

SWI: Speaking with Intent in Large Language Models

Yuwei Yin¹, EunJeong Hwang^{1,2}, Giuseppe Carenini¹

¹University of British Columbia; ²Vector Institute for AI
{yuweiyin, ejhwang, carenini}@cs.ubc.ca

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 is hypothesized to enhance the reasoning capabilities and generation quality of LLMs. Extensive experiments on text summarization, multi-task question answering, and mathematical reasoning benchmarks consistently demonstrate the effectiveness and generalizability of Speaking with Intent over direct generation without explicit intent. Further analysis corroborates the generalizability of SWI under different experimental settings. Moreover, human evaluations verify the coherence, effectiveness, and interpretability of the intent produced by SWI. The promising results in enhancing LLMs with explicit intents pave a new avenue for boosting LLMs’ generation and reasoning abilities with cognitive notions.¹

1 Introduction

Intent, the goal-oriented intention in our mind (Adams, 1986; Mele, 1989; Mele and Moser, 1994), serves as a critical component in communication and a guiding framework for problem-solving. As illustrated in Figure 1(a), human thinking (Kahneman, 2011) 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. Hence, we hypothesize that enabling AI systems to speak with their intent explicitly can replicate this

¹Source code: <https://github.com/YuweiYin/SWI>

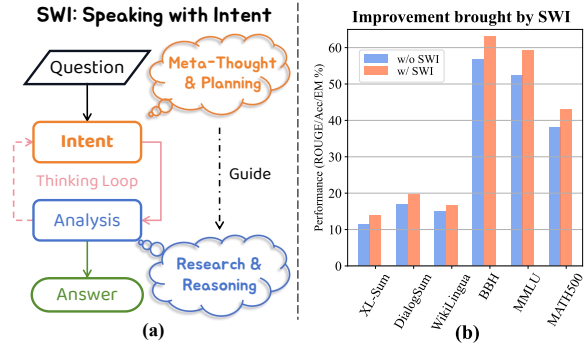


Figure 1: **SWI Overview.** (a) The intent, functioning as meta-thought and planning, guides the analysis with reasoning to answer the question. (b) The performance improvements brought by SWI on various tasks.

meta-cognitive process, thereby improving their generation quality and reasoning ability.

In recent years, large language models (LLMs) (Zhao et al., 2023; Min et al., 2023; Minaee et al., 2024) have revolutionized Natural Language Processing (NLP) with their excellent generative capabilities (OpenAI, 2024; Anthropic, 2024; Gemini, 2024). Enhancing LLMs on various language understanding and logical reasoning tasks (Hendrycks et al., 2021; Suzgun et al., 2023; Lightman et al., 2024; Vendrow et al., 2025) is vital for their ongoing development (Huang and Chang, 2023; Qiao et al., 2023; Patil, 2025).

This work introduces Speaking with Intent (SWI), requiring LLMs to articulate their own intent as a planning mechanism during generation, which makes LLM intentionality more explicit, in a way reminiscent of the long tradition of intent-driven generation in classic NLG (Grosz and Sidner, 1986; Mann and Thompson, 1988; Moore and Paris, 1993; Reiter and Dale, 2000). In essence, we hypothesize that, due to the autoregressive nature of LLMs (Radford et al., 2019) and the attention mechanism (Vaswani et al., 2017), explicitly stated intent provides high-level guidance for subsequent analysis and reasoning. For example, when applying SWI to summarization tasks, each analytical

step in summarizing an article is effectively guided by a preceding intent statement, which is a piece of free-form text generated by instruction-following LLMs (Ouyang et al., 2022; Rafailov et al., 2023; Llama, 2024) instead of a predefined class as in traditional intent modeling (Weld et al., 2022).

In this work, we verify the proposed hypothesis by comprehensively evaluating the effectiveness and generalizability of the proposed SWI method. Specifically, the experimental results across three diverse task categories (i.e., text summarization, multi-task question answering, and mathematical reasoning) demonstrate that speaking with intent in LLMs consistently outperforms directly generating responses without explicit intent. In summarization tasks, SWI produces summaries that are more accurate, concise, and factually reliable, with fewer hallucinations (Ji et al., 2023; Li et al., 2024a) in the output. In math reasoning tasks, SWI surpasses the LLM reasoning method Chain-of-Thought (CoT) (Kojima et al., 2022) and LLM planning method Plan-and-Solve (PS) (Wang et al., 2023), and SWI can work synergistically with these methods to further improve them. Additionally, we perform human evaluations to assess the coherence, effectiveness, and interpretability of the intent generated by our SWI method. Evaluators largely agree on the quality of the generated intent, particularly for mathematical reasoning tasks. The evaluation results confirm that SWI enhances task performance and output explainability.

The key contributions of this work are as follows:

- ❶ We introduce Speaking with Intent in LLMs, where the generated intent effectively guides problem analysis, logical reasoning, and text generation, boosting performance across various benchmarks.
- ❷ Extensive experiments and analyses across diverse task types and multiple datasets, including text summarization, multi-task QA, and mathematical reasoning, demonstrate the consistent effectiveness and generalizability of SWI.
- ❸ Human evaluations validate the coherence, effectiveness, and interpretability of the intent generated by SWI, with our evaluation practice providing standards for assessing freely generated intents.

2 Speaking with Intent

This section presents the problem-solving workflow of LLMs and introduces Speaking with Intent (SWI), enabling LLMs to explicitly articulate their intent during response generation.

2.1 Problem-solving Workflow using LLMs

Let $\mathcal{D} = \{\mathcal{X}, \mathcal{Y}\}$ be a dataset, where $\mathcal{X} = \{X_1, X_2, \dots, X_n\}$ is the input information (questions), $\mathcal{Y} = \{y_1, y_2, \dots, y_n\}$ is the corresponding references (correct answers), and n is the number of instances in $\mathcal{D}$. For text summarization datasets, X_i is the source article, and y_i is one of the reference summaries. For multiple-choice QA datasets, X_i contains the question and options, and y_i is the answer label such as (A)/(B)/(C) or Yes/No. For mathematical reasoning datasets, X_i is the math problem, and y_i is the correct answer (usually an integer number).

