LREC-COLING 2024

LatEval: An Interactive LLMs Evaluation Benchmark with Incomplete Information from Lateral Thinking Puzzles

Shulin Huang*, Shirong Ma*, Yinghui Li*, Mengzuo Huang, Wuhe Zou,

Weidong Zhang and Hai-Tao Zheng

sl-huang21@mails.tsinghua.edu

Shenzhen International Graduate School, Tsinghua University.

Netease Games AI Lab, HangZhou.

Peng Cheng Laboratory, Shenzhen.

Exist Evaluation Methods for LLMs

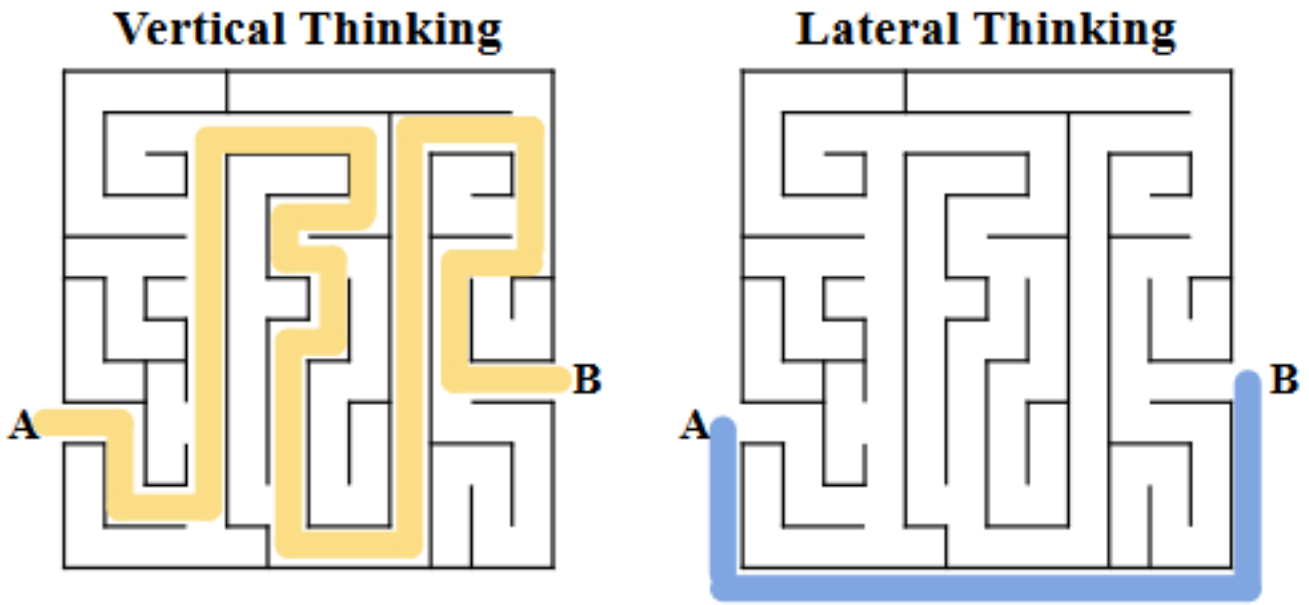

Figure 1: Vertical Thinking V.S. Lateral Thinking

- **MMLU**[1], **C-Eval**[2], **GSM8K**[3] → mathematics, science, law, and general problems → **Vertical Thinking**
- **Vertical Thinking**: typically refers to thinking within established or conventional thought patterns, following the known rules.
- **Lateral Thinking**: also known as **divergent thinking**, involves breaking out of traditional thought patterns and employing innovative approaches to explore non-conventional solutions.

