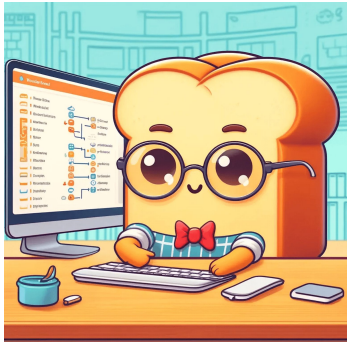


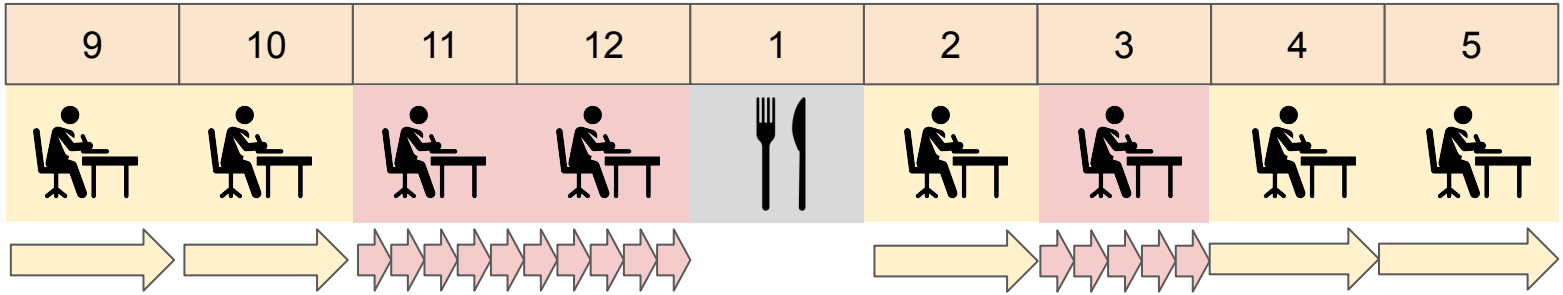
WONDERBREAD

A Benchmark for Evaluating Multimodal Foundation Models on Business Process Management Tasks

Michael Wornow, Avanika Narayan, Ben Viggiano, Ishan S. Khare, Tathagat Verma, Tibor Thompson, Miguel Angel Fuentes Hernandez, Sudharsan Sundar, Chloe Trujillo, Krrish Chawla, Rongfei Lu, Justin Shen, Divya Nagaraj, Joshua Martinez, Vardhan Agrawal, Althea Hudson, Nigam H. Shah, Christopher Ré

NeurIPS 2024

Problem: People spend **too much time** on tedious workflows



92%

of jobs require digital skills

3 hrs/day

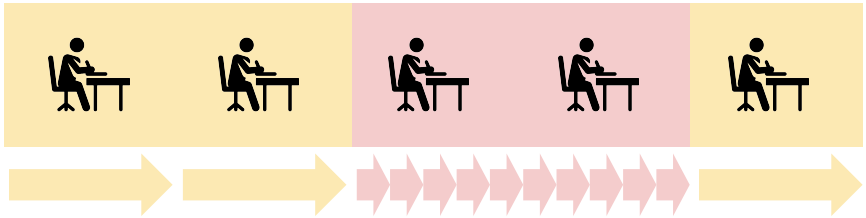
spent on repetitive tasks unrelated to core job

Solution: Business Process Management (**BPM**)

BPM is a systematic approach to make organizations more efficient by measuring and optimizing workflows.

Solution: Business Process Management (**BPM**)

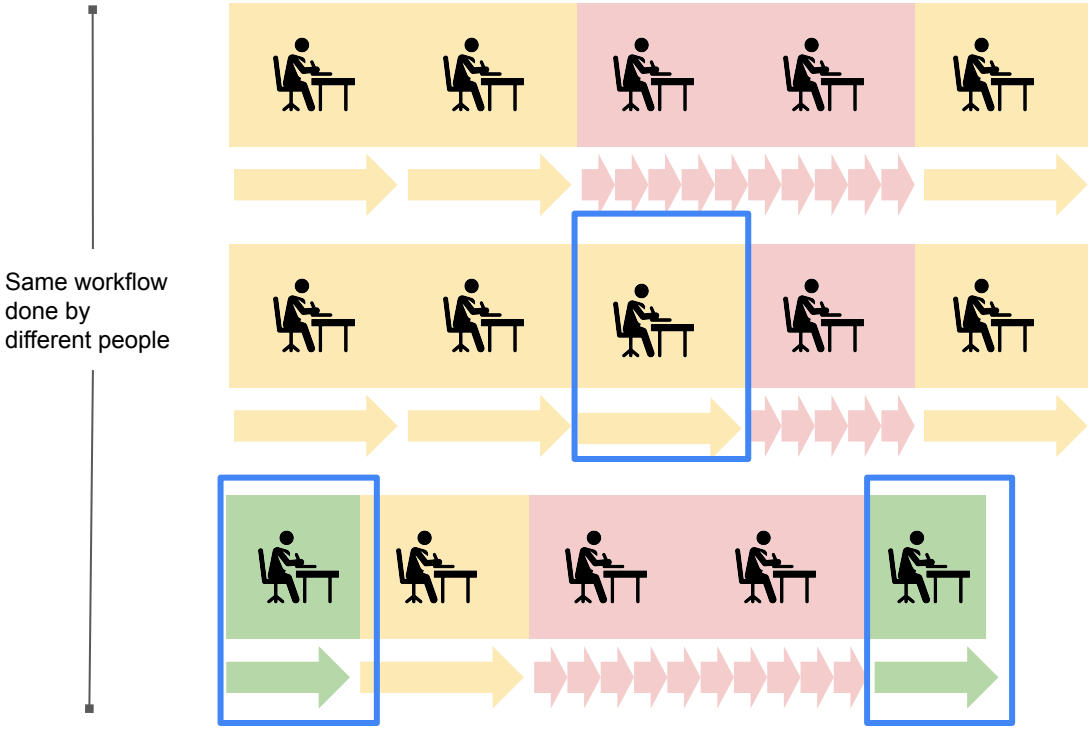
BPM is a systematic approach to make organizations more efficient by measuring and optimizing workflows.



