中国科学院大学 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

ISCAS

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

Jiawei Chen

2025.11.07

EMNLP 2025 Suzhou, China 中国苏州

Preliminary Analysis

What happens when chatbots talk longer?


How do I improve my sleep quality?


Try maintaining a regular schedule... ✅


What about diet?

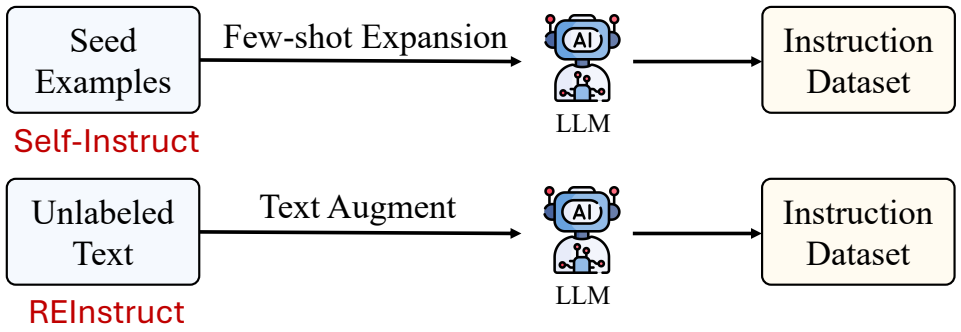

Mediterranean diet is healthy for overall wellness... ✅


Any specific foods?


Speaking of nutrition, let me tell you about supplements... ❌

The conversation *drifts away* from the original sleep topic.

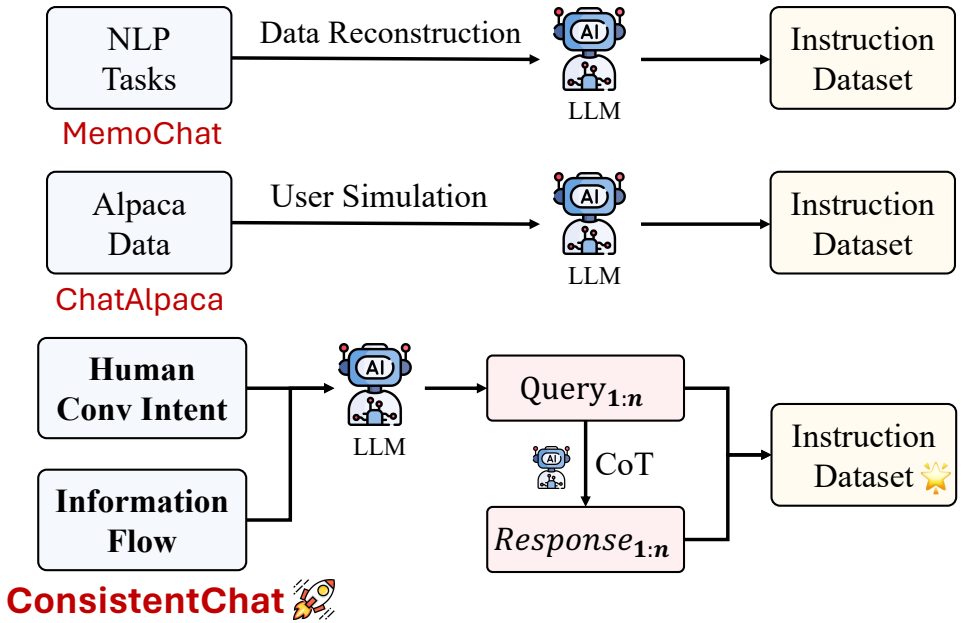
Why Does This Happen ?



(a) **Single-Turn** Instruction Synthesis Manners.

Single-Turn Focus:

- Self-Instruct, Evolve-Instruct → <instruction, response> pairs
- Missing: Sequential dependencies & conversational flow



