

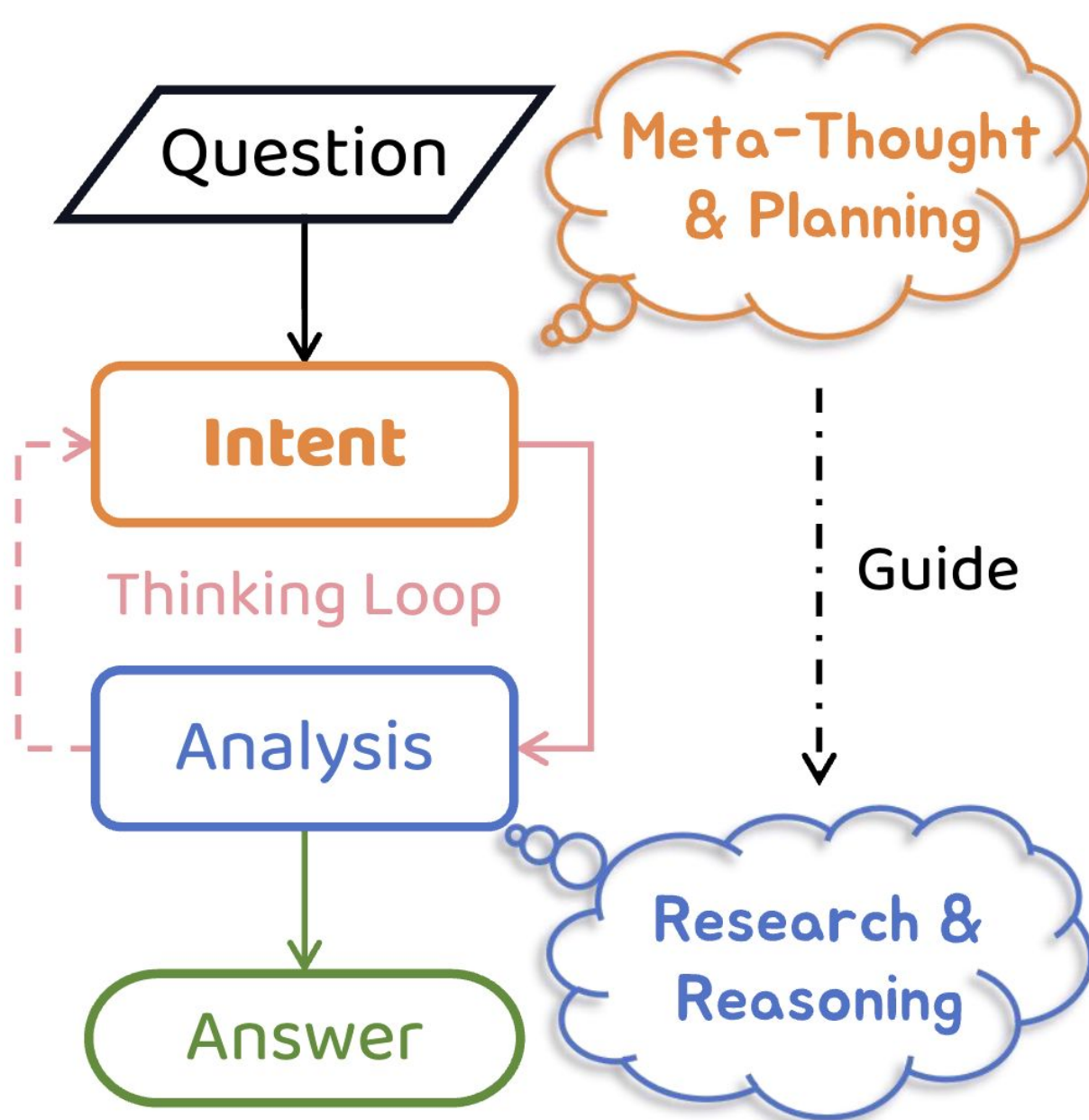


SWI: Speaking with Intent in Large Language Models

Yuwei Yin, EunJeong Hwang, Giuseppe Carenini

Abstract: Intent, typically clearly formulated and planned, functions as a cognitive framework for communication and problem-solving. This paper introduces the concept of Speaking with Intent (SWI) in large language models (LLMs), where the explicitly generated intent encapsulates the model's underlying intention and provides high-level planning to guide subsequent analysis and action. By emulating deliberate and purposeful thoughts in the human mind, SWI enhances the reasoning capabilities and generation quality of LLMs. The promising results in enhancing LLMs with explicit intents pave a new avenue for boosting LLMs' generation and reasoning abilities with cognitive notions.

Why SWI should be helpful?



Human thinking typically follows a structured loop where intent--a mental state or proactive commitment to perform a specific action or produce a particular outcome--directs problem analysis and logical reasoning, therefore also facilitating communication and interaction.