1 Document

Solution: Business Process Management (**BPM**)

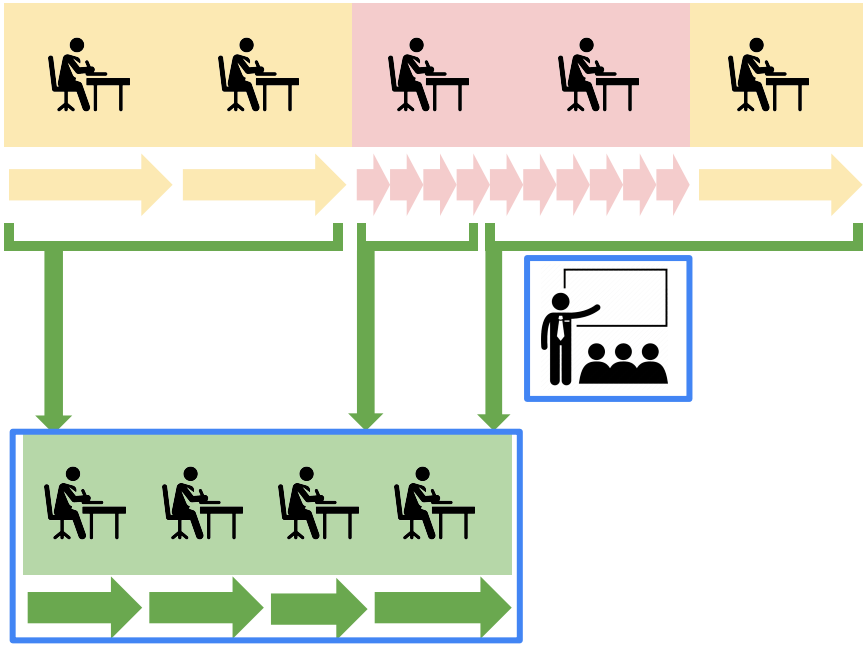
BPM is a systematic approach to make organizations more efficient by measuring and optimizing workflows.



1. Document
2. Knowledge Transfer

Solution: Business Process Management (**BPM**)

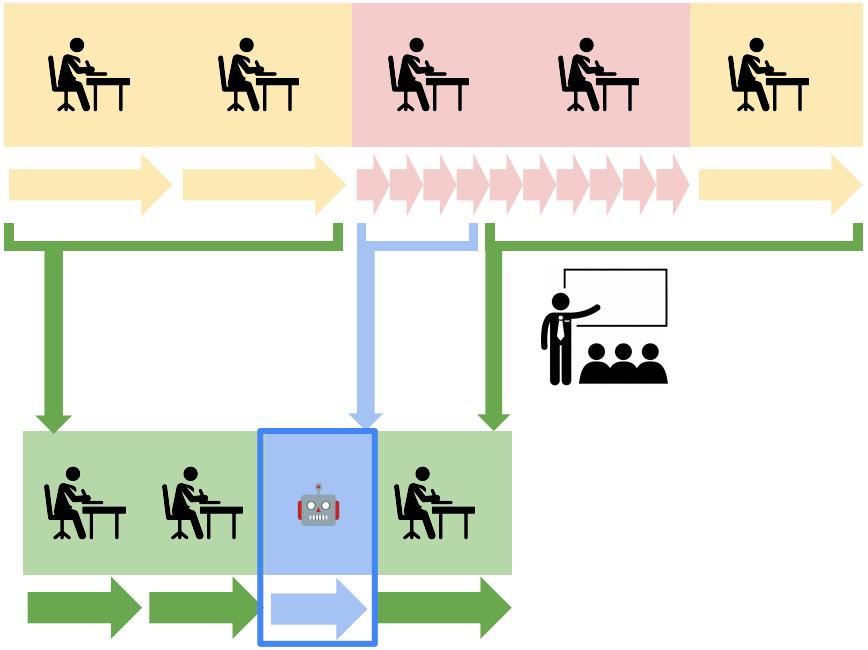
BPM is a systematic approach to make organizations more efficient by measuring and optimizing workflows.



1. Document
2. Knowledge Transfer
3. Improve

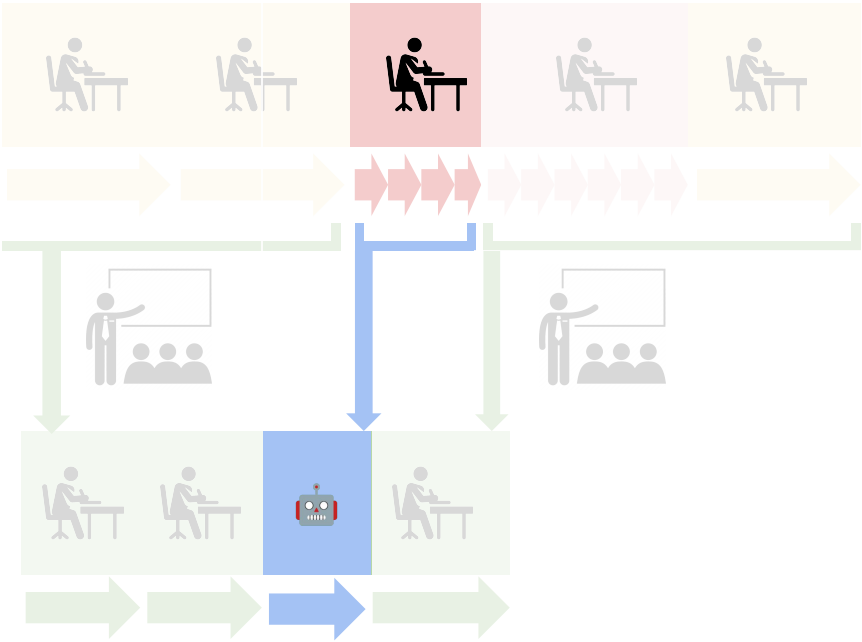
Solution: Business Process Management (**BPM**)

BPM is a systematic approach to make organizations more efficient by measuring and optimizing workflows.