Study Motivation

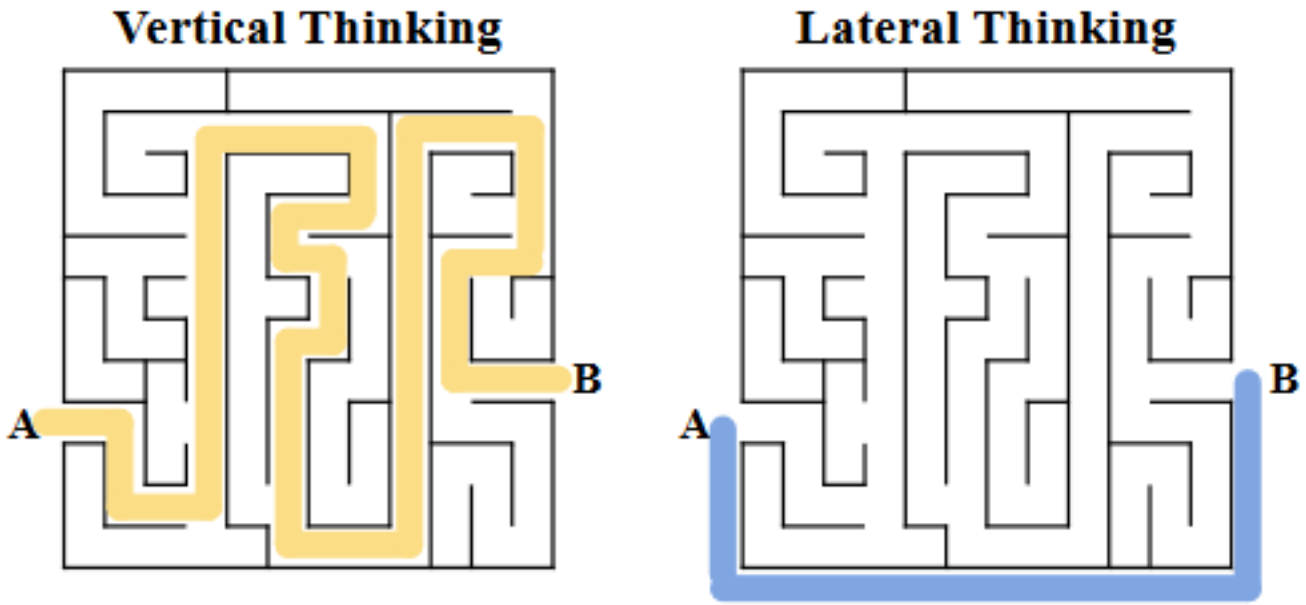

Figure 1: Vertical Thinking V.S. Lateral Thinking

- **Lateral Thinking**
  - Is widely recognized in **Psychology** and **Education**.
  - Encourages individuals to construct creative ideas [4-6].

Lateral Thinking Puzzles

- **Rules:**
  - With one host and multiple players.
  - **Player:** Ask the host a series of “yes” or “no” questions to gather information.
  - **Host:** Use “yes”, “no” or “irrelevant” to response the question.
- Involves unexpected twists, hidden clues, or counter-intuitive solutions.
- Requires players to think beyond traditional patterns and consider unconventional possibilities.

Our Contributions

- We propose a novel interactive benchmark to evaluate the lateral thinking abilities of LLMs.
- We construct an evaluation dataset both in Chinese and English.
- Experimental results indicate that most LLMs perform poorly.
- Our benchmark is challenging and has the capacity to distinguish LLMs' performance.

Dataset Construction

	English	Chinese
# Entries	225	100
Puzzles' Average Length	29.4	42.0
Truths' Average Length	51.4	66.6
Average Key Clues	3.7	3.8

Table 1: Statistics of LatEval dataset

- Collect 2000 samples.
- Remove duplicate entries and irrational entries.
- Exclude excessively negative entries.
- Select some positive and heartwarming entries.
- Perform a model awareness test to filter the data.
- Label the key clues of each entry.

Dataset Introduction

- Chinese and English data capture the cultural characteristics of each language.

- **Annotation** :
  - 12% is easily associated with the truth.
  - 88% is relatively challenging.
  - underscores the high quality of LatEval dataset.

