University of Chinese Academy of Sciences

中文信息处理实验室–让机器理解语言
Chinese Information Processing Laboratory

EMNLP 2025
Suzhou, China | 中国苏州

ConsistentChat: Building Skeleton-Guided Consistent Multi-Turn Dialogues for Large Language Models from Scratch

Jiawei Chen, Xinyan Guan, Qianhao Yuan, Guozhao Mo,
Weixiang Zhou, Yaojie Lu, Hongyu Lin, Ben He, Le Sun, Xianpei Han
Chinese Information Processing Laboratory, Institute of Software,
Chinese Academy of Sciences, Beijing, China
University of Chinese Academy of Sciences, Beijing, China

Paper

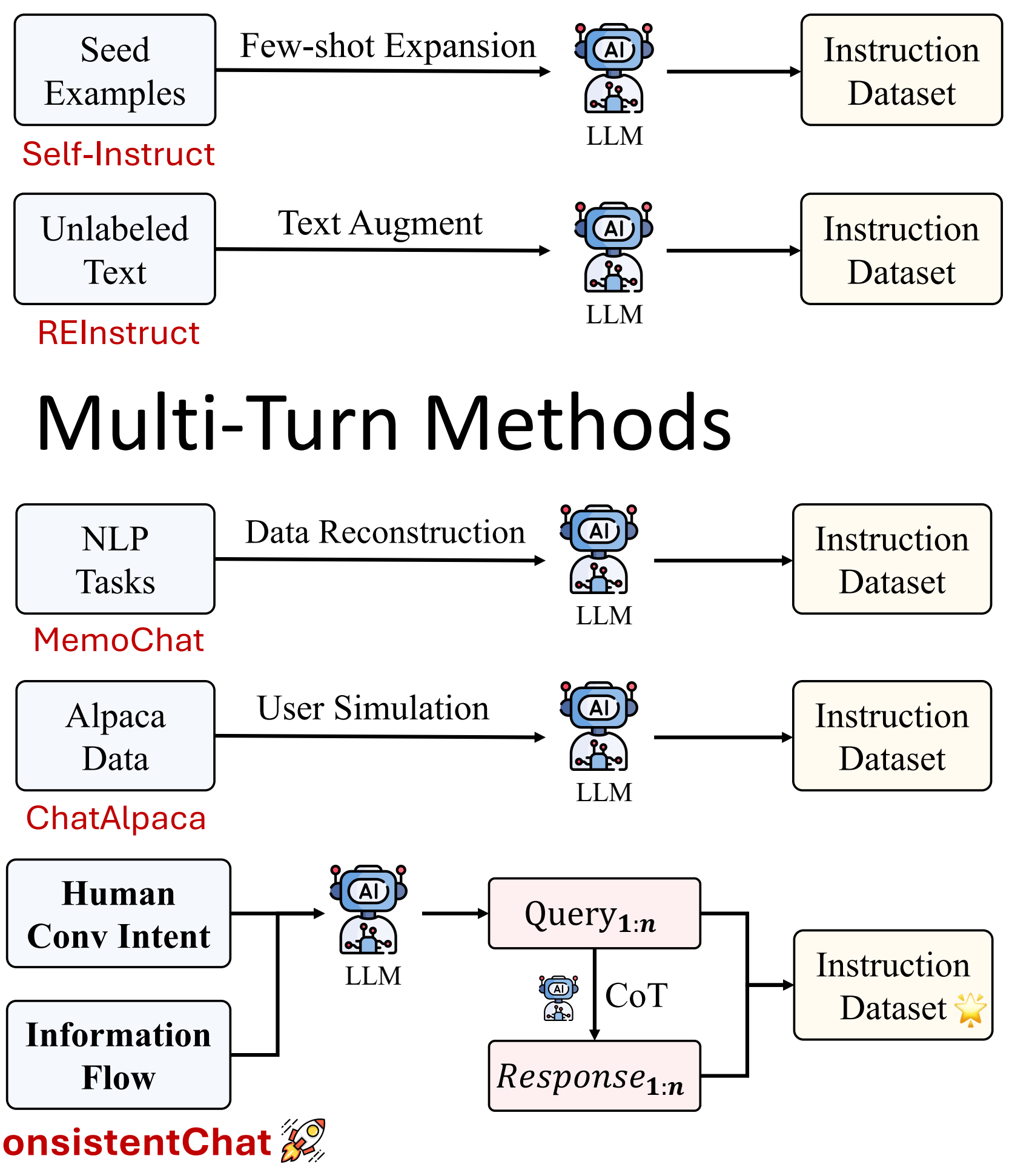
Code

Dataset

Why Consistency in Dialogue Matters?