1. Document
2. Knowledge Transfer
3. Improve
4. Automate

Prior work focuses on only **one aspect** of BPM -- **automation**



60%

of a typical BPM project is spent simply defining the workflow

Prior work focuses on only **one aspect** of BPM -- automation

Benchmark	Workflows			Evaluation			Human Demonstrations					
	# Tasks	# Envs	Env Type	Auto	Doc	KT	Imp	Action	Video	SOP	Ranking	Demos/Task
AITW	30,378	357	M	✔	–	–	–	✔	✔	–	–	23.5
Mind2Web	2,350	137	W	✔	–	–	–	✔	✔	–	–	1
MoTIF	6,100	125	M	–	–	–	–	✔	✔	–	–	0.77
WebArena	812	4	W	✔	–	–	–	✔	✔	–	–	0.22
OmniAct	9,802	65	D + W	✔	–	–	–	✔	–	–	–	1
WebShop	12,087	1	W	✔	–	–	–	✔	–	–	–	0.13
VWA	910	3	W	✔	–	–	–	–	–	–	–	0
WorkArena	23,150	5	W	✔	–	–	–	–	–	–	–	0
WebLINX	2,337	155	W	✔	–	–	–	✔	✔	–	–	1
OSWorld	369	13	D + W	✔	–	–	–	✔	✔	–	–	1

Evaluations ignore rest of BPM process

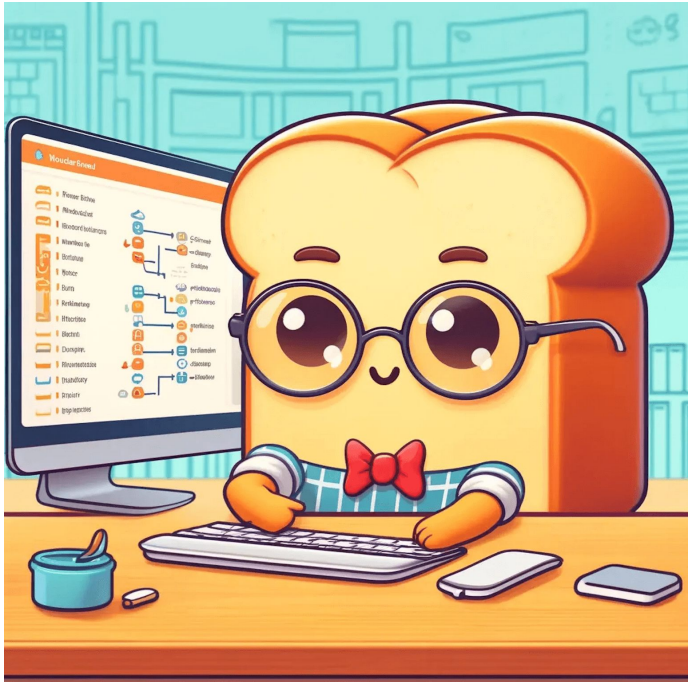
Data does not support BPM tasks

Our work fills these gaps in the agentic workflow literature!

Benchmark	Workflows			Evaluation				Human Demonstrations				
	# Tasks	# Envs	Env Type	Auto	Doc	KT	Imp	Action	Video	SOP	Ranking	Demos/Task
AITW	30,378	357	M	✔	–	–	–	✔	✔	–	–	23.5
Mind2Web	2,350	137	W	✔	–	–	–	✔	✔	–	–	1
MoTIF	6,100	125	M	–	–	–	–	✔	✔	–	–	0.77
WebArena	812	4	W	✔	–	–	–	✔	✔	–	–	0.22
OmniAct	9,802	65	D + W	✔	–	–	–	✔	–	–	–	1
WebShop	12,087	1	W	✔	–	–	–	✔	–	–	–	0.13
VWA	910	3	W	✔	–	–	–	–	–	–	–	0
WorkArena	23,150	5	W	✔	–	–	–	–	–	–	–	0
WebLINX	2,337	155	W	✔	–	–	–	✔	✔	–	–	1
OSWorld	369	13	D + W	✔	–	–	–	✔	✔	–	–	1
Wonderbread	598	4	W	✔	✔	✔	✔	✔	✔	✔	✔	4.9

WONDERBREAD

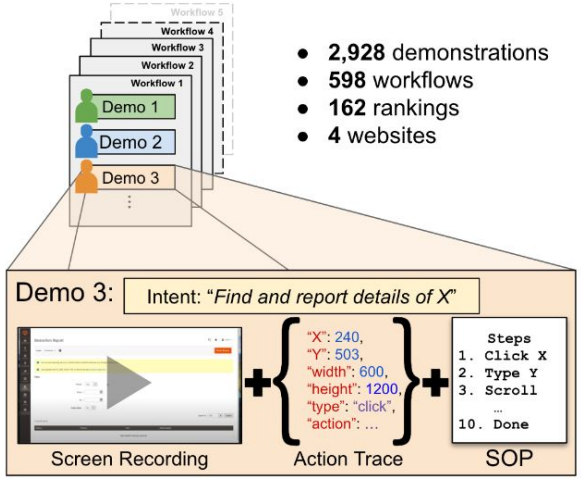
A WOrkflow uNDERstanding BenchmaRk, EvAluation harness, and Dataset



WONDERBREAD is a **benchmark** and **dataset** for studying the **workflow understanding** capabilities of models