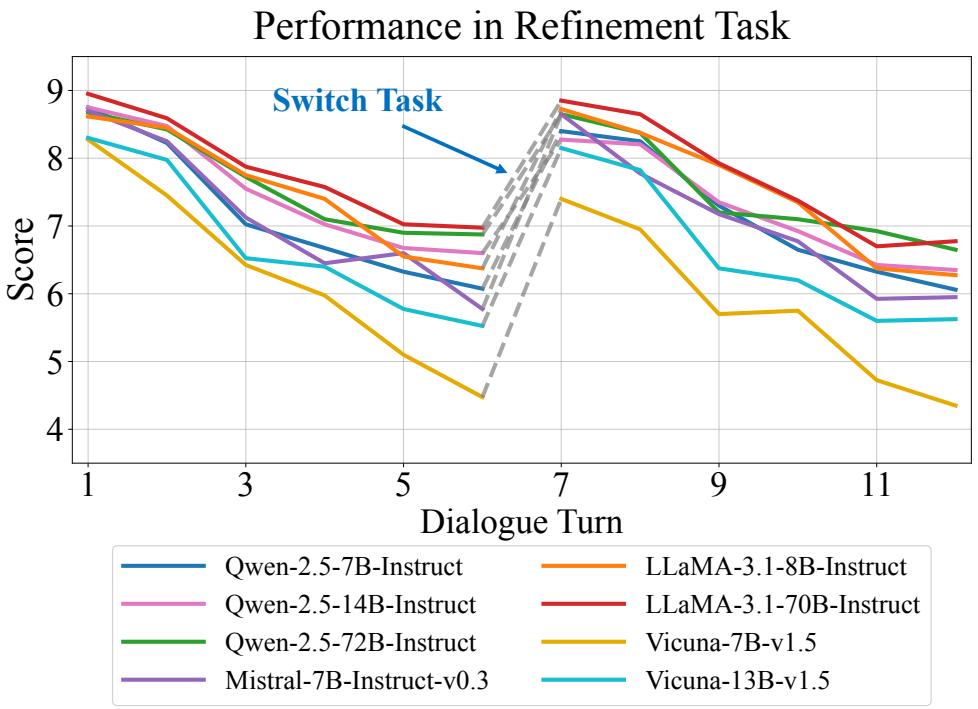
(b) **Multi-Turn** Instruction Synthesis Manners.

Multi-Turn Attempts:

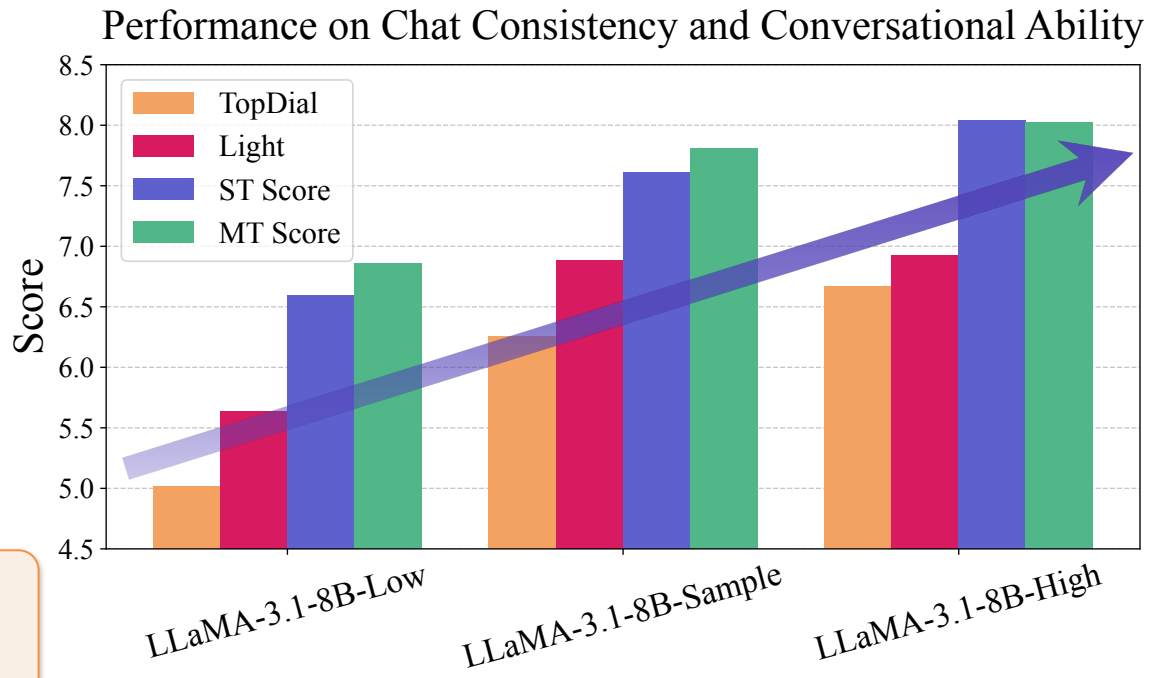
- MemoChat, ChatAlpaca, GLAN
- Problem: Unconstrained generation
- Each turn is locally coherent BUT...
- No global intent modeling → Topic drift

Why Consistency Matters ?

