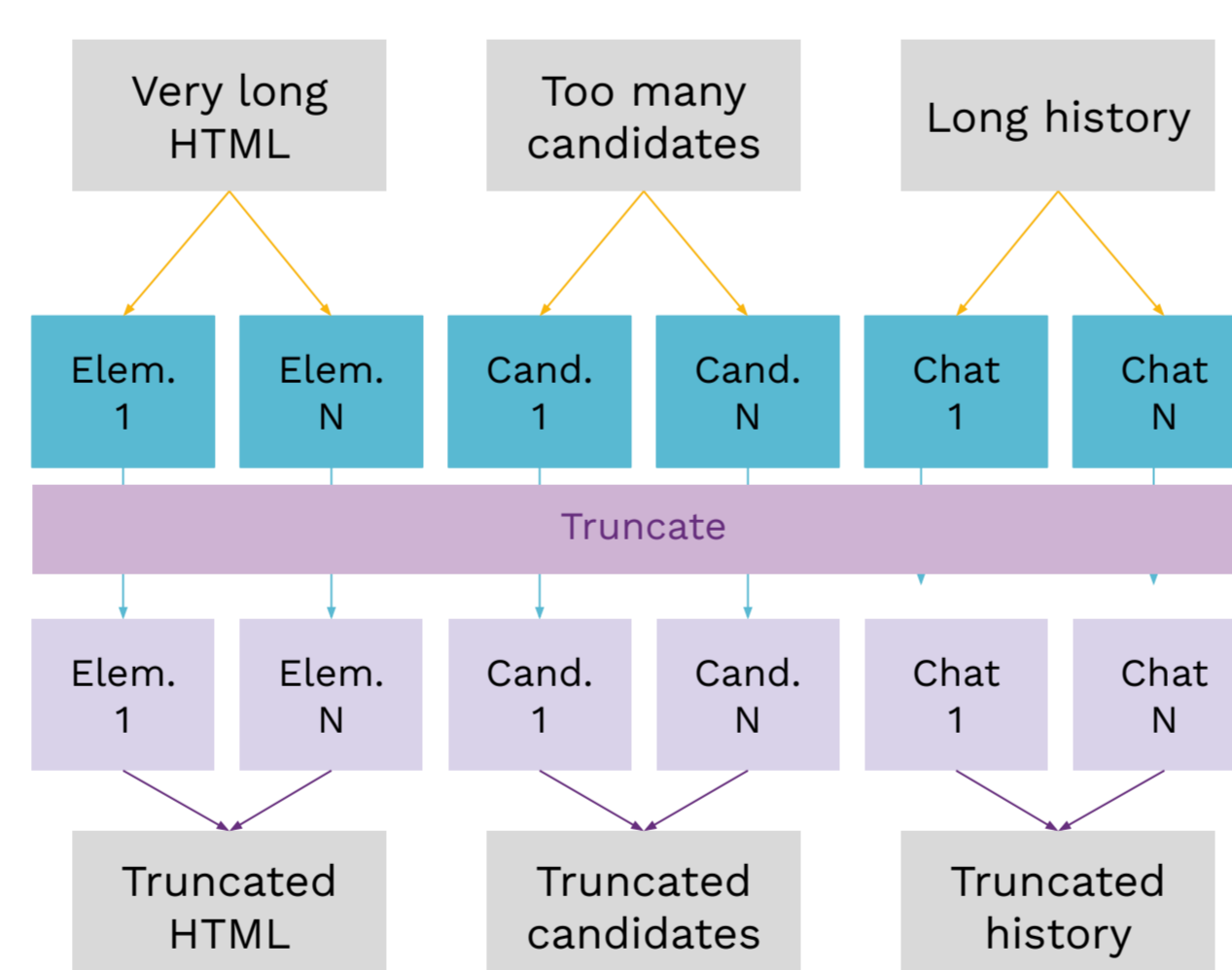

😀 Create for me a poster on Google Drawings

🖥 **say**("Sure!")

🖥 **load**("https://docs.google.com/drawings")

Untitled drawing ☆
File Edit View Insert Format Arrange Tools Help

🖥 **say**("What is the paper name?")

😀 WebLINX: Real-World [...]

🖥 **click**(`<div>`)

Untitled drawing ☆
File Edit View Insert Format Arrange Tools Help

🖥 **input**(`<div>`, "WebLINX [...]")

WebLINX: Real-World Website Navigation with Multi-Turn Dialogue
File Edit View Insert Format Arrange Tools Help

🖥 **say**("Done. Anything else?")

😀 That's all, thanks!

➡ We recorded **2300 expert demonstrations** across *150* real-world websites.

Train 1,404 · Valid 140 · Test OOD 1,692 · Test ID 146
AI Tools 603 · Booking 983 · Composing 295 · Information Lookup 391 · Productivity 218 · Shopping 330 · Social Interaction 276 · Summarizing 286

➡ We designed an **action prediction framework** for large action models.

Pruned HTML · Candidate Elements · Action History · Screenshots *(vision-only)*
Template with instruction → Language Action Model → Predicted Action

➡ We proposed improvements for **handling real-world inputs**.

Very long HTML → Elem. 1 ... Elem. N → Truncate → Truncated HTML
Too many candidates → Cand. 1 ... Cand. N → Truncate → Truncated candidates
Long history → Chat 1 ... Chat N → Truncate → Truncated history

◆ Closed-source models in zero-shot *surpassed by finetuned models.*
◆ For finetuned models: text-only models surpass multimodal models.
◆ All the models make *unjustified errors*.

➡ We evaluated **SotA action models** with turn-based evaluation metrics.

Models	Size	Test$_{OOD}$	Test$_{IID}$
Zero-shot			
Llama-2	13B	5.2	5.6
GPT-3.5T	–	8.5	10.3
GPT-4T	–	10.7	12.2
GPT-4V 👁	–	10.4	12.9
Finetuned			
Pix2Act 👁	1.3B	16.9	23.9
S-LLaMA	2.7B	25.0	**37.4**
MindAct	3B	20.9	25.7
Flan-T5	3B	23.8	31.1
Fuyu 👁	8B	20.0	30.9
Llama-2	13B	**25.2**	37.0
GPT-3.5F	–	21.2	30.8
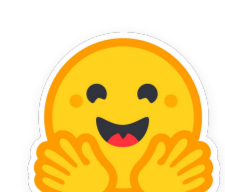
😀 Benchmark & models + 💻 Code + 🔻 Data explorer