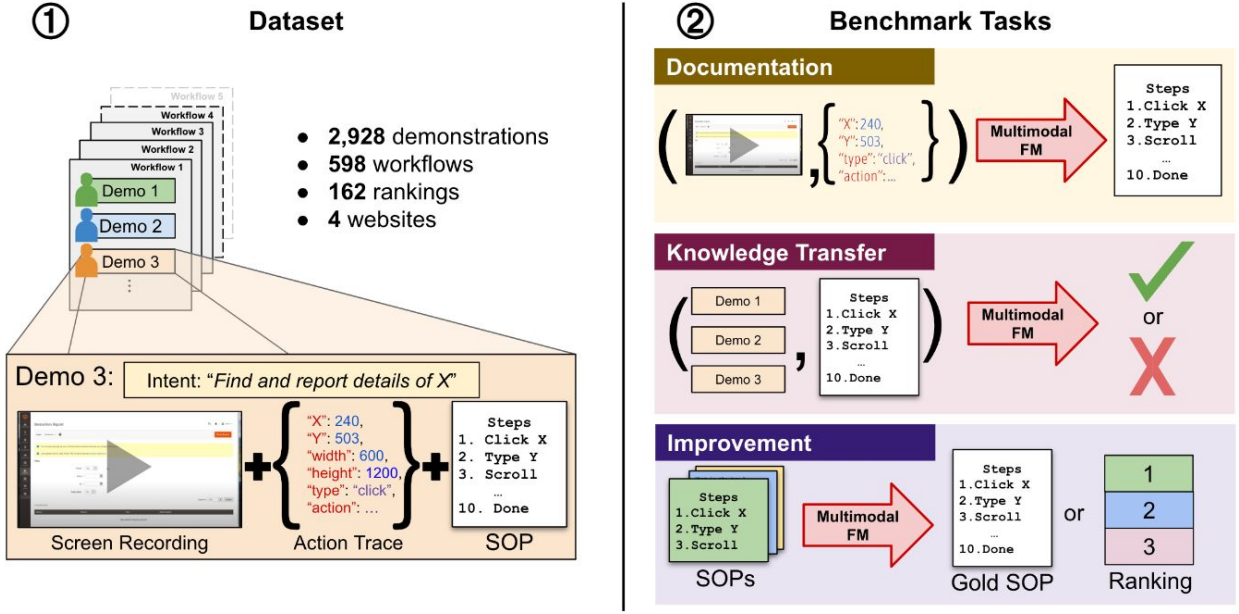
WONDERBREAD is a **benchmark** and **dataset** for studying the **workflow understanding** capabilities of models

① **Dataset**



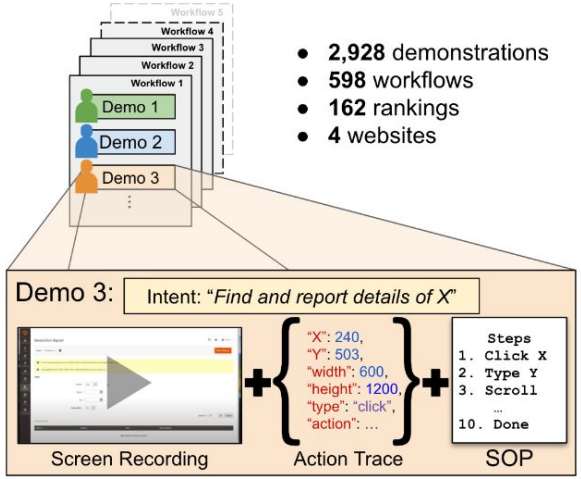
- **2,928** demonstrations
- **598** workflows
- **162** rankings
- **4** websites

WONDERBREAD is a benchmark and dataset for studying the workflow understanding capabilities of models



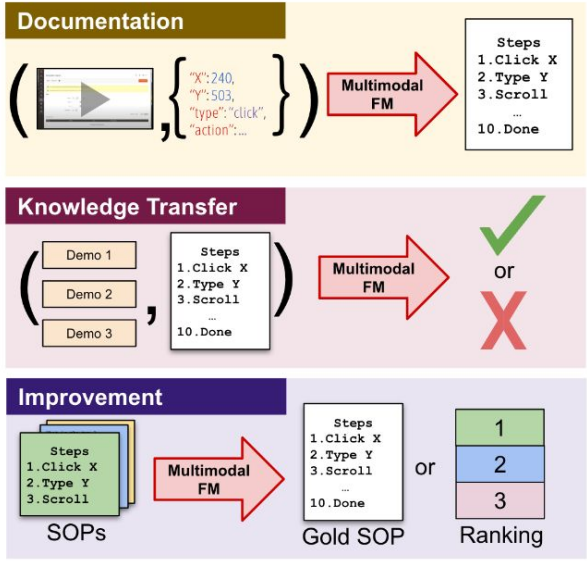
WONDERBREAD is a benchmark and dataset for studying the workflow understanding capabilities of models

① Dataset

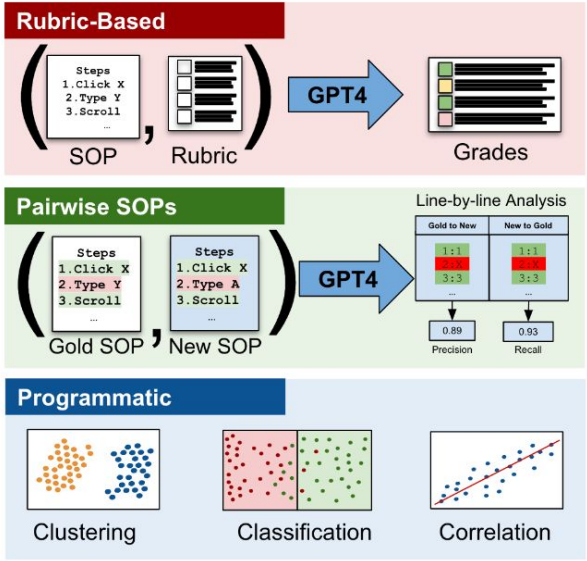


- **2,928** demonstrations
- **598** workflows
- **162** rankings
- **4** websites