1. Popular models exhibit degradation in conv abilities as turns increases.



2. The consistency of training data affects the performance of fine-tuned models, with higher consistency data yielding superior results.



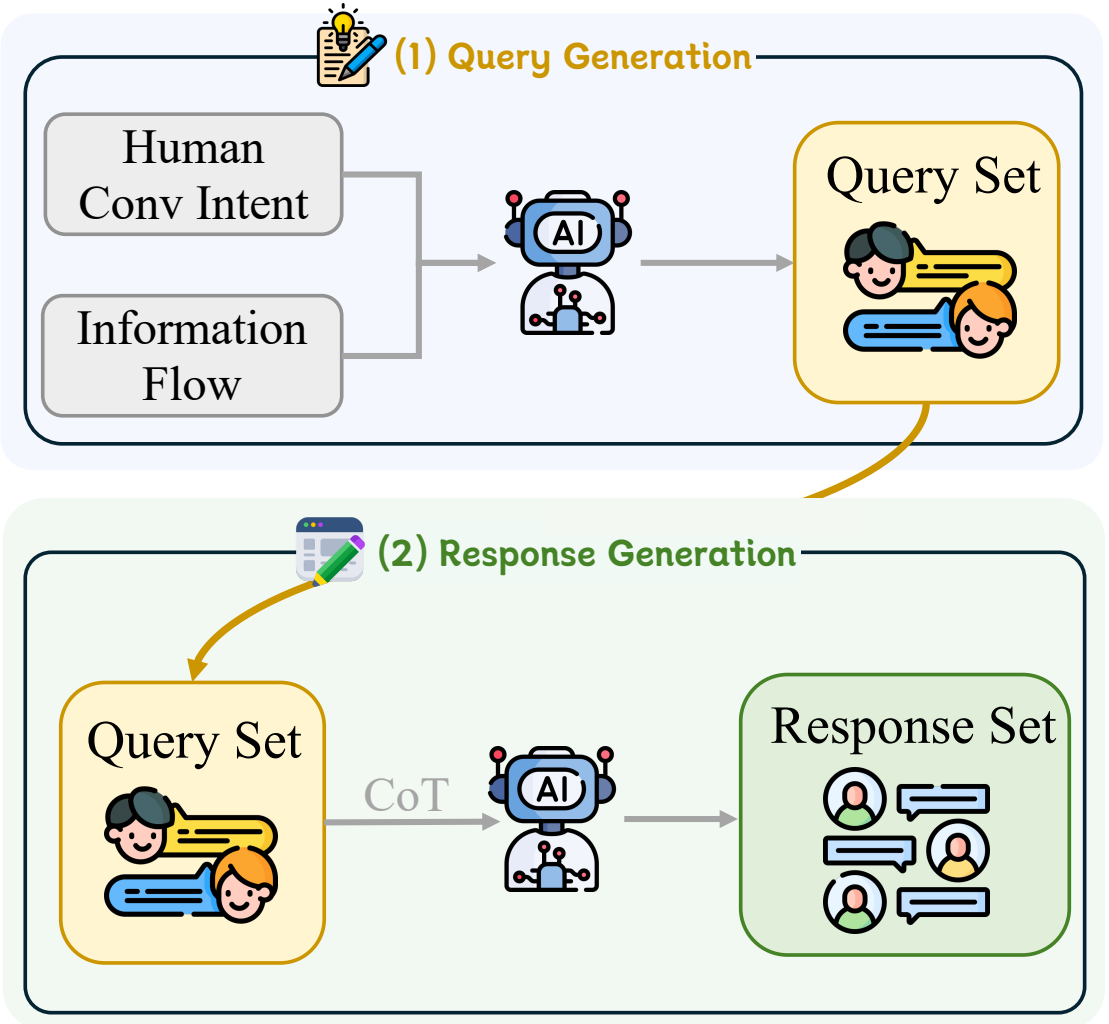
Consistency:

(1) Alignment with the initial conversational intent,

(2) Smooth and logically connected information flow,

(3) Contextual relevance of each response across turns.

How It Works ?



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**.