In this work, we employ instruction-following LLMs (aka Chat LLMs) $\mathcal{M}$ for experiments and apply the chat template with the system prompt P_s and user prompt P_u . The system prompt specifies the general behavior of the model (assistant), and the user prompt poses questions to the model. Therefore, the generated output $\hat{y}_i$ is obtained by

$$\hat{y}_i = \mathcal{M}(P_s, P_u, X_i; \Theta, \zeta), \quad (1)$$

where P_s is the system prompt, P_u is the user prompt, and X_i is the task input. These string objects are concatenated using line breaks (“\n”) as the delimiter. With parameters Θ and hyperparameters ζ , the LLM $\mathcal{M}$ generates new tokens one by one until reaching the generation limit or generating the “end-of-text” special token provided by the tokenizer.

2.2 LLM Speaking with Intent

SWI is a novel method that brings a significant cognitive concept (i.e., intent) into LLM generation. SWI is implemented in a simple and reproducible approach, i.e., we require LLMs to speak with intent (SWI) by presenting detailed instructions in the system prompts P_s and restating the SWI requirement in the user prompt P_u . Table 1 presents the SWI instructions for QA and math tasks. For summarization tasks, the fourth requirement in the system prompt is *At last, clearly and concisely give your final summary starting with "Final Summary:"* and the user prompt becomes *Speak with intent and summarize the following article. {{article}}*.

As showcased in Figure 2 (math) and Figure 3 (summarization), LLMs that speak with intent articulate their **intents** during the thinking, reasoning and communication process and then provide the final answer based on the analysis.

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}}

Table 1: SWI instructions for QA and math tasks.

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

Figure 2: An example of LLMs speaking with intent for mathematical reasoning tasks.

2.3 Result Evaluation

We extract the final answer (denoted as $\tilde{y}_i$) from the model outputs $\hat{y}_i$ and compute the overall performance of $\mathcal{M}$ on the dataset $\mathcal{D}$ by

$$s = \frac{1}{n} \sum_{i=1}^n \mathcal{S}(y_i, \tilde{y}_i), \quad (2)$$

where the score function $\mathcal{S}(\cdot, \cdot)$ returns a value in the range of $[0, 1]$. Different tasks adopt different score functions to evaluate the model performance.

To evaluate the quality of summaries, we apply the standard ROUGE (Lin, 2004) as the automatic evaluation metric $\mathcal{S}$ and complement it with a more sophisticated fact-checking analysis as described in Section 4.1. For multiple-choice QA tasks, we adopt the Option Selection metric introduced by Yin and Carenini (2025), which evaluates the LLM perplexity of different option concatenations and selects the one with lowest perplexity as the model’s

Task	Dataset	Split	Size
Sum	CDM	Test	11,490
	XSum	Test	11,334
	XL-Sum	Test	11,535
	DialogSum	Test	1,500
	WikiLingua	Test	3,000
QA	BBH	Test	5,511
	MMLU	Test	13,842
	MMLU-Pro	Test	12,032
Math	GSM8K	Test	1,319
	GSM8K-P	Test	1,209
	MATH500	Test	500

Table 2: Dataset Statistics.

choice. For mathematical reasoning tasks, we first extract numbers in $\tilde{y}_i$ and apply text normalization to both $\tilde{y}_i$ and the reference y_i , and then conduct exact match to check if the generated answer $\tilde{y}_i$ is correct.

3 Experimental Setup

This section presents the experimental setup, including Tasks and Datasets (§ 3.1), Model Settings (§ 3.2), and Baseline Settings (§ 3.3).

3.1 Tasks and Datasets

The proposed SWI method aims to enhance the generation quality and reasoning ability of LLMs. To comprehensively study the effectiveness and generalizability of SWI, we conduct extensive experiments on various challenging benchmarks of three diverse task types: text summarization (Sum), multi-task question answering (QA), and mathematical reasoning (Math). The dataset statistics are presented in Table 2.

Text Summarization (Sum). We hypothesize that Speaking with Intent benefits natural language generation tasks like summarization, where the generated intent can guide the model in summarizing the source article point by point in an orderly fashion. Hence, we test the effect of SWI on the following text summarization datasets covering different genres: CNN/DailyMail (CDM) (Hermann et al., 2015; See et al., 2017), Extreme summarization (XSum) (Narayan et al., 2018), XL-Sum (Hasan et al., 2021), DialogSum (Chen et al., 2021), and WikiLingua (Ladhak et al., 2020).

Multi-task Question Answering (QA). Our SWI method is also evaluated on various multi-task question answering datasets, including BIG-Bench Hard (BBH) (Suzgun et al., 2023),

MMLU (Hendrycks et al., 2021), and MMLU-Pro (Wang et al., 2024b). They are all reasoning-intensive benchmarks designed as multiple-choice QA tasks, where the model is asked to select the most appropriate one from the given options to answer the question. Here, the hypothesis is that generating its intent explicitly as text improves the system’s question analysis abilities.

Mathematical Reasoning (Math). Beyond multiple-choice QA tasks, we also consider mathematical reasoning benchmarks, where the model is asked to solve the given math problem and present the final answer (numerical values). We consider representative and high-quality math benchmarks, including Grade School Math 8K (GSM8K) (Cobbe et al., 2021), GSM8K-Platinum (Vendrow et al., 2025), and MATH500 (Lightman et al., 2024). Again, our hypothesis is that speaking with explicit intent improves the model’s reasoning abilities.

3.2 Model Settings

Language Models. By default, we employ LLaMA3-8B-Instruct (Llama, 2024) as the language model $\mathcal{M}$ for generation and evaluation. It is an open-weights, instruction-following, and Transformer-based (Vaswani et al., 2017) LLM with 8 billion model parameters. We load the model checkpoint and tokenizer provided by Hugging Face Transformers (Wolf et al., 2020). To further assess the generality of SWI, we also evaluate the efficacy of SWI with different LLMs in § 5.3. In the fact-checking evaluation for summaries (§ 5.1), we adopt GPT-4o-mini (OpenAI, 2024) to decompose the generated summary and reference summary into two sets of atomic facts.

Generation Configurations. Each experiment session was conducted on a single NVIDIA H100 GPU, and all the models were loaded in a half-precision mode (float16). The input sequence is not truncated to avoid losing context information, while we set the maximum number of newly generated tokens to 4096 during generation.

Reproducibility Statement. To guarantee reproducibility, we fixed the seeds to 42 for all random modules, set the LLM generation temperature to 0 for deterministic generation without sampling, and ran all experiments three times, obtaining reproducible generation outputs and evaluation scores. The source code is available on GitHub.

3.3 Baseline Settings

The main comparison is LLM generation with intent or without intent, and the effectiveness of SWI is verified if the former outperforms the latter.