LatEval

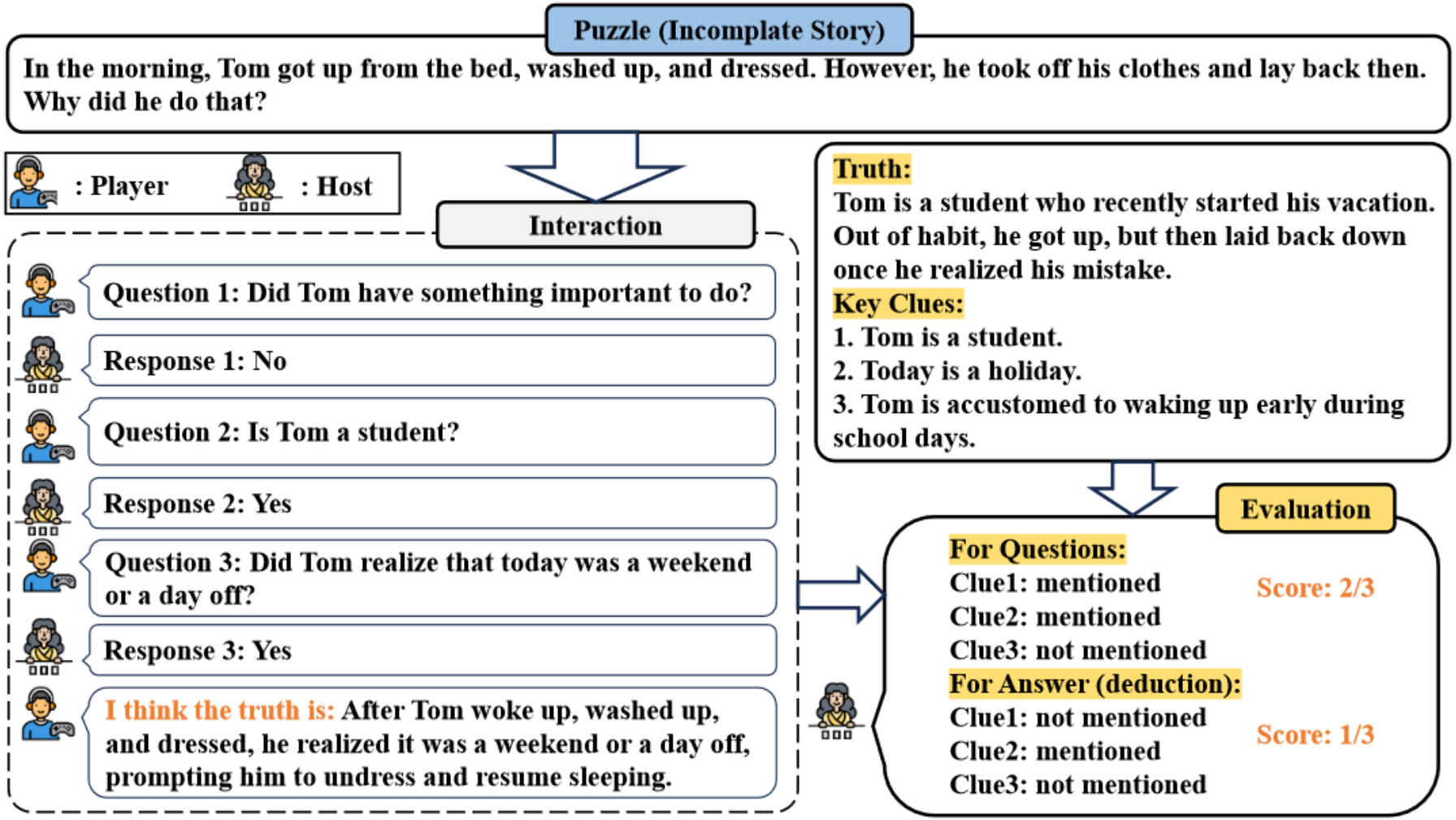

Figure 2: An example of a Lateral Thinking Puzzle in LatEval benchmark

Evaluation Metrics

Expectation:

(1) Posing high-quality questions that break out of conventional norms but are beneficial for puzzle-solving.

(2) Incorporating available information to progressively arrive at the truth through reasoning.

- **Answer Consistency (AC)**

$$\text{AC} = \frac{1}{|c|} \sum_{i=1}^{|c|} I(c_i, a),$$

- **Question Relevance (QR)**

$$\text{QR} = \frac{1}{|c|} \sum_{i=1}^{|c|} \max_{j=1}^{|q|} I(c_i, q_j),$$

- **Question Divergence (QD)**

$$\text{QD} = 1 - \frac{2}{|q| \cdot (|q| - 1)} \sum_{i=1}^{|q|} \sum_{j=i+1}^{|q|} \text{sim}(q_i, q_j),$$

↓

Jaccard Index

- **Average Turns (AT)**

Lateral Thinking Performance

	Players	English				Chinese			
		AC	QR	QD	AT	AC	QR	QD	AT
Closed-source Models	GPT-4	**34.2**	**58.4**	79.8	10.2	**36.6**	**83.8**	**77.1**	12.4
	GPT-3.5	17.1	53.3	77.2	13.0	9.5	67.6	76.7	10.3
	Claude	14.5	49.6	**81.1**	8.3	16.1	54.2	68.6	7.4
Open-source Models	Llama2-chat-7B*	11.7	30.8	74.2	15.2	8.6	66.6	62.3	15.7
	Llama2-chat-13B*	6.7	**42.5**	78.5	13.0	9.1	56.2	57.5	12.5
	Llama2-chat-70B*	11.9	38.4	73.0	14.3	10.4	**72.6**	61.2	15.2
	Baichuan-chat-13B	7.4	22.9	**91.4**	8.5	4.5	12.0	53.0	9.5
	Baichuan2-chat-13B	9.7	36.0	82.5	9.6	4.8	39.8	**72.7**	8.7
	InternLM-chat-20B	**13.7**	32.0	61.1	4.7	**11.0**	23.8	34.7	5.0
	ChatGLM2-6B	1.3	13.6	40.9	19.1	3.2	24.5	55.2	11.6
	Bloomchat-176B	1.8	7.9	45.7	3.4	0.0	5.0	0.0	4.1