How do LLMs speak with intent?

Problem:

A robe takes 2 bolts of blue fiber and half that much white fiber. How many bolts in total does it take?

LLMs speaking with Intent:

To calculate the total number of bolts of fiber required to make a robe.

To find the total number of bolts of fiber required, we need to determine the amount of white fiber needed and add it to the amount of blue fiber.

To determine the amount of white fiber needed.

Since the robe takes half as much white fiber as blue fiber, and there are 2 bolts of blue fiber, we need to divide 2 by 2 to find the amount of white fiber. $2 / 2 = 1$

To add the amount of blue and white fiber together.

Now that we know the amount of white fiber needed, we can add it to the amount of blue fiber to find the total number of bolts required. 2 (blue fiber) + 1 (white fiber) = 3

Final Answer: 3

SWI Output (Intent + Analysis + Final Summary)

<INTENT> To provide a concise summary of the article about a dog's impressive trick. </INTENT>
A dog named Pakak, a German Shepherd from Alaska, has been filmed performing a skillful trick by Kristina Patrick.

<INTENT> To describe the specific details of the trick. </INTENT>
Pakak lies on her back with a tennis ball between her front paws, lifts it into the air without dropping it, freezes for a few seconds, and then places the ball between her teeth, returning it to its original position.

<INTENT> To highlight the dog's other interests. </INTENT>
Pakak also enjoys playing with balls and has a passion for swimming and being by the sea, as shown in other videos.

<INTENT> To conclude the summary. </INTENT>

Final Summary:

A German Shepherd named Pakak from Alaska has been filmed performing a skillful trick by lifting a tennis ball into the air and placing it between her teeth, showcasing her talent.

Effectiveness of Speaking with Intent

Method	Text Summarization (Average ROUGE-1/2/L/Sum %)					Multi-task QA (Accuracy %)			Math Reasoning (Accuracy %)		
	CDM	XSum	XL-Sum	DialogSum	WikiLingua	BBH	MMLU	MMLU-Pro	GSM8K	GSM8K-P	MATH500
① w/o SWI	23.38	11.90	11.29	16.92	15.01	56.65	52.40	39.27	79.08	81.89	38.20
② SWI	23.91	13.90	13.80	19.57	16.53	63.11	59.22	43.72	80.06	82.88	43.00

Extensive experiments with LLaMA-3-8B on text summarization, multi-task question answering, and math reasoning benchmarks consistently demonstrate the effectiveness and generalizability of Speaking with Intent (SWI) over direct generation without explicit intent (w/o SWI).

System Prompt:

You are a helpful assistant who speaks with intent. During generation, follow all the requirements below:
1. Always explicitly state your own intent before speaking each sentence.
2. Each intent statement should explain the sentence that follows.
3. Your intent must start with the "<INTENT>" tag and end with the "</INTENT>" tag. The content within the intent tags must begin with "To" followed by a verb, such as "To accomplish a task."
4. At last, clearly and concisely give your final answer starting with "Final Answer:"

User Prompt: Speak with intent and answer the following question.
{question}

Analysis & Findings & Insights

Dataset	Method	Precision	Recall	F1
CDM	① w/o SWI	26.06	76.28	36.37
	② SWI	34.22	55.89	37.79
XSum	③ w/o SWI	11.06	48.38	15.15
	④ SWI	14.77	37.30	16.29
XL-Sum	⑤ w/o SWI	8.96	61.88	13.79
	⑥ SWI	12.96	46.72	16.51
DialogSum	⑦ w/o SWI	23.99	57.08	29.55
	⑧ SWI	34.92	45.19	31.20
WikiLingua	⑨ w/o SWI	23.33	65.55	30.63
	⑩ SWI	32.40	54.98	35.78