In § 5.2, we also investigate the synergy between SWI and existing LLM reasoning & planning methods, i.e., Chain-of-Thought (CoT) (Kojima et al., 2022) and Plan-and-Solve (PS) (Wang et al., 2023). CoT aims to elicit LLM reasoning using the answer-trigger prompt Φ_i^{CoT} as “Let’s think step by step”, while PS applies the following prompt Φ_i^{PS} to construct plans before problem-solving: “Let’s first understand the problem and devise a plan to solve the problem. Then, let’s carry out the plan and solve the problem step by step.” With such answer-trigger prompts Φ_i , the generation process (Eq. 1) is given by

$$\hat{y}_i = \mathcal{M}(P_s, P_u, X_i, \Phi_i; \Theta, \zeta), \quad (3)$$

where P_s is the system prompt, P_u is the user prompt, and X_i is the task input. Θ and ζ are the parameters and hyper-parameters of the LLM $\mathcal{M}$, respectively.

4 Main Results

This section presents the experimental results to verify the effectiveness of SWI on diverse generation and reasoning tasks.

4.1 Text Summarization

First, we demonstrate that SWI benefits natural language generation tasks like summarization by more explicitly analyzing the source document point by point and better planning the generation of the final summary.

We evaluate the quality of summaries using the ROUGE score (Lin, 2004), which counts the overlaps of the generated summaries and reference summaries. Specifically, we average the ROUGE-1 (unigrams), ROUGE-2 (bigrams), ROUGE-L (longest common subsequences), and ROUGE-LSum (sentence-level ROUGE-L) scores as the final ROUGE score. As shown in Table 3, our SWI method consistently surpasses the direct generation baseline (“w/o SWI”), confirming its effectiveness in enhancing the quality of text summaries.

4.2 Multi-task Question Answering

Beyond text understanding and generation, we also consider multi-task QA tasks. We test the

Method	Text Summarization (Average ROUGE-1/2/Lsum %)					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

Table 3: **Main results** on text summarization, multi-task QA, and mathematical reasoning tasks.

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

Table 4: **Fact Checking Evaluation of Summaries.** We compare the atomic facts drawn from the LLM-generated summaries and the golden references, and compute recall, precision, and F1 scores (%).

effect of SWI on three large-scale and challenging benchmarks designed as multiple-choice QA tasks, where the model is asked to select the most appropriate one from the given options to answer the question. Table 3 shows that our SWI method consistently improves the direct generation baseline by a large margin. The results demonstrate the efficacy of SWI in reasoning-intensive QA tasks.

4.3 Mathematical Reasoning

Additionally, we explore the efficacy of SWI on three high-quality math benchmarks. Unlike multiple-choice QA in § 4.2, where the model picks an option from the given list, math tasks require LLMs to generate numerical values as the answer. As shown in Table 3, SWI consistently improves the model performance over direct generation, showing its effectiveness in enhancing LLM on problem analysis and mathematical reasoning.

5 Analysis

5.1 Fact Checking of Summaries

LLMs frequently generate hallucinated content (Ji et al., 2023; Li et al., 2024a), which can not be detected by lexical metrics like ROUGE. To assess this issue, we adopt a more semantically sophisticated fact-checking metric (Hwang et al., 2025), which quantifies factual consistency by calibrating the extent of fabricated statements (low precision) and omitted factual information (low re-

	Method	GSM8K	GSM8K-P	MATH500	Avg.
CoT	① w/o SWI	77.86	80.07	42.00	66.64
	② SWI	80.21	82.55	42.80	68.52
PS	③ w/o SWI	72.56	75.35	40.00	62.64
	④ SWI	79.45	82.54	41.40	67.80

Table 5: **LLM Reasoning & Planning with SWI.** When additional LLM reasoning (CoT) and planning (PS) methods are adopted, the exact matching scores (%) on multiple **math** datasets with or without SWI.

	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

Table 6: **LLM Reasoning & Planning with SWI.** When additional LLM reasoning (CoT) and planning (PS) methods are adopted, the ROUGE scores (%) on multiple **summarization** datasets with or without SWI.

call). Specifically, we use GPT-4o-mini (OpenAI, 2024) to decompose both generated and reference summaries into atomic fact sets and measure their coverage to quantify factual consistency.

We evaluate 100 random samples from each summarization dataset using this fact-checking metric, with results presented in Table 4. Directly generated summaries (“w/o SWI”) tend to be more lengthy and verbose, resulting in higher recall scores. In contrast, SWI-generated summaries exhibit greater accuracy, conciseness, and factual correctness, with fewer hallucinations. Overall, SWI consistently outperforms the direct generation baseline in terms of F1 score.

5.2 Synergy with Other Methods

In recent years, various methods have been proposed to boost the reasoning and planning abilities of LLMs. Since our SWI method is orthogonal to previous work, it is necessary to compare the performance and study the synergy between SWI and them. As mentioned in § 3.3, we adopt representative LLM reasoning method Chain-of-Thought (CoT) (Kojima et al., 2022) and LLM planning method Plan-and-Solve (PS) (Wang et al., 2023).

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

Table 7: **Generalizability of SWI to different LLMs.** When different sizes and types of LLMs are adopted, the exact matching scores (%) on multiple mathematical reasoning datasets with or without SWI.

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

Table 8: **Results of Different SWI Prompt Variants.** When different paraphrases of SWI prompts are adopted, the ROUGE scores (%) on various text summarization datasets with or without Speaking with Intent (SWI).

When additional reasoning (CoT) and planning (PS) methods are adopted, Table 5 and Table 6 present the LLM performance on multiple math and summarization tasks, respectively. Comparing the results in Table 3 and Table 5, our SWI method (② in Table 3) beats CoT (① in Table 5) and PS (③ in Table 5). Moreover, the combination of CoT+SWI (② in Tables 5 and 6) boosts the CoT method (i.e., ②>①), and the synergy of PS+SWI (④ in Tables 5 and 6) also improves the PS-alone performance (i.e., ④>③). These results verify that SWI works synergistically with existing LLM reasoning & planning methods.

5.3 Generalizability to Different LLMs

To further validate the generalizability of our SWI method, we evaluate its effect on different sizes and types of LLMs. Aside from the results of LLaMA3-8B (Table 3), we present the results on multiple mathematical reasoning tasks using LLaMA3-3B (with 3B parameters) and LLaMA3-8B-R1, which is fine-tuned using reasoning data distilled from DeepSeek R1 (Guo et al., 2025). As observed in Table 7, our SWI method brings consistent improvements over the direct generation baseline, verifying the effectiveness of SWI when applied to models of different model sizes and LLM types (i.e., chat and reasoning models).