② Benchmark Tasks



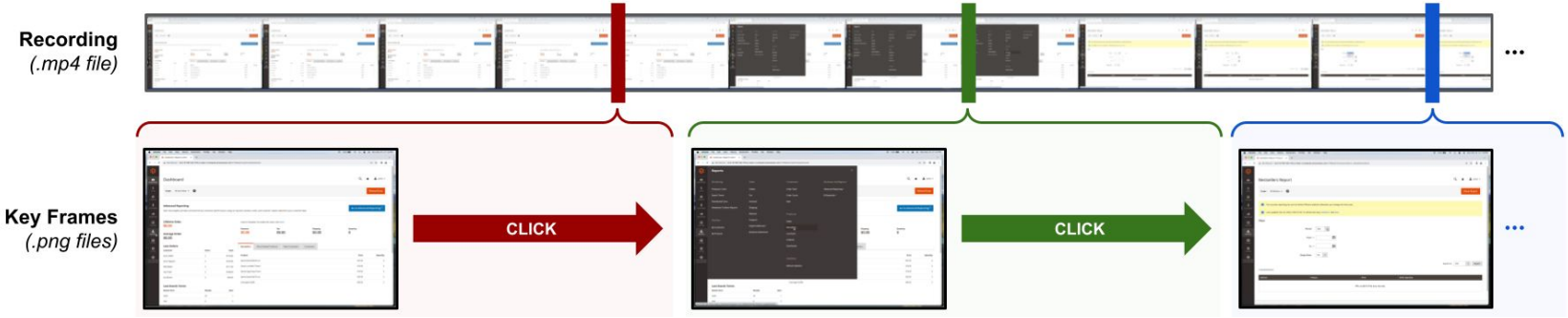
③ Evaluation



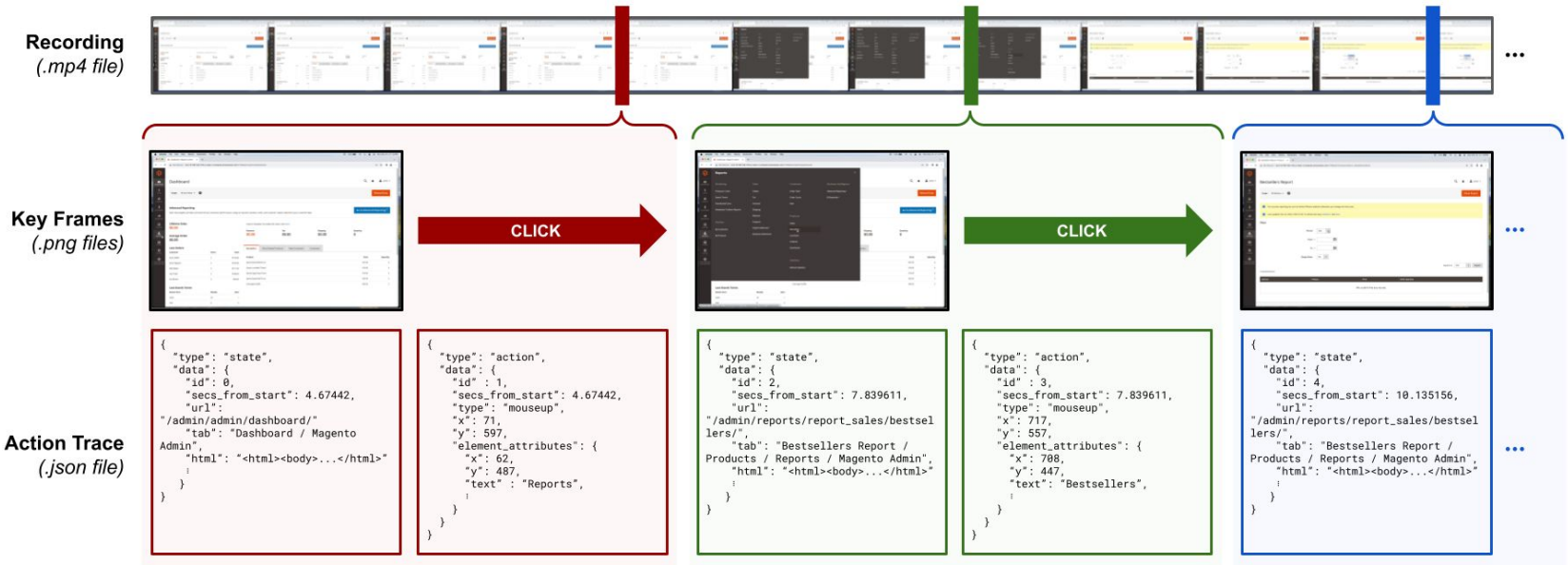
Each **demonstration** contains a full screen recording, extracted key frames, an action log, and a Standard Operation Procedure (SOP)

Each **demonstration** contains a full screen recording, extracted key frames, an action log, and a Standard Operation Procedure (SOP)

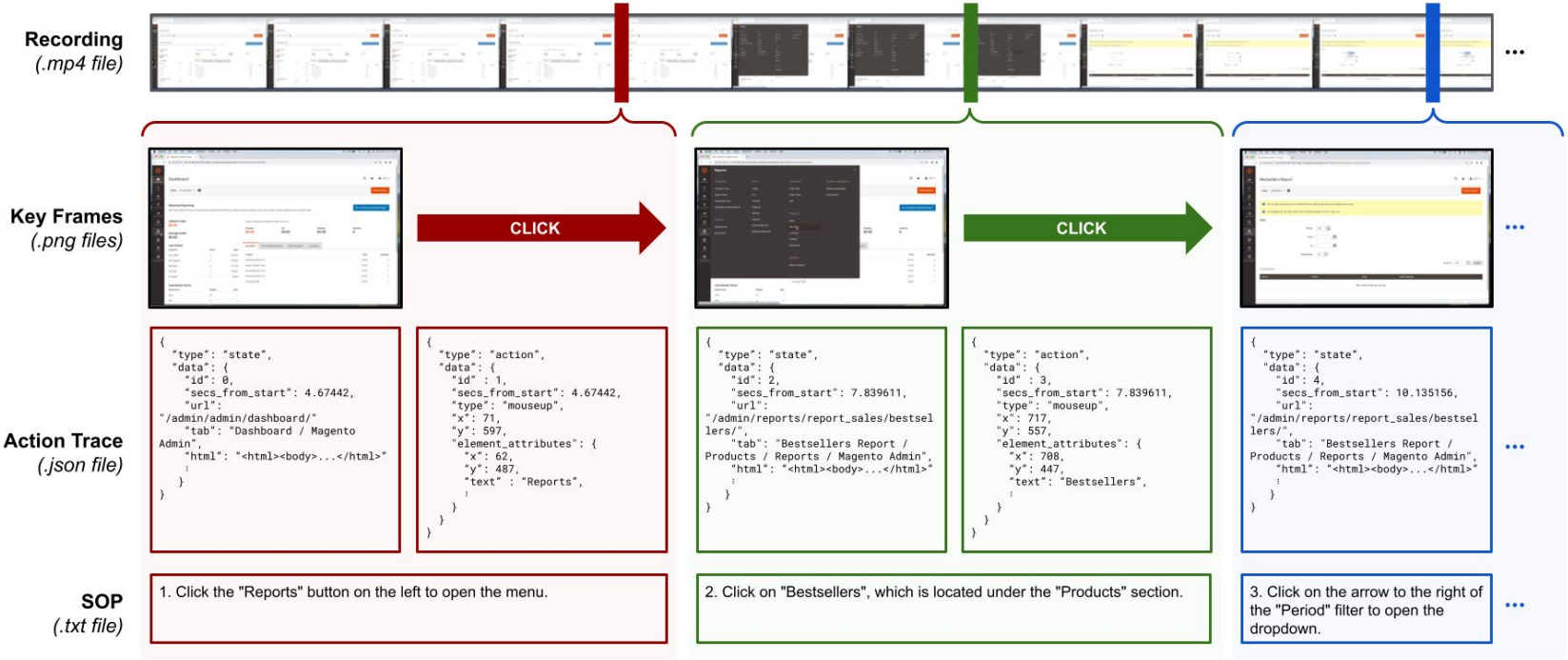
Each **demonstration** contains a full screen recording, extracted key frames, an action log, and a Standard Operation Procedure (SOP)