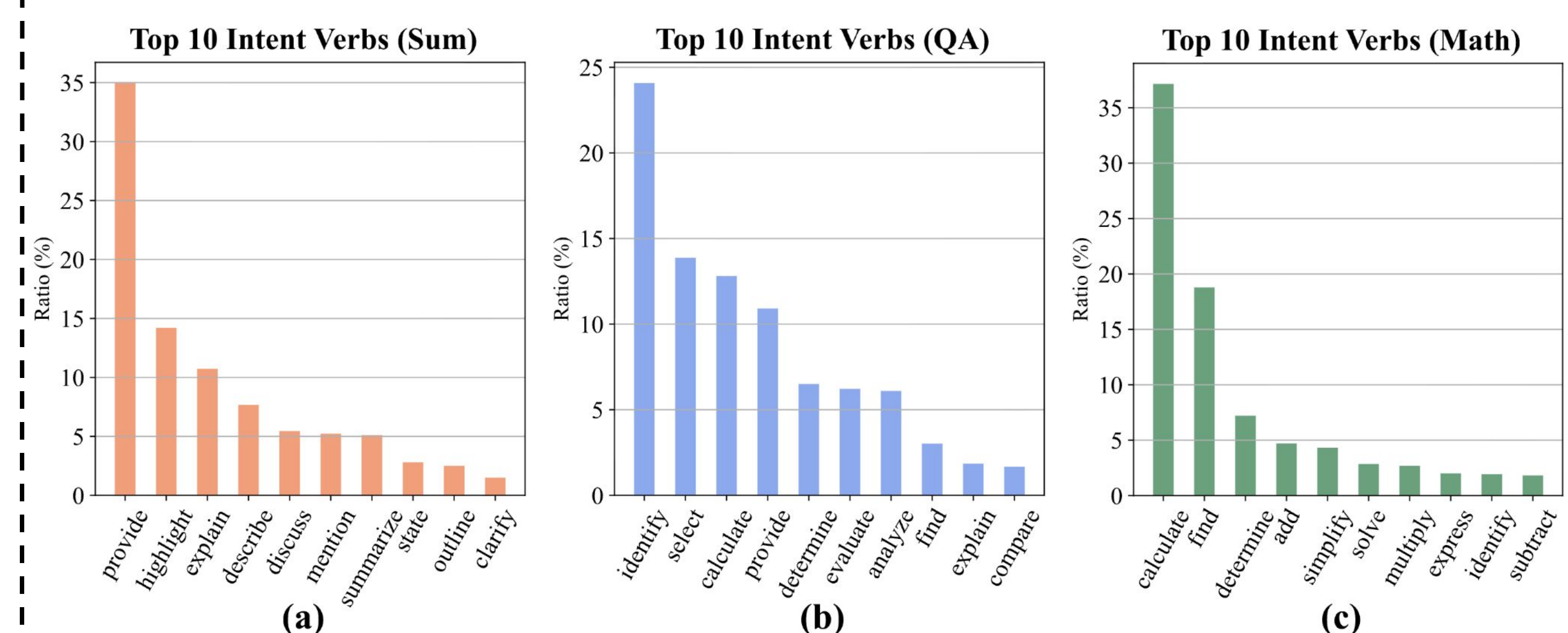
Sec 5.1: SWI-generated summaries exhibit greater accuracy, conciseness, and factual correctness, with fewer hallucinations.

Method	News Article			Dialogue	Wiki Article
	CDM	XSum	XL-Sum	DialogSum	WikiLingua
① w/o SWI	23.38	11.90	11.29	16.92	15.01
② SWI (V0)	23.91	13.90	13.80	19.57	16.53
③ SWI (V1)	24.27	14.12	14.10	19.43	17.34
④ SWI (V2)	24.04	14.69	14.66	19.24	17.06
⑤ SWI (V3)	24.17	13.83	14.02	18.82	16.10

Sec 5.4: As a general framework for problem-solving, SWI brings consistent improvement regardless of its prompt implementation.

Method		News Article			Dialogue	Wiki Article
		CDM	XSum	XL-Sum	DialogSum	WikiLingua
CoT	① w/o SWI	23.17	11.54	11.11	15.77	14.44
	② SWI	24.25	13.86	13.73	19.49	16.88
PS	③ w/o SWI	24.12	12.21	11.91	17.92	15.86
	④ SWI	24.43	12.46	12.28	18.95	16.76

Sec 5.2: SWI works synergistically with existing LLM reasoning & planning methods like Chain-of-Thought (CoT) & Plan-Solve (PS).



Sec 5.7: The intent statistics indicate the need for versatile intent skill sets when performing different tasks (Summarization, QA, & Math).

Model	Method	GSM8K	GSM8K-P	MATH500	Avg.
LLaMA3-3B	① w/o SWI	45.64	46.82	27.20	39.89
	② SWI	65.05	67.58	32.80	55.14
LLaMA3-8B-R1	③ w/o SWI	68.08	70.72	56.40	65.07
	④ SWI	75.44	79.24	57.00	70.56

Sec 5.3: SWI remains effective when applied to models of different model sizes and LLM types (i.e., chat and reasoning models).

Task	Dataset	Coherence		Effectiveness		Interpretability	
		Score	Agree	Score	Agree	Score	Agree
Summarization	CDM	2.83	80%	2.77	70%	2.83	75%
	XSum	2.70	70%	2.60	65%	2.57	65%
Math Reasoning	GSM8K	2.90	85%	2.97	95%	2.97	95%
	MATH500	2.87	80%	2.87	80%	2.83	80%
Multi-task QA	BBH	2.37	55%	2.37	50%	2.33	45%
	MMLU	2.67	75%	2.53	55%	2.37	45%

Sec 5.8: Human evaluations verify the coherence, effectiveness, and interpretability of the intent produced by SWI. (score: 1--3 best)