Current Paradigms

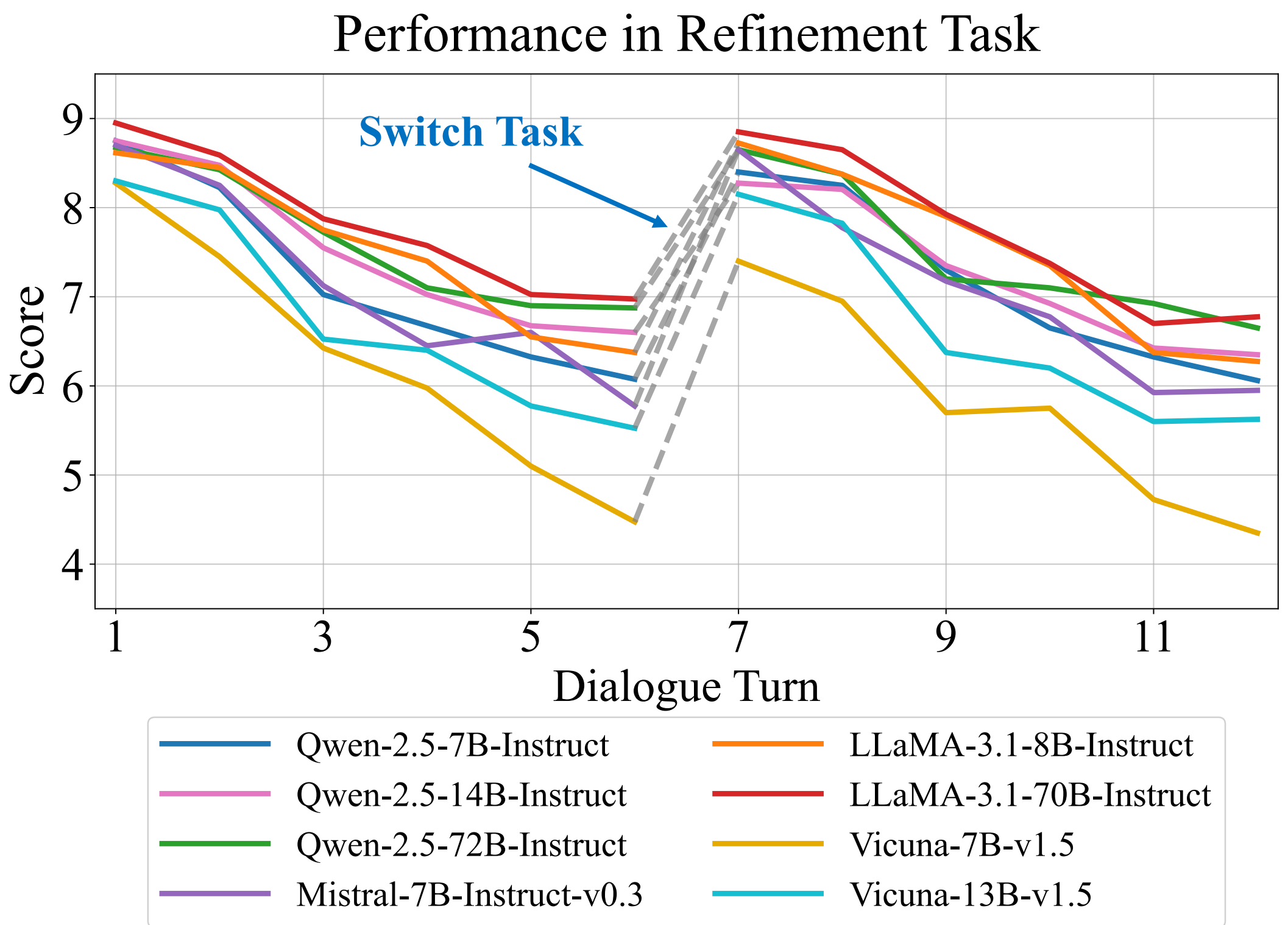
- Single-Turn Methods

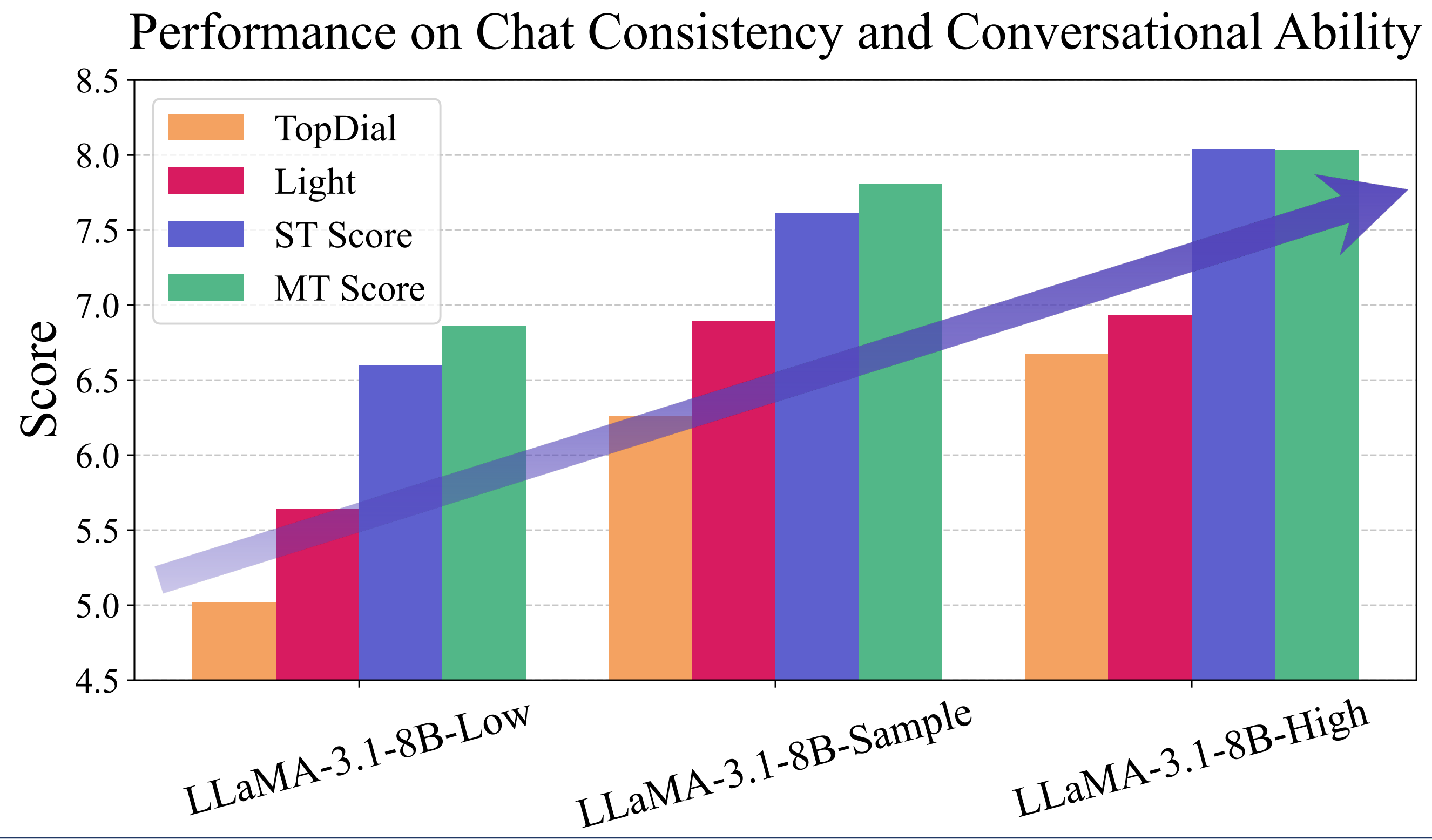


- Multi-Turn Methods