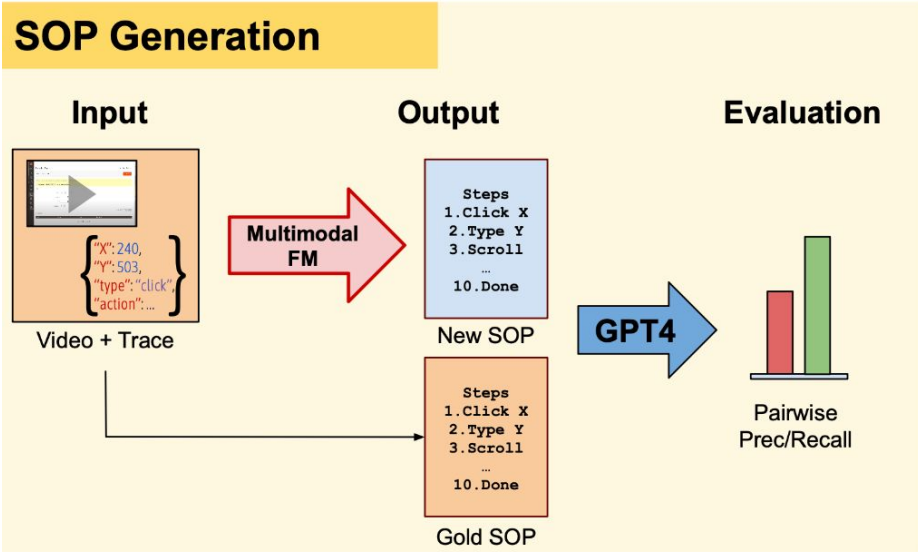
Each **demonstration** contains a full screen recording, extracted key frames, an action log, and a Standard Operation Procedure (SOP)



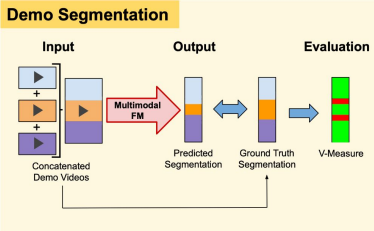
Each **demonstration** contains a full screen recording, extracted key frames, an action log, and a Standard Operation Procedure (SOP)



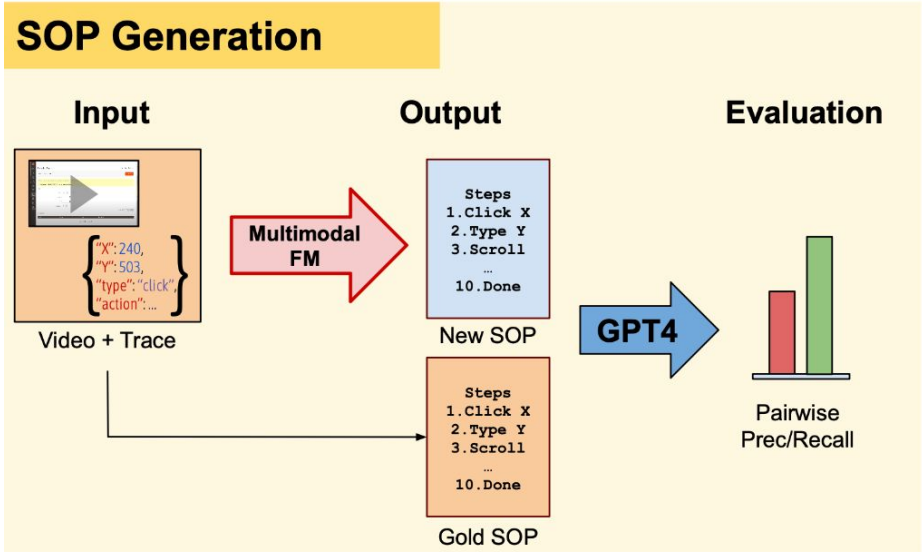
Task Group 1: Automatically **documenting** workflows