5.4 SWI Prompt Variants

In this work, we implement SWI in a straightforward prompting way for simplicity and repro-

Task	Dataset	# Input Tokens			# Output Tokens		
		w/o SWI	SWI	Δ	w/o SWI	SWI	Δ
Sum	CDM	920	1028	+108	166	434	+161%
	XSum	542	650	+108	135	374	+177%
	XL-Sum	613	721	+108	150	405	+170%
	DialogSum	263	371	+108	77	238	+209%
	WikiLingua	525	633	+108	145	386	+166%
QA	BBH	203	313	+110	96	244	+154%
	MMLU	193	303	+110	63	229	+263%
	MMLU-Pro	270	380	+110	493	703	+43%
Math	GSM8K	123	233	+110	211	280	+33%
	GSM8K-P	123	233	+110	213	273	+28%
	MATH500	133	243	+110	810	922	+14%

Table 9: **Efficiency Study.** The number of input and output tokens with or without SWI.

ducibility. To demonstrate that SWI works effectively irrespective of specific prompt design, we conduct experiments on different SWI prompt variants. Specifically, the original SWI prompt (“V0”) is paraphrased into three different versions (details in the appendix) by GPT-4o (OpenAI, 2024). Table 8 shows the performance of each SWI variants over multiple summarization tasks. As illustrated, the proposed SWI method, regardless of its prompt implementation, maintains a consistent advantage over the direct generation baseline, substantiating that SWI is an effective and general framework that brings intent—a key cognitive concept for reasoning and communication—into LLM generation.

5.5 Efficiency Study

To study the efficiency cost of requiring LLMs to speak with intent, we present the number of input and output tokens with or without SWI in Table 9. SWI adds about 110 extra input tokens in the system prompt (as in Table 1) for additional instructions, and the extra output tokens are not more than 300 tokens (percentages in Table 9) on average compared to direct generation. Hence, this cost should be traded off against the substantially improved performance, transparency, and interpretability of SWI.

5.6 Case Study

In addition to the math example in Figure 2, we conduct a case study on the summarization task to provide more insights into the benefits of SWI. Figure 3 showcases the SWI outputs, where the generated intent is well formulated and articulated, guiding a point-by-point summarization process that leads to a final summary that is accurate, concise, and abstractive, while effectively capturing the key information of the source article. This probably leads to the high Precision and F1 scores of

Source Article

Shhh! Don't make a sound. This dog is lost in deep concentration. Kristina Patrick from Alaska filmed her German Shepherd Pakak performing a very skillful trick. Video footage shows the pup lying on her back with a tennis ball neatly clutched between her front paws. Keeping her eyes fixed on the toy, she then proceeds to lift it into the air being careful not to drop it. After freezing still for several seconds, Pakak lowers the ball slowly down. She then places it in between her teeth, back to where it started. Pakak's owner says she loves playing with balls. Other videos reveal the pup also has a passion for swimming and being by the sea. Pet's got talent: Kristina Patrick from Alaska filmed her German Shepherd, Pakak, performing a skillful trick. Easy does it: Video footage shows the pup lying on her back with a tennis ball neatly clutched between her front paws - she then proceeds to lift it into the air being careful not to drop it.

Reference Summary

Kristina Patrick from Alaska filmed her German Shepherd Pakak performing a very skillful trick. Footage shows the pup taking the ball from her mouth with her paws and holding it up high in the air to admire it. She then carefully lowers it back down to the starting point.

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.

Figure 3: **Case Study.** The source article, reference summary, and SWI output of a text summarization example.

Task	Dataset	Total	Unique	Per Ins.
Sum	CDM	53,086	86	4.6
	XSum	50,697	92	4.5
	XL-Sum	56,100	84	4.9
	DialogSum	4,739	96	3.2
	WikiLingua	12,463	91	4.2
QA	BBH	7,425	49	1.3
	MMLU	19,380	106	1.4
	MMLU-Pro	21,916	122	1.8
Math	GSM8K	5,237	45	4.0
	GSM8K-P	4,719	43	3.9
	MATH500	3,619	84	7.2

Table 10: **Intent Statistics.** The number of total and unique intents (verbs) of each dataset, as well as the average number of intents per instance (“Per Ins.”).

SWI in Table 4, suggesting the validity of SWI in text generation tasks. Similarly, SWI enables LLM to have progressive planning when solving math problems. This is a critical ability when dealing with complex problems that require the divide-and-conquer strategy.

5.7 Intent Statistics

To further analyze the pattern and variability of the intents generated by our SWI method, we present the intent statistics across different tasks. Specifically, we extract and count the verbs in the generated intent statements, which are required to be in the format “To do something.”. Table 10 shows the number of total verbs (“Total”) and unique verbs (“Unique”) of each dataset, as well as the average number of intents per instance (“Per Ins.”). We observe that the number of unique intents in the summarization task is larger, indicating that summarizing documents demands a higher intent vari-

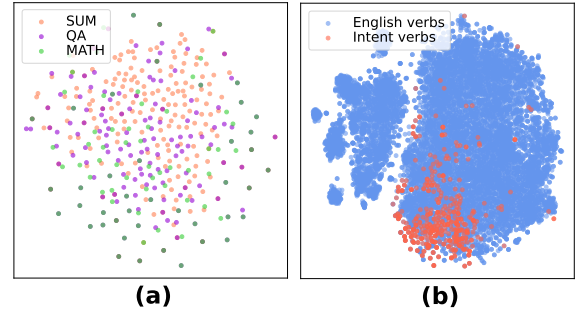


Figure 4: **The semantic distribution of intents** across different task types and among all English verbs.

ability. In addition, summarization and math tasks generally have more intents per instance than QA, likely due to their longer outputs, as also observed in the efficiency study (Table 9). Among the three math datasets, MATH500 is relatively harder, as the model performance is lower in Table 3. Thus, solving the problems in MATH500 requires more thinking steps and longer reasoning chains, which is consistent with the observation that its number of unique intent verbs and the average number of intents per instance are larger than GSM8K.

