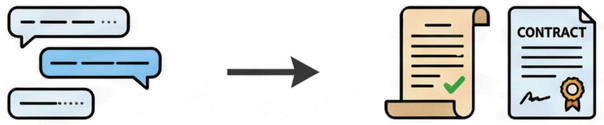
DiscoX: Benchmarking Discourse-Level Translation Tasks in Expert Domains

ByteDance | Seed Xiying Zhao*, Zhoufutu Wen*, Zhixuan Chen, Jingzhe Ding, Shuai Li, Xi Li, Shengda Long

WHY WE DID IT & CORE CONTRIBUTIONS

THE GAP: Is Translation Solved? 🤔 NO.

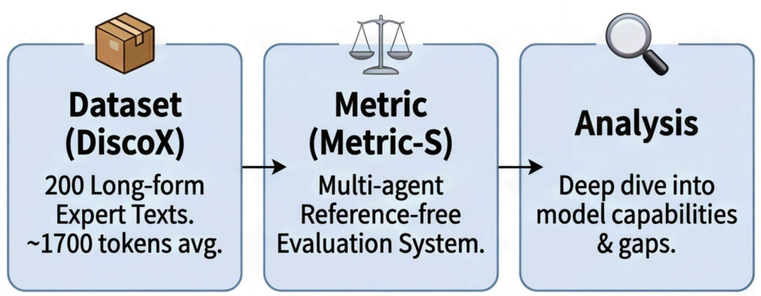


Sentence-Level
(e.g., FLORES)
✅ Near Perfect Scores

Discourse-Level & Expert
(e.g., Papers, Industrial)
❌ Not Delivery-Ready. Fails on Coherence & Terminology.

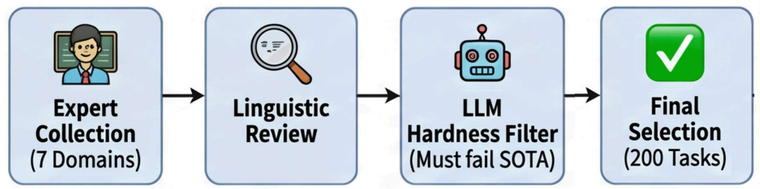
Current models struggle with long-context coherence and specialized domains.

CORE CONTRIBUTIONS 🚀



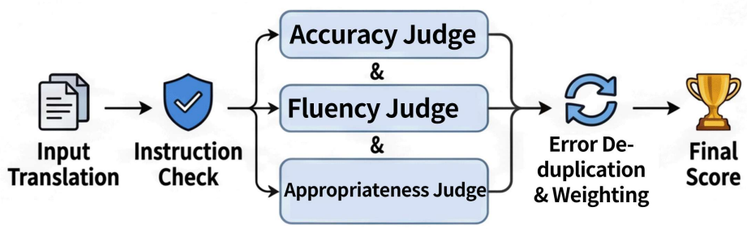
METHODOLOGY: DATASET &METRIC

DISCO-X DATASET 📚



Dataset: 7 Domains (Academic & Non-Academic), ~1.7k tokens/text. High Difficulty.

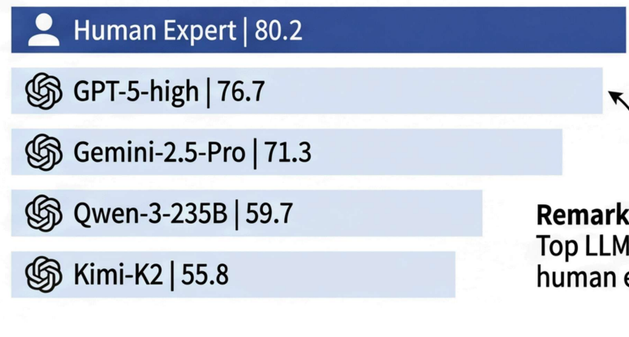
METRIC-S EVALUATION SYSTEM ⚙️



Effectiveness: High human alignment (70.3% consistency vs XCOMET-QE 34.7%).

RESULTS & KEY OBSERVATIONS

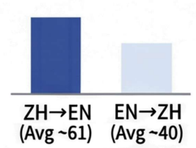
LEADERBOARD (Overall Scores) 🏆



Remarkable Gap: Top LLMs still trail human experts.

KEY OBSERVATIONS 🧠

🔁 Language Asymmetry



Models are significantly better at translating into English.

🏛️ Domain Differences

Stronger in structured Academic Papers; Weaker in literary/creative texts.

📉 Dimensional Imbalance

Models miss different dimensions (e.g., good Accuracy but poor Fluency).