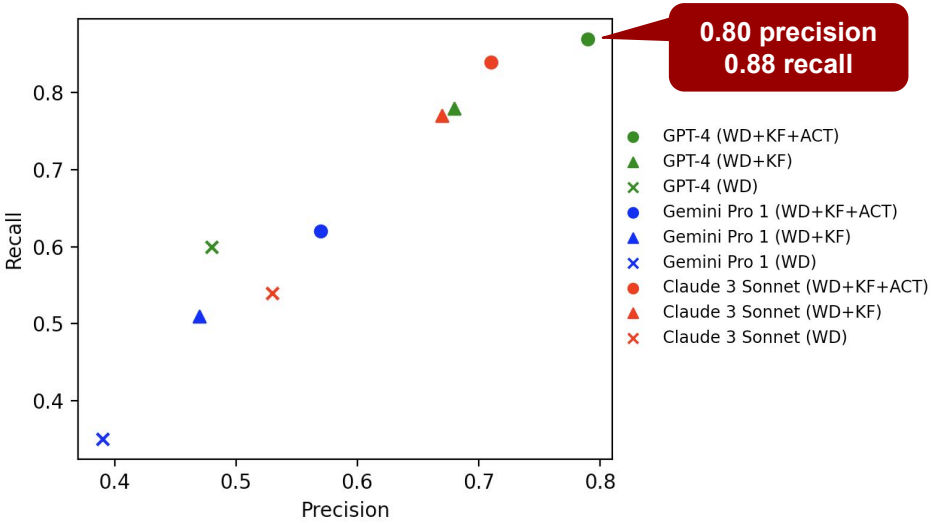


See paper for

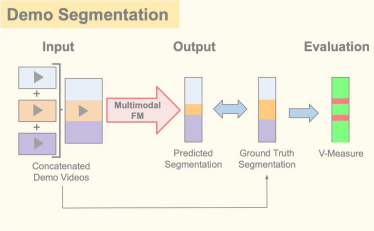


Task Group 1: Automatically **documenting** workflows

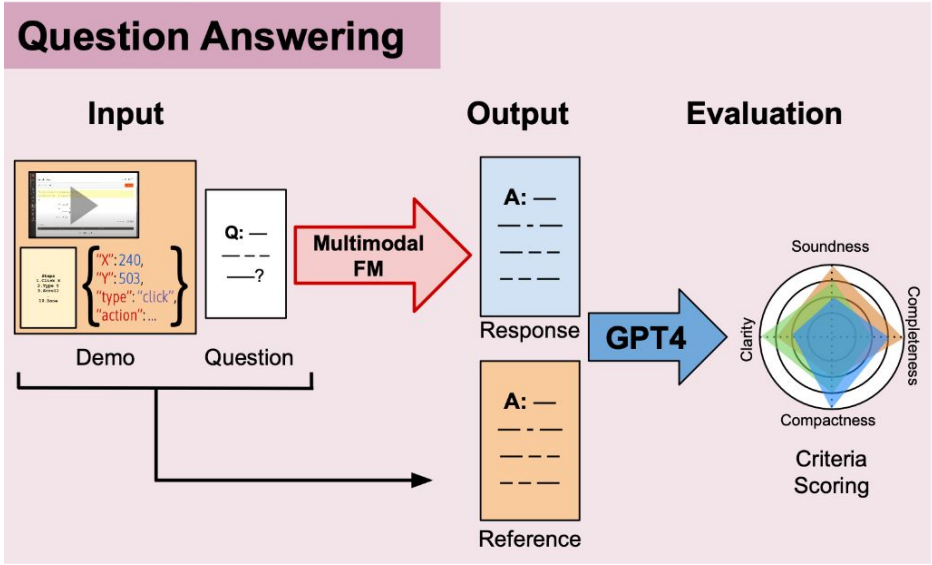




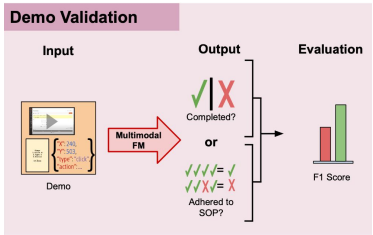
See paper for



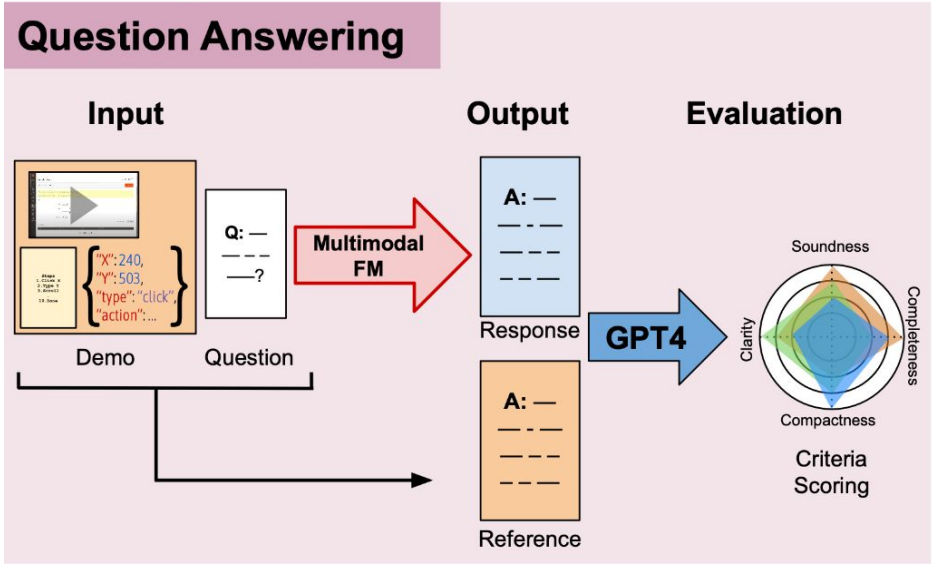
Task Group 2: Facilitating **knowledge transfer**



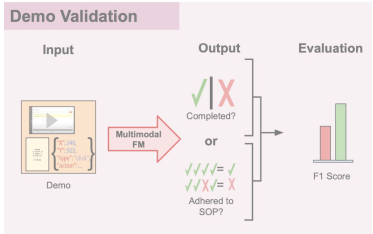
See paper for



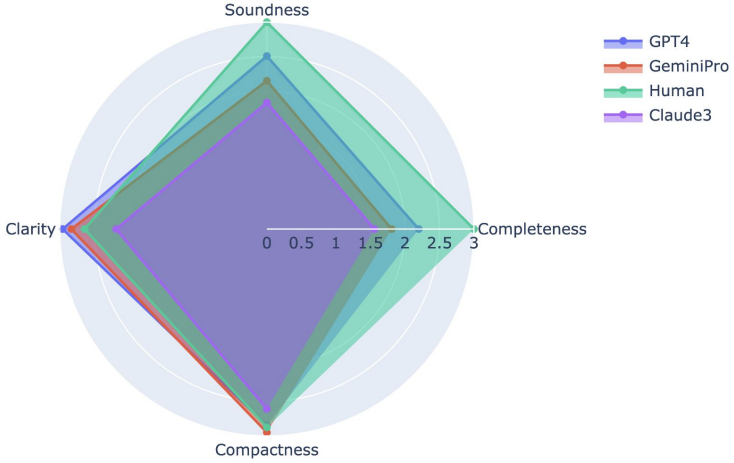
Task Group 2: Facilitating **knowledge transfer**