Preliminary Analyses

- Popular models exhibit degradation in conv abilities as turns increases.
- The consistency of training data affects the performance of fine-tuned models, with higher consistency data yielding superior results.





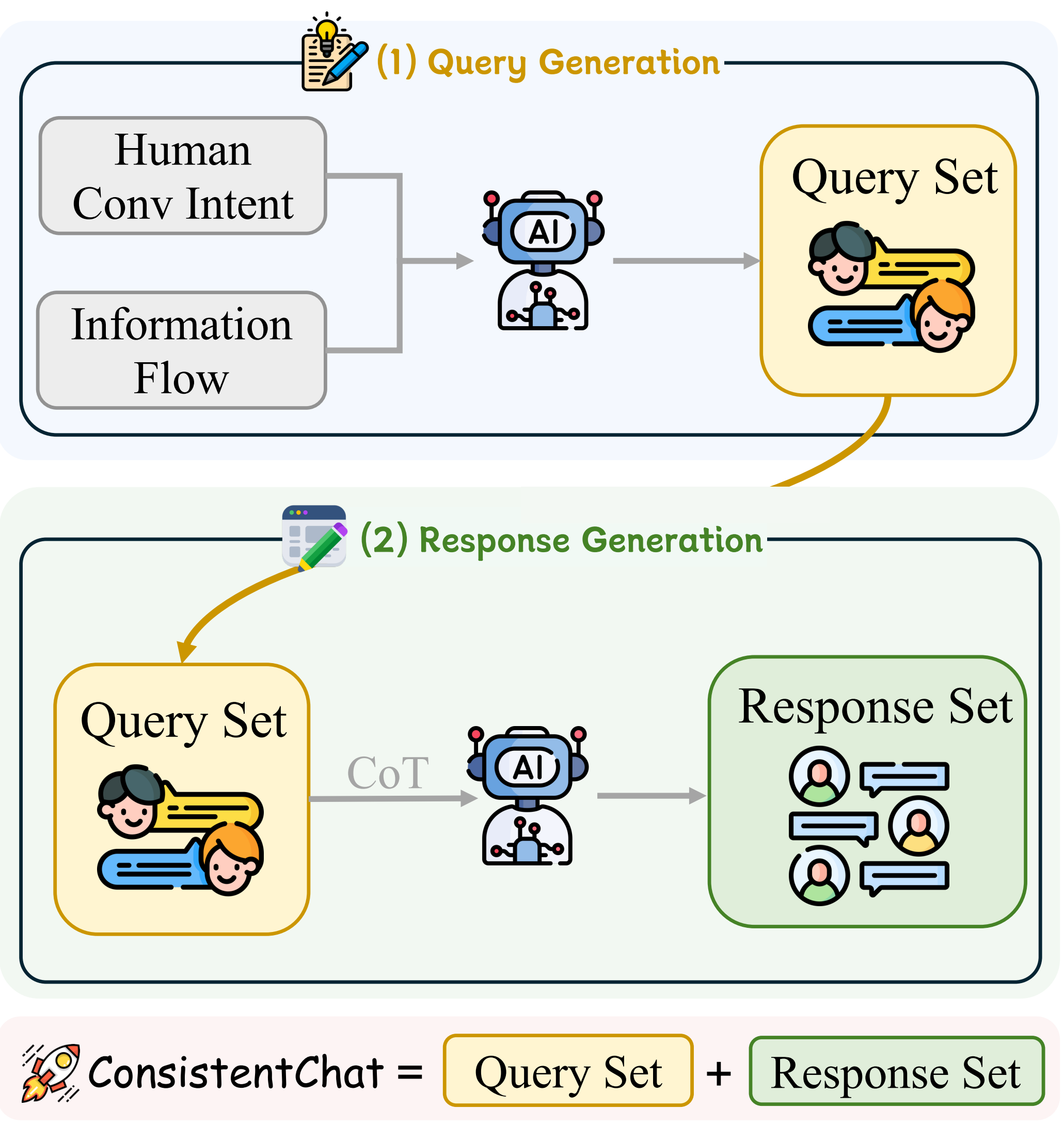
Skeleton-Guided Multi-Turn Dialogue Generation