Chat Consistency Results

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	7.32	7.03	7.33	7.10
Qwen-2.5-7B-ChatAlpaca	6.11	6.97	6.70	6.87	6.66
Qwen-2.5-7B-UltraChat	6.78	7.23	7.14	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	6.42	6.66	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	7.14	6.84	6.67
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	6.33	**6.71**	6.71	5.61	6.34
Mistral-7B-v0.3-ChatAlpaca	5.65	6.18	6.22	5.20	5.81
Mistral-7B-v0.3-UltraChat	5.49	6.08	6.83	6.36	6.19
Mistral-7B-v0.3-LmsysChat	5.08	5.52	6.01	5.37	5.50
Mistral-7B-v0.3-*ConsistentChat*	**6.62**	6.21	**7.09**	**6.67**	**6.65**

Table 1: The overall experimental results of *ConsistentChat* and other baselines on the LIGHT and TOPDIAL benchmarks, based on chat consistency metric. QWEN Score and LLAMA Score indicate the results obtained from *Qwen-2.5-72B-Instruct* and *LLaMA-3.1-70B-Instruct*, respectively. The best/second scores are **bolded**/underlined.

Settings

Fine-tuned models:

- Qwen-2.5-7B
- LLaMA-3.1-8B
- Mistral-7B-v0.3

We use different popular multi-turn dialogue datasets, with ~15,000 conversations for each model.

Result

Fine-tuning with ConsistentChat achieves the highest average scores on chat consistency metrics.

Multi-Turn Capability Results

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	7.86	8.12 (+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	7.40	7.60 (+0.20)
LLaMA-3.1-8B-ChatAlpaca	7.37	7.73 (+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	6.94 (+0.55)
Mistral-7B-v0.3-ChatAlpaca	6.47	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)

Result

Models trained with ConsistentChat also demonstrate improved multi-turn conversational capabilities. These models outperform all the models trained on baseline datasets in both single-turn and multi-turn conditions, **even outperforming Qwen-2.5-14B-Instruct.**

Positive Correlation

Enhancing consistency substantially improves overall multi-turn competence, underscoring its importance for building robust conversational AI systems.

Beyond Consistency

➢ Commonsense Reasoning

Models	Reasoning			Examination		Knowledge	Code	Avg.
	Hellaswag	MATH	GPQA	MMLU	Gaokao	TriviaQA	HumanEval	
Qwen-2.5-7B	80.20	49.80	**36.40**	**74.20**	61.05	64.93	57.90	60.64
Qwen-2.5-7B-ShareGPT	79.96	52.42	29.80	70.42	71.96	60.79	70.73	62.30
Qwen-2.5-7B-ChatAlpaca	80.40	36.68	32.32	73.66	74.13	66.49	67.68	61.62
Qwen-2.5-7B-UltraChat	70.56	46.26	32.83	71.36	71.88	60.94	58.54	58.91
Qwen-2.5-7B-LmsysChat	48.19	14.40	30.81	59.48	30.66	46.29	32.32	37.44
Qwen-2.5-7B-*ConsistentChat*	**83.09**	**64.96**	35.35	74.02	**75.54**	**67.31**	**77.44**	**68.24**

Table 3: Results of fine-tuning Qwen models on commonsense reasoning tasks. Qwen-2.5-7B-*ConsistentChat* exhibits stronger general capabilities, with notable improvements in MATH(↑ **30.4%**) and HumanEval(↑ **33.7%**) compared to the pre-trained model, indicating an average gain of **12.5%** across all commonsense benchmarks.

Beyond Consistency

Human Evaluation

Does LLM evaluation match human judgment?

Task	Pearson	Spearman
LIGHT	0.73	0.70
TOPDIAL	0.69	0.66
Avg.	0.71	0.68

✅ Reliable evaluation

Data Quality

Semantic consistency within dialogues

Dataset	DistilBERT	MPNet	MiniLM
ShareGPT	0.892	0.895	0.857
ChatAlpaca	0.851	0.874	0.862
UltraChat	0.907	0.911	0.915
LmsysChat	0.817	0.847	0.813
ConsistentChat	**0.915**	**0.919**	**0.923**

✅ Superior Semantic Consistency

Impact & Takeaways

ConsistentChat Dataset:

- ✓ 15,000 Conversations, 224K Utterances
- ✓ 9 Intent categories, publicly available

Key Takeaways:

- ✓ Modeling human conversational intent is crucial
- ✓ Skeleton-Guided generation improves consistency
- ✓ We need to **take *consistency* into account as well**

Thanks for Your Attention!

Contact me at jiawei.ucas@gmail.com

https://github.com/chenjiawei30/ConsistentChat