See paper for



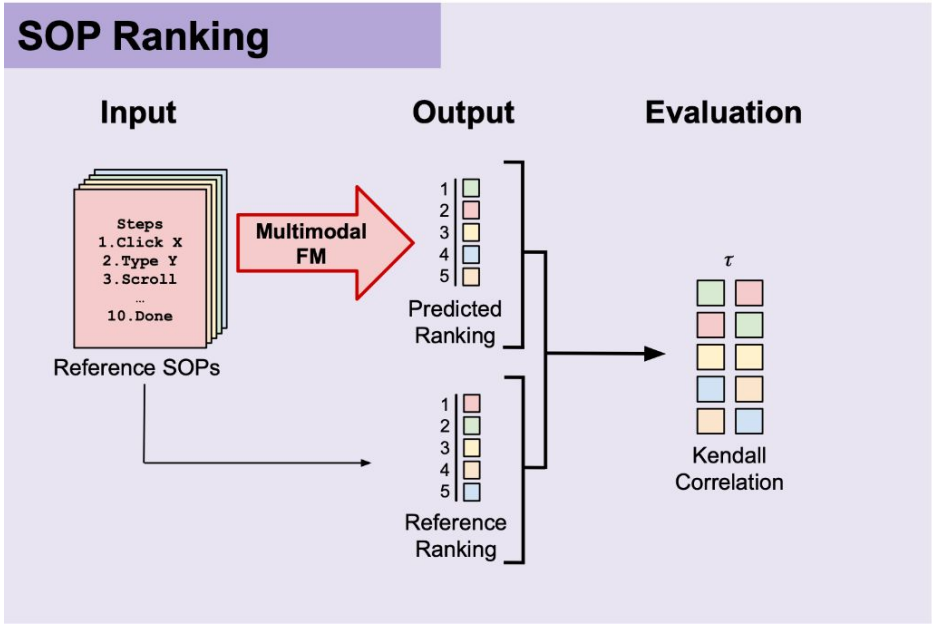
Answers scored by LLM on a scale of **1 (bad)** to **3 (good)**…



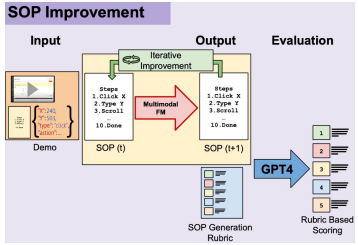
Example questions

- *"Explain what the most common failure modes might be for a user performing this task."*
- *"Here are two demonstrations, one of which is more efficient than the other. Please describe ways to improve the less optimal workflow."*
- *"What is the purpose of doing this workflow?"*

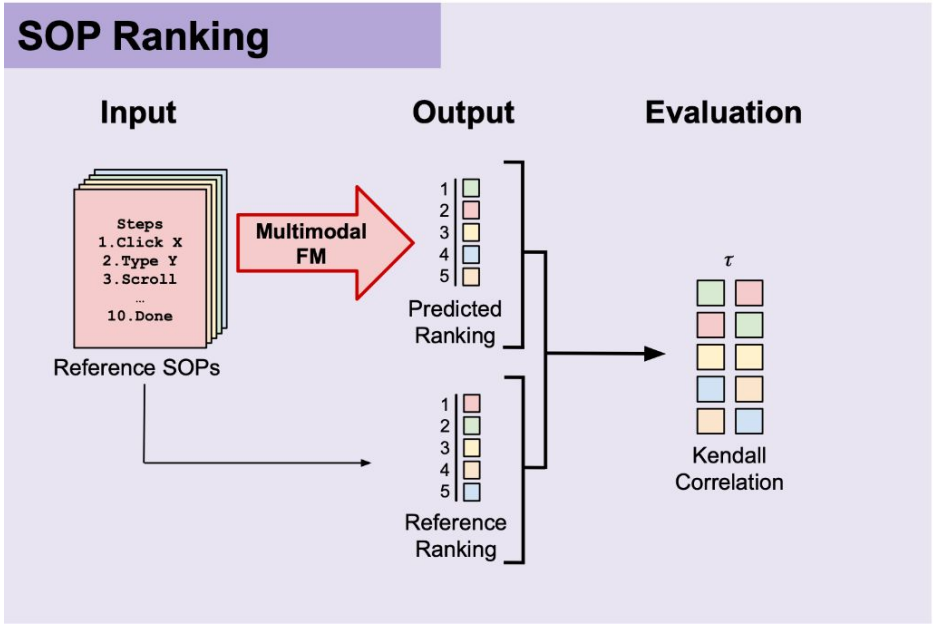
Task Group 3: Ranking and **improving** processes



See paper for

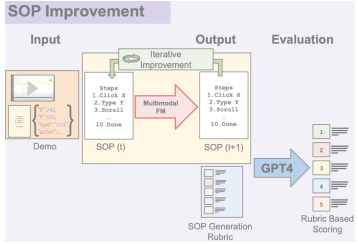


Task Group 3: Ranking and **improving** processes



Model	Spearman ρ	Kendall τ
GPT-4	**0.07 ± 0.58**	**0.06 ± 0.49**
Claude3 Sonnet	0.06 ± 0.59	0.03 ± 0.50
Gemini Pro 1	0.03 ± 0.58	0.03 ± 0.49

See paper for



Future Work + Next Steps

There are many opportunities for future work!

Data Collection

- Increasing **task diversity** + sourcing from **real-world enterprise** applications
- **Automated annotation** of screen recordings

Model Training

- **Fine-tuning AI agents** on our **dataset of 3k high-quality** human demonstrations
- Improving **human-model alignment** of workflow preference

Inference

- Enhancing model **self-validation** to unlock **self-improvement**

…

Thank you!

WONDERBREAD

🌐 Website: https://wonderbread.stanford.edu

📄 Paper: https://arxiv.org/abs/2406.13264

💾 Dataset: https://zenodo.org/records/12671568

💻 Code: https://github.com/HazyResearch/wonderbread

Thanks to our amazing lab mates, advisors & collaborators!

Michael Wornow, Avanika Narayan, Ben Viggiano, Ishan S. Khare, Tathagat Verma, Tibor Thompson, Miguel Angel Fuentes Hernandez, Sudharsan Sundar, Chloe Trujillo, Krrish Chawla, Rongfei Lu, Justin Shen, Divya Nagaraj, Joshua Martinez, Vardhan Agrawal, Althea Hudson, Nigam H. Shah, Christopher Ré

Contact: mwornow@stanford.edu