Furthermore, we investigate the intents across different task types and the distribution of intent verbs among all English verbs. First, we feed the same model that generates the outputs (i.e., LLaMA3-8B) with each verb and extract the last-layer hidden states, which indicate how the generator perceives and utilizes the intents. Then, t-SNE (van der Maaten and Hinton, 2008) is applied to visualize the semantic representations of each unique intent verb and all 11,531 English verbs drawn from WordNet (Miller, 1992). We observe

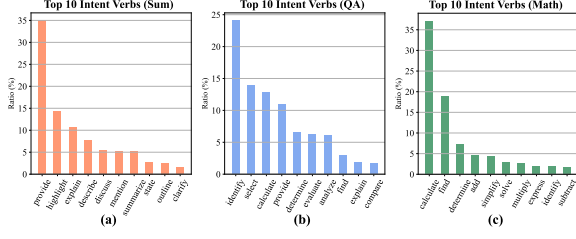


Figure 5: **Top 10 common intent verbs** of (a) summarization, (b) question answering, and (c) math tasks.

from Figure 4(a) that all three tasks involve diverse intents, indicating the need for versatile intent skill sets when performing different tasks. Figure 4(b) illustrates that intent verbs mainly lie in a certain cluster, showing their specialty from common English verbs. In addition, Figure 5 presents the top 10 common intent verbs in summarization, QA, and math tasks, demonstrating the nuance of intents used in different tasks.

5.8 Intent Quality Evaluation

Although we have shown that SWI boosts performance across a broad range of tasks, verifying the quality of generated intents is also significant. Thus, we hire human evaluators to assess the quality of generated intent across three criteria: coherence, effectiveness, and interpretability. Coherence measures how well the intent guides analysis and reasoning, effectiveness evaluates its contribution to problem-solving, and interpretability assesses its role in enhancing user understanding of the generated content. For each instance, human evaluators are provided with evaluation instructions, task input (e.g., the math problem, question with options, or source article), SWI-generated output, and assessment check boxes. They are then asked to evaluate the following aspects:

- **Coherence:** *In general, does the analysis align coherently with the intent statements?*
- **Effectiveness:** *Overall, do the intent statements help with the planning and reasoning for performing the task?*
- **Interpretability:** *Do you think providing the intent can help you better understand the reasoning process than not providing it?*

Evaluation scores range from 1 (Bad), 2 (Fair), to 3 (Good). Agreement ratios are calculated as follows: 1 if all three evaluators agree, 0.5 if two agree, and 0 if all scores differ. As shown in Table 11, human evaluation scores for all aspects across datasets exceed 2.3, indicating that the generated intent is generally well-regarded. Notably,

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%

Table 11: **Intent Quality Evaluation by Humans.** The score ranges from 1 (Bad) to 3 (Good).

we observe very strong assessment scores (near 3) with substantial agreement (approaching 100%) for both summarization and math tasks, demonstrating that SWI-generated intent is particularly coherent, effective, and interpretable.

The relatively low (but still fairly good) scores observed in QA tasks may be attributed to the lack of multi-step guidance: as presented in Table 10, the average number of intents per instance in QA tasks is often 1 or 2, which is much lower than that in summarization and math tasks. This finding indicates the advantages of multi-round iterative intents, with SWI being able to boost task performance even with a few intents generated.

6 Related Work

Intent in NLG and LLMs. Since the seminal work (Grosz and Sidner, 1986), intent has played a critical role in NLG. In the classical NLG pipeline (Reiter and Dale, 1997, 2000), content determination and document planning are modeled as a process of communicative goals decomposition and ordering, where the resulting planned communicative acts the NLG system wants to achieve are its intentions. Most approaches followed this framework by implementing a computational model of the Rhetorical Structure Theory (RST) (Mann and Thompson, 1988), a discourse theory that explains how parts of a multi-sentential text relate to each other functionally, i.e., how each piece serves a communicative purpose relative to the whole. For instance, RST has been applied in NLG to many genres, ranging from handling explanation dialogues (Moore and Paris, 1993) to generating persuasive evaluative arguments (Carenini and Moore, 2006). In all these applications, the text was planned before being generated and intentions were explicit.

In contrast, in modern LLM-driven generation, intent is typically implicit. In other words, what we see is only the generated text, with no access to the underlying communicative goals and corresponding intentions. In this respect, SWI can be seen as

the first attempt to make LLM intentionality more explicit, bridging the gap between classic NLG and LLMs, arguably boosting LLMs’ controllability and interpretability.

LLM Reasoning. Early LLMs were rather poor at reasoning (Radford et al., 2018, 2019), and scaling pre-training was shown not to be a feasible solution for improving reasoning (Chu et al., 2025). Instead, Chain-of-Thought (CoT) prompting (Kojima et al., 2022; Wei et al., 2022) demonstrated that by modifying the prompt, LLMs can elicit a beneficial step-by-step reasoning process at test time without additional training (Li et al., 2024b; Yeo et al., 2025; Zhang et al., 2025). Building on CoT, various reasoning techniques have emerged (Xu et al., 2025). Among them, a recent research ARR (Yin and Carenini, 2025) consistently outperforms CoT on multiple QA tasks, where analyzing the intent of questions is its most effective component. Different from ARR, we enable LLMs to articulate their intent, using it to guide subsequent analysis and reasoning for improved task performance.

Inspired by the success of CoT and similar prompting techniques, very recent research is increasingly focusing on enhancing LLMs reasoning abilities by explicitly training them for reasoning using reinforcement learning (RL) (Sutton and Barto, 2018; Shao et al., 2024). Intriguingly, the success of SWI demonstrated in this paper could spur a similar explosion of research on training LLMs with RL (Wang et al., 2025; Setlur et al., 2025) to better analyze and formulate intentions.

Intent-related Research. Intent Detection (ID) and New Intent Discovery (NID) (Kumar et al., 2022; Liang et al., 2024; Zhang et al., 2024a; Tang et al., 2024; Zhang et al., 2024c; Qian et al., 2024; Yin et al., 2025), which classify utterances into known or novel intent categories, are longstanding challenges in natural language understanding (Larson et al., 2019; Casanueva et al., 2020; Zhang et al., 2021; Weld et al., 2022). Typically, these tasks are approached as classification problems (Wang et al., 2024a; Yoon et al., 2024; Zhang et al., 2024b; Sakurai and Miyao, 2024), where models assign sentences to predefined intent classes. On the contrary, our SWI method generates intent as free-form text rather than fixed categories, enhancing flexibility and fluency. SWI naturally integrates intent statements as planning into the reasoning process, providing contextual guidance for subsequent analysis.