Stage 1: Intent Modeling

- Identify *nine* types of human conv Intents.
- Each intent is formalized as an information flow schema describing how topics, user goals evolve across turns.

Stage 2: Skeleton Generation

- Generate user query skeletons aligned with intents under well-defined information flow.
- Produce all responses in a single pass with CoT prompting, improving coherence and preventing topic drift.
- This generation yields dialogues are globally consistent, context-aware, and human-like.



Consistent Multi-turn Dialogue

User (q_1): Hey, I've been feeling really anxious lately, especially when I'm alone. **Any tips on how to handle it**?
GPT (r_1): **One effective way to manage** is to create a routine includes activities you enjoy, like reading, listening to music, or walking. Would you like some more specific **techniques or tips**?
User (q_2): **I tried meditating**, but it seems a bit overwhelming. **Am I doing it wrong**?
GPT (r_2): **Meditation** can be challenging at first, and it's okay to find it overwhelming.… Be patient and not to **judge your progress harshly**.
User (q_3): Sometimes I just **can't shake off** the feeling of worry. It's always there in the back of my mind. **How can I manage this**?
GPT (r_3): **One effective strategy** is to practice mindfulness, which involves focusing on the present moment without judgment. This can help you gain perspective and reduce the intensity of the worry.

……

User (q_n): I'm thinking about trying **therapy** for help. How do I find the right therapist, and what should I expect from the sessions? Thank you for your kind advice!
GPT (r_n): Start by looking for **therapists** who specialize in anxiety disorders and have experience with the techniques you're interested in, such as cognitive-behavioral therapy…

Overall, the dialogue is **highly consistent, context-aware**, and **emotionally supportive**.

Experiments & Results

Consistency Improvement

- Models fine-tuned on ConsistentChat achieve a *20–30%* increase in consistency and significantly reduce topic drift across turns on the LIGHT and TOPDIAL.

Multi-Turn Conversational Capability

- ConsistentChat models show up to a *15%* higher task success rate and stronger resilience as conversations deepen, even surpassing larger models (Qwen-2.5-14B-Instruct) on MT-EVAL benchmark.

Models	LIGHT		TOPDIAL		Avg.
	QWEN Score	LLAMA Score	QWEN Score	LLAMA Score	
Qwen-2.5-72B-Instruct	7.48	7.92	7.87	8.05	7.83
Qwen-2.5-7B	6.36	5.69	6.98	6.42	6.36
Qwen-2.5-7B-ShareGPT	6.71	<u>7.32</u>	7.03	<u>7.33</u>	<u>7.10</u>
Qwen-2.5-7B-ChatAlpaca	6.11	6.97	6.70	6.87	6.66
Qwen-2.5-7B-UltraChat	<u>6.78</u>	7.23	<u>7.14</u>	6.90	7.01
Qwen-2.5-7B-LmsysChat	6.00	6.07	6.44	5.83	6.09
Qwen-2.5-7B-*ConsistentChat*	**6.94**	**7.50**	**7.34**	**7.51**	**7.32**
LLaMA-3.1-70B-Instruct	7.44	7.86	7.57	7.62	7.62
LLaMA-3.1-8B	4.55	3.76	5.83	5.34	4.87
LLaMA-3.1-8B-ShareGPT	<u>6.42</u>	<u>6.66</u>	6.62	6.39	6.52
LLaMA-3.1-8B-ChatAlpaca	6.38	6.56	6.85	6.77	6.64
LLaMA-3.1-8B-UltraChat	6.15	6.55	<u>7.14</u>	<u>6.84</u>	<u>6.67</u>
LLaMA-3.1-8B-LmsysChat	5.66	5.43	6.24	4.59	5.48
LLaMA-3.1-8B-*ConsistentChat*	**6.71**	**6.72**	**7.22**	**7.06**	**6.93**
Mistral-7B-v0.3	3.09	2.49	4.09	4.00	3.42
Mistral-7B-v0.3-ShareGPT	<u>6.33</u>	**6.71**	6.71	5.61	<u>6.34</u>
Mistral-7B-v0.3-ChatAlpaca	5.65	6.18	6.22	5.20	5.81
Mistral-7B-v0.3-UltraChat	5.49	6.08	<u>6.83</u>	<u>6.36</u>	6.19
Mistral-7B-v0.3-LmsysChat	5.08	5.52	6.01	5.37	5.50
Mistral-7B-v0.3-*ConsistentChat*	**6.62**	<u>6.21</u>	**7.09**	**6.67**	**6.65**

(a) Consistency result

Models	ST Score	MT Score
Qwen-2.5-14B-Instruct	8.01	7.95 (-0.06)
Qwen-2.5-7B	5.66	5.83 (+0.17)
Qwen-2.5-7B-ShareGPT	7.81	7.86 (+0.05)
Qwen-2.5-7B-ChatAlpaca	<u>7.86</u>	<u>8.12</u> (+0.26)
Qwen-2.5-7B-UltraChat	6.18	6.65 (+0.47)
Qwen-2.5-7B-LmsysChat	5.61	5.74 (+0.13)
Qwen-2.5-7B-*ConsistentChat*	**8.07**	**8.38** (+0.31)
LLaMA-3.1-8B	4.86	4.38 (–0.48)
LLaMA-3.1-8B-ShareGPT	<u>7.40</u>	7.60 (+0.20)
LLaMA-3.1-8B-ChatAlpaca	7.37	<u>7.73</u> (+0.36)
LLaMA-3.1-8B-UltraChat	6.89	6.85 (–0.04)
LLaMA-3.1-8B-LmsysChat	5.66	5.78 (+0.12)
LLaMA-3.1-8B-*ConsistentChat*	**7.71**	**7.93** (+0.22)
Mistral-7B-v0.3	4.41	5.71 (+1.30)
Mistral-7B-v0.3-ShareGPT	6.39	<u>6.94</u> (+0.55)
Mistral-7B-v0.3-ChatAlpaca	<u>6.47</u>	6.68 (+0.21)
Mistral-7B-v0.3-UltraChat	5.97	6.23 (+0.26)
Mistral-7B-v0.3-LmsysChat	5.48	5.06 (–0.42)
Mistral-7B-v0.3-*ConsistentChat*	**6.67**	**7.14** (+0.47)

(b) Multi-Turn conv ability result

Conclusions

- We propose ConsistentChat, generated by a simple yet effective Skeleton-Guided framework for supervised fine-tuning, which can be applied in broad downstream dialogue scenarios.
- Extensive experiments show that ConsistentChat outperforms existing popular multi-turn datasets in terms of chat consistency, as well as both single-turn and multi-turn conversational ability.