Table 2: Lateral thinking performance of various LLMs with GPT-4 as the host

The results suggest that the majority of existing models still lack lateral thinking and are unable to perfectly complete this task.

Lateral Thinking Performance

	Players	English				Chinese			
		AC	QR	QD	AT	AC	QR	QD	AT
Closed-source Models	GPT-4	**27.5**	**72.7**	**79.6**	11.7	**25.7**	**70.4**	**77.9**	12.8
	GPT-3.5	23.9	59.3	78.3	13.2	12.6	61.2	68.1	12.6
	Claude	19.6	50.2	77.9	8.4	19.6	52.2	64.7	7.2
Open-source Models	Llama2-chat-7B*	5.7	43.2	69.6	9.5	5.7	28.0	44.2	7.9
	Llama2-chat-13B*	6.4	34.7	58.7	4.3	3.6	20.1	18.4	2.4
	Llama2-chat-70B*	13.4	**52.2**	75.5	11.0	9.9	**46.1**	61.7	8.4
	Baichuan-chat-13B	6.3	20.7	**82.3**	5.4	1.5	5.5	33.4	7.0
	Baichuan2-chat-13B	11.8	45.0	77.0	9.3	10.8	42.1	**62.0**	9.4
	InternLM-chat-20B	**13.8**	28.1	54.8	5.2	**13.3**	27.4	34.7	16.3
	ChatGLM2-6B	1.0	3.1	55.8	12.1	5.1	28.4	41.6	8.8
	Bloomchat-176B	1.5	3.4	39.9	3.8	0.7	3.5	13.3	6.4

Table 3: Lateral thinking performance of various LLMs with GPT-3.5 as the host

With GPT-4 and GPT-3.5 as hosts, the overall trends are generally similar.

Human evaluation

Models	Answer Consistency		Question Relevance		Host Correction Accuracy	
	Score	Correlation	Score	Correlation	GPT-4	GPT-3.5
GPT-4	43.9	0.84†	67.0	0.80†	98.4	89.8
GPT-3.5	9.7	0.71†	46.3	0.77†	96.8	94.5
Baichuan2-chat-13B	7.0	0.69†	24.3	0.81†	94.6	86.8
Baichuan-chat-13B	5.7	0.69†	15.3	0.87†	90.4	84.5
Llama2-chat-70B	12.0	0.71†	45.5	0.81†	94.7	89.0
Bloomchat-176B	0.0	-	5.4	0.76†	90.1	81.0

Table 4: Human Evaluation Results

- Spearman Correlation Coefficient
- † represents $p < 0.01$

- Annotation:
  - (1) The relevance of the player model's questions to the truth's key clues.
  - (2) The consistency between the player model's answer and truth's key clues.
  - (3) The accuracy of the host model's responses to the player model's questions.

Human evaluation

Models	Answer Consistency		Question Relevance		Host Correction Accuracy	
	Score	Correlation	Score	Correlation	GPT-4	GPT-3.5
GPT-4	43.9	0.84†	67.0	0.80†	98.4	89.8
GPT-3.5	9.7	0.71†	46.3	0.77†	96.8	94.5
Baichuan2-chat-13B	7.0	0.69†	24.3	0.81†	94.6	86.8
Baichuan-chat-13B	5.7	0.69†	15.3	0.87†	90.4	84.5
Llama2-chat-70B	12.0	0.71†	45.5	0.81†	94.7	89.0
Bloomchat-176B	0.0	-	5.4	0.76†	90.1	81.0

Table 4: Human Evaluation Results

- The results demonstrates a strong correlation between the human evaluation and the GPT-4 automated evaluation.
- GPT-4 achieves an accuracy rate of over 90% as a host responding to questions.
- GPT-3.5 achieves an accuracy rate of over 80% as a host responding to questions

Fine-grained Analysis

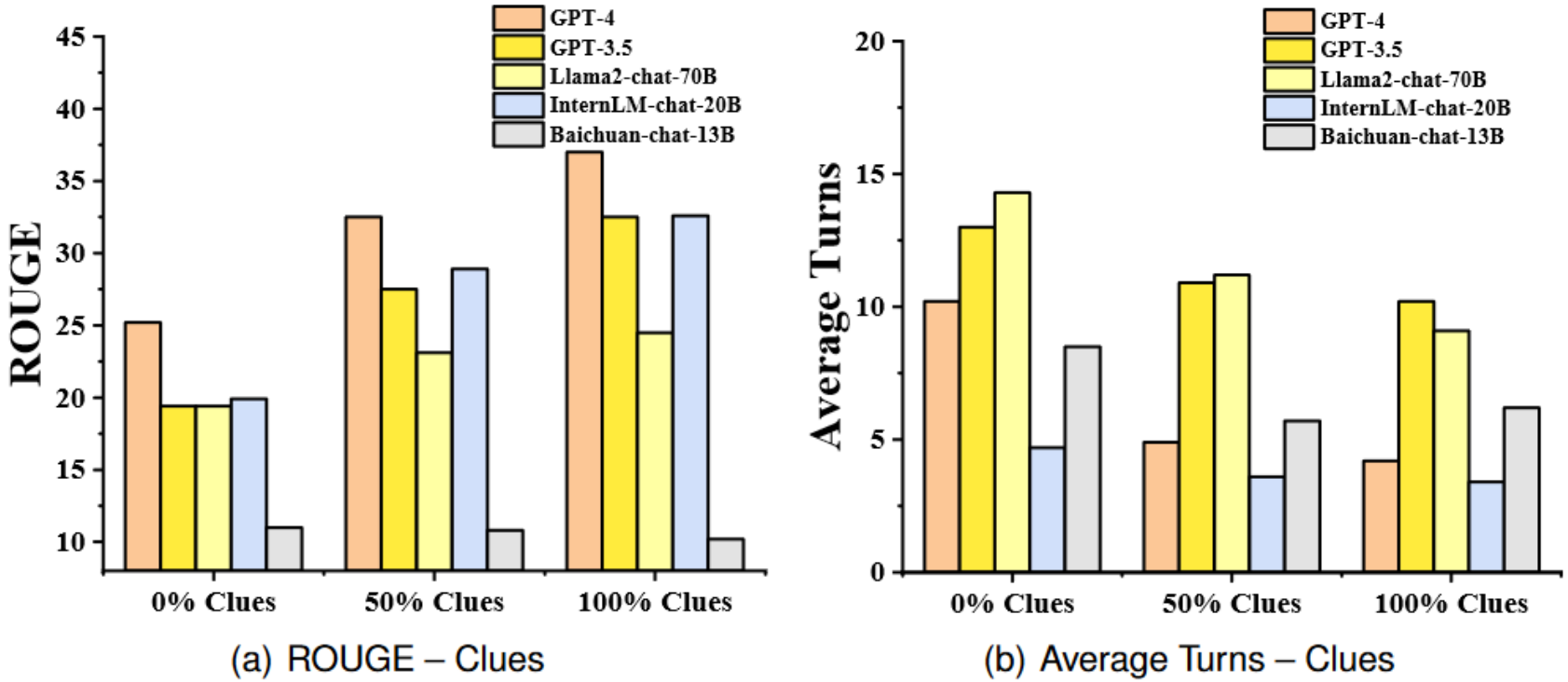

Figure 3: Lateral thinking performance of various LLMs under various difficulty setting

- As the amount of provided clues increases, the deduction becomes closer to the truth, and models that are skilled at integrating information exhibit the decrease in Average Turns.
- The performance under 100% clues setting represents the upper limit of each model's performance in Lateral Thinking Puzzle.

References

[1] Hendrycks D, Burns C, Basart S, et al. Measuring Massive Multitask Language Understanding[C]//International Conference on Learning Representations. 2020.
[2] Huang Y, Bai Y, Zhu Z, et al. C-Eval: A Multi-Level Multi-Discipline Chinese Evaluation Suite for Foundation Models[J]. arXiv e-prints, 2023: arXiv: 2305.08322.
[3] Cobbe K, Kosaraju V, Bavarian M, et al. Training verifiers to solve math word problems[J]. arXiv preprint arXiv:2110.14168, 2021.
[4] Hendrycks D, Burns C, Basart S, et al. Measuring