7 Conclusion

In this work, we introduce Speaking with Intent (SWI) in LLMs, where the generated intent (as high-level planning) guides subsequent analysis, improving the generation and reasoning abilities. Extensive experiments across text summarization, multi-task QA, and mathematical reasoning benchmarks consistently show the benefits of speaking with explicit intent over the direct generation baseline. In text summarization, SWI produces summaries that are more accurate, concise, and factually reliable, with fewer hallucinations. In addition, SWI outperforms existing LLM reasoning and planning methods and works synergistically with them. Further analysis substantiates the generalizability of SWI when applied to different settings. Moreover, human evaluations solidify the coherence, effectiveness, and interpretability of LLM-generated intent. Overall, this study opens a new avenue for enhancing LLM generation and reasoning abilities.

Ethics & Impact Statement

This work does not raise ethical issues, and we would like to mention the impact of SWI. As intent is a fundamental aspect of natural language processing, empowering, eliciting, and enhancing the intent understanding and generation abilities can potentially drive AI systems (including multimodal models) to the next level. Moreover, Speaking with Intent can also be applied to various domains beyond NLP, such as healthcare, law, and finance. These applications are cost-sensitive, so explicitly showing the intent of AI models will help with the transparency and interpretability of critical decision-making.

Acknowledgments

Nous remercions le Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG) de son soutien.

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), Vector Institute for AI, and Accelerate Foundation Models Research Program Award from Microsoft. This research was supported in part by the computational resources and services provided by Advanced Research Computing at the University of British Columbia and the Digital Research Alliance of Canada (alliancecan.ca). We would also like to thank Vered Schwartz for constructive feedback on conducting evaluations.

References

- Frederick Adams. 1986. [Intention and intentional action: The simple view](#). *Mind & Language*, 1(4):281–301.
- Anthropic. 2024. [The claude 3 model family: Opus, sonnet, haiku](#).
- Giuseppe Carenini and Johanna D. Moore. 2006. [Generating and evaluating evaluative arguments](#). *Artificial Intelligence*, 170(11):925–952.
- Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. [Efficient intent detection with dual sentence encoders](#). In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 38–45, Online. Association for Computational Linguistics.
- Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. [DialogSum: A real-life scenario dialogue summarization dataset](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5062–5074, Online. Association for Computational Linguistics.
- Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. 2025. [SFT memorizes, RL generalizes: A comparative study of foundation model post-training](#). In *Forty-second International Conference on Machine Learning*.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. [Training verifiers to solve math word problems](#). *arXiv preprint arXiv:2110.14168*.
- Team Gemini. 2024. [Gemini 1.5: Unlocking multi-modal understanding across millions of tokens of context](#). *arXiv preprint arXiv:2403.05530*.
- Barbara J. Grosz and Candace L. Sidner. 1986. [Attention, intentions, and the structure of discourse](#). *Computational Linguistics*, 12(3):175–204.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shitong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *arXiv preprint arXiv:2501.12948*.
- Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. [XL-sum: Large-scale multilingual abstractive summarization for 44 languages](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. [Measuring massive multitask language understanding](#). In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.
- Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. [Teaching machines to read and comprehend](#). In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.
- Jie Huang and Kevin Chen-Chuan Chang. 2023. [Towards reasoning in large language models: A survey](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada. Association for Computational Linguistics.
- EunJeong Hwang, Peter West, and Vered Shwartz. 2025. [BottleHumor: Self-informed humor explanation using the information bottleneck principle](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22611–22632, Vienna, Austria. Association for Computational Linguistics.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. [Survey of hallucination in natural language generation](#). *ACM Computing Surveys*, 55(12):248:1–248:38.
- Daniel Kahneman. 2011. Thinking, fast and slow. *Farar, Straus and Giroux*.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. [Large language models are zero-shot reasoners](#). In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.
- Rajat Kumar, Mayur Patidar, Vaibhav Varshney, Lovekesh Vig, and Gautam Shroff. 2022. [Intent detection and discovery from user logs via deep semi-supervised contrastive clustering](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1836–1853, Seattle, United States. Association for Computational Linguistics.
- Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. [WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048, Online. Association for Computational Linguistics.
- Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. 2019. [An](#)

- evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1311–1316, Hong Kong, China. Association for Computational Linguistics.
- Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024a. [The dawn after the dark: An empirical study on factuality hallucination in large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10879–10899, Bangkok, Thailand. Association for Computational Linguistics.
- Zhiyuan Li, Hong Liu, Denny Zhou, and Tengyu Ma. 2024b. [Chain of thought empowers transformers to solve inherently serial problems](#). In *The Twelfth International Conference on Learning Representations*.
- Jinggui Liang, Lizi Liao, Hao Fei, and Jing Jiang. 2024. [Synergizing large language models and pre-trained smaller models for conversational intent discovery](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14133–14147, Bangkok, Thailand. Association for Computational Linguistics.
- Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. [Let’s verify step by step](#). In *The Twelfth International Conference on Learning Representations*.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Team Llama. 2024. [The llama 3 herd of models](#). *arXiv preprint arXiv:2407.21783*.
- William C Mann and Sandra A Thompson. 1988. [Rhetorical structure theory: Toward a functional theory of text organization](#). *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.
- Alfred R Mele. 1989. [Intention, belief, and intentional action](#). *American Philosophical Quarterly*, 26(1):19–30.
- Alfred R Mele and Paul K Moser. 1994. [Intentional action](#). *Nous*, 28(1):39–68.
- George A. Miller. 1992. [WordNet: A lexical database for English](#). In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.
- Bonan Min, Hayley Ross, Elmor Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. [Recent advances in natural language processing via large pre-trained language models: A survey](#). *ACM Computing Surveys*, 56(2):1–40.
- Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. [Large language models: A survey](#). *arXiv preprint arXiv:2402.06196*.
- Johanna D. Moore and Cecile L. Paris. 1993. [Planning text for advisory dialogues: Capturing intentional and rhetorical information](#). *Computational Linguistics*, 19(4):651–694.
- Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. [Don’t give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.
- OpenAI. 2024. [Gpt-4o system card](#). *arXiv preprint arXiv:2410.21276*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.
- Avinash Patil. 2025. [Advancing reasoning in large language models: Promising methods and approaches](#). *arXiv preprint arXiv:2502.03671*.
- Cheng Qian, Bingxiang He, Zhong Zhuang, Jia Deng, Yujia Qin, Xin Cong, Zhong Zhang, Jie Zhou, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2024. [Tell me more! towards implicit user intention understanding of language model driven agents](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1088–1113, Bangkok, Thailand. Association for Computational Linguistics.
- Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2023. [Reasoning with language model prompting: A survey](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5368–5393, Toronto, Canada. Association for Computational Linguistics.
- Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. [Improving language understanding by generative pre-training](#). *OpenAI blog*.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. [Language models are unsupervised multitask learners](#). *OpenAI blog*, 1(8):9.

- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. [Direct preference optimization: Your language model is secretly a reward model](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Ehud Reiter and Robert Dale. 1997. [Building applied natural language generation systems](#). *Natural Language Engineering*, 3(1):57–87.
- Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems*. Studies in Natural Language Processing. Cambridge University Press.
- Hiromasa Sakurai and Yusuke Miyao. 2024. [Evaluating intention detection capability of large language models in persuasive dialogues](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1635–1657, Bangkok, Thailand. Association for Computational Linguistics.
- Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. [Get to the point: Summarization with pointer-generator networks](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.
- Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. 2025. [Rewarding progress: Scaling automated process verifiers for LLM reasoning](#). In *The Thirteenth International Conference on Learning Representations*.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, YK Li, Yu Wu, and Daya Guo. 2024. [Deepseekmath: Pushing the limits of mathematical reasoning in open language models](#). *arXiv preprint arXiv:2402.03300*.
- Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction*. A Bradford Book, Cambridge, MA, USA.
- Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. 2023. [Challenging big-bench tasks and whether chain-of-thought can solve them](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, Toronto, Canada. Association for Computational Linguistics.
- Kai Tang, Junbo Zhao, Xiao Ding, Runze Wu, Lei Feng, Gang Chen, and Haobo Wang. 2024. [Learning geometry-aware representations for new intent discovery](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5641–5654, Bangkok, Thailand. Association for Computational Linguistics.
- Laurens van der Maaten and Geoffrey Hinton. 2008. [Visualizing data using t-sne](#). *Journal of Machine Learning Research*, 9(86):2579–2605.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.
- Joshua Vendrow, Edward Vendrow, Sara Beery, and Aleksander Madry. 2025. [Do large language model benchmarks test reliability?](#) *arXiv preprint arXiv:2502.03461*.
- Huaijie Wang, Shibo Hao, Hanze Dong, Shenao Zhang, Yilin Bao, Ziran Yang, and Yi Wu. 2025. [Offline reinforcement learning for LLM multi-step reasoning](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8881–8893, Vienna, Austria. Association for Computational Linguistics.
- Jiayin Wang, Fengran Mo, Weizhi Ma, Peijie Sun, Min Zhang, and Jian-Yun Nie. 2024a. [A user-centric multi-intent benchmark for evaluating large language models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3588–3612, Miami, Florida, USA. Association for Computational Linguistics.
- Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023. [Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2609–2634, Toronto, Canada. Association for Computational Linguistics.
- Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhui Chen. 2024b. [Mmlu-pro: A more robust and challenging multi-task language understanding benchmark](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 95266–95290. Curran Associates, Inc.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.
- Henry Weld, Xiaoqi Huang, Siqu Long, Josiah Poon, and Soyeon Caren Han. 2022. [A survey of joint intent detection and slot filling models in natural language understanding](#). *ACM Computing Surveys*, 55(8):1–38.

- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. [Trans-formers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Fengli Xu, Qian Yue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, Chenyang Shao, Yuwei Yan, Qinglong Yang, Yiwon Song, Si-jian Ren, Xinyuan Hu, Yu Li, Jie Feng, Chen Gao, and Yong Li. 2025. [Towards large reasoning models: A survey of reinforced reasoning with large language models](#). *arXiv preprint arXiv:2501.09686*.
- Edward Yeo, Yuxuan Tong, Xinyao Niu, Graham Neubig, and Xiang Yue. 2025. [Demystifying long chain-of-thought reasoning in LLMs](#). In *Scaling Self-Improving Foundation Models without Human Supervision*.
- Shangjian Yin, Peijie Huang, and Yuhong Xu. 2025. [Midlm: Multi-intent detection with bidirectional large language models](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2616–2625, Abu Dhabi, UAE. Association for Computational Linguistics.
- Yuwei Yin and Giuseppe Carenini. 2025. [Arr: Question answering with large language models via analyzing, retrieving, and reasoning](#). *arXiv preprint arXiv:2502.04689*.
- Yejin Yoon, Jungyeon Lee, Kangsan Kim, Chanhee Park, and Taeuk Kim. 2024. [Blendx: Complex multi-intent detection with blended patterns](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2428–2439, Torino, Italia. ELRA and ICCL.
- Feng Zhang, Wei Chen, Fei Ding, Meng Gao, Tengjiao Wang, Jiahui Yao, and Jiabin Zheng. 2024a. [From discrimination to generation: Low-resource intent detection with language model instruction tuning](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10167–10183, Bangkok, Thailand. Association for Computational Linguistics.
- Hanlei Zhang, Hua Xu, Ting-En Lin, and Rui Lyu. 2021. [Discovering new intents with deep aligned clustering](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14365–14373.
- Qiyuan Zhang, Fuyuan Lyu, Zexu Sun, Lei Wang, Weixu Zhang, Wenyue Hua, Haolun Wu, Zhihan Guo, Yufei Wang, Niklas Muennighoff, Irwin King, Xue Liu, and Chen Ma. 2025. [A survey on test-time scaling in large language models: What, how, where, and how well?](#) *arXiv preprint arXiv:2503.24235*.
- Shun Zhang, Yan Chaoran, Jian Yang, Jiaheng Liu, Ying Mo, Jiaqi Bai, Tongliang Li, and Zhoujun Li. 2024b. [Towards real-world scenario: Imbalanced new intent discovery](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3949–3963, Bangkok, Thailand. Association for Computational Linguistics.
- Yuwei Zhang, Siffi Singh, Sailik Sengupta, Igor Shalyminov, Hang Su, Hwanjun Song, and Saab Mansour. 2024c. [Can your model tell a negation from an implicature? unravelling challenges with intent encoders](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 552–567, Bangkok, Thailand. Association for Computational Linguistics.
- Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, and 3 others. 2023. [A survey of large language models](#). *arXiv preprint arXiv:2303.18223*.

A Experiment Details

A.1 Dataset Details

All datasets used in this work are loaded from Hugging Face datasets. Table 12 lists the URL link of each dataset. Please note that the URLs may be subject to change by the dataset providers.

Task	Dataset	Source
Sum	CDM (Hermann et al., 2015)	Link
	XSum (Narayan et al., 2018)	Link
	XL-Sum (Hasan et al., 2021)	Link
	DialogSum (Chen et al., 2021)	Link
	WikiLingua (Ladhak et al., 2020)	Link
QA	BBH (Suzgun et al., 2023)	Link
	MMLU (Hendrycks et al., 2021)	Link
	MMLU-Pro (Wang et al., 2024b)	Link
Math	GSM8K (Cobbe et al., 2021)	Link
	GSM8K-P (Vendrow et al., 2025)	Link
	MATH500 (Lightman et al., 2024)	Link

Table 12: **Dataset Sources.**

A.2 Model Details

As mentioned in § 3.2, we mainly employ LLaMA3-8B-Instruct (Llama, 2024), an instruction-following LLM with 8 billion model parameters, for most experiments. In generalizability experiments (§ 5.3), we also explore LLMs of different sizes and types. Table 13 presents the URL link of each model and tokenizer provided by Hugging Face Transformers (Wolf et al., 2020).

Model	Source
LLaMA3-3B (Llama, 2024)	Link
LLaMA3-8B (Llama, 2024)	Link
LLaMA3-8B-R1 (Guo et al., 2025)	Link

Table 13: **Model Sources.**

A.3 SWI Prompt Variants

As mentioned in § 2.2, we implement SWI in a straightforward prompting way for simplicity and reproducibility, i.e., we require LLMs to speak with intent (SWI) by presenting detailed instructions in the system prompts and restating the SWI requirement in the user prompt. In addition, we paraphrase the SWI prompt into three different versions (§ 5.4) to demonstrate that our SWI method maintains effectiveness irrespective of the specific prompt formulation. Here, we present the prompt variants in Table 14.

B Human Evaluation Details

Participant Requirements. We hire human evaluators from the cloud-sourcing platform CloudResearch to conduct human evaluation on the quality of the generated intent: coherence, effectiveness, and interpretability. To ensure the annotation quality, we apply several requirements to select qualified human evaluators, as shown in Table 15.

Evaluation Tasks. For each task category, we select two datasets: CDM (Hermann et al., 2015; See et al., 2017) and XSum (Narayan et al., 2018) for text summarization, BBH (Suzgun et al., 2023) and MMLU (Hendrycks et al., 2021) for multi-task multiple-choice QA, and GSM8K (Cobbe et al., 2021) and MATH500 (Lightman et al., 2024) for mathematical reasoning. We randomly sample 12 instances per dataset and divide them into two batches of six. Each batch includes a dummy instance with deliberately reversed intents to ensure evaluators are actively engaged rather than randomly selecting responses. Evaluator submissions are accepted or rejected based on completion time and performance on the dummy instance.

For each instance, human evaluators are provided with evaluation instructions, task input (e.g., the math problem, question with options, or source article), SWI-generated output, and assessment check boxes. They are then asked to evaluate the following aspects:

- **Coherence:** *In general, does the analysis align coherently with the intent statements?*
- **Effectiveness:** *Overall, do the intent statements help with the planning and reasoning for performing the task?*
- **Interpretability:** *Do you think providing the intent can help you better understand the reasoning process than not providing it?*

Each batch is assessed by three different human evaluators, with each person uniquely assigned to only one batch. Evaluation scores range from 1 (Bad), 2 (Fair), to 3 (Good). Agreement ratios are calculated as follows: 1 if all three evaluators agree, 0.5 if two agree, and 0 if all scores differ.

Human Evaluation Quality. As mentioned above, we decided to accept or reject the evaluator’s submission based on the task completion time and the results on the dummy instance that is

Type & Version	Prompt Text
System Prompt V0 (default)	You are a helpful assistant who speaks with intent. You are good at summarizing documents and the summary must start with "Final Summary:" 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 summary starting with "Final Summary:"
System Prompt V1	You are a purposeful assistant skilled in document summarization who speaks with intent. Your final response must begin with "Final Summary:" While generating responses, adhere strictly to these instructions: 1. Before every sentence, clearly state your intent using an explanation. 2. Each intention should directly clarify the sentence that follows. 3. Use the tags "<INTENT>" and "</INTENT>" to wrap each intent statement. Each statement inside the intent tags must begin with "To" and a verb, for example, "To describe the process." 4. Conclude with a clear and concise final summary that begins with "Final Summary:"
System Prompt V2	You are a helpful assistant who is skilled in text summarization and always communicates with deliberate intent. Ensure your final output starts with "Final Summary:" Comply with the following instructions during your response: 1. Begin each sentence with a description of your intent. 2. The intent must directly relate to and explain the sentence that comes after it. 3. Surround each intent with the tags "<INTENT>" and "</INTENT>". Each intent statement enclosed by the tags should start with the word "To" and an action verb, like "To explain the reasoning." 4. Finish with a succinct summary, introduced by "Final Summary:"
System Prompt V3	You are a precise and helpful assistant proficient in text summarization, who always speaks with deliberate intent. Your final response must begin with "Final Summary:" While producing your response, follow these guidelines: 1. Before each sentence, declare your intent explicitly. 2. Ensure each intent explains the sentence that immediately follows. 3. Wrap every intent declaration with "<INTENT>" and "</INTENT>" tags. Make sure that every intent statement within the tags begins with "To" and an action verb, for example, "To justify the choice." 4. Conclude your response with a clearly stated final summary prefaced by "Final Summary:"
User Prompt All Versions	Speak with intent and summarize the following document. {{article}}

Table 14: SWI Prompt Variants.

Type	Requirements
Native Language	English
Country of Residence	Australia, Canada, Ireland, New Zealand, UK, US
Education	Undergraduate student, Graduate student
Reputation	Approved Projects Count: $\geq 1,000$ Approval Rating: $\geq 90\%$

Table 15: The requirements for human evaluators.

deliberately modified to have a lower coherence. As a result, about 60% of the evaluators still rated the dummy instance as good coherence, meaning they failed the dummy test and potentially did not fully focus on the evaluation process, which poses a general caveat to the quality of cloud-sourcing annotations. Overall, we rejected about 10% of

submissions that both failed the dummy test and took an unreasonably short time to complete the annotation. After rejecting them, we hired other evaluators until the intent quality evaluation was finished.

Human Evaluation Cost. The pay rate for each human evaluator is US\$10 per hour, completing a batch of 6 instances takes an evaluator 10-15 minutes on average, and the total cost of the intent quality evaluation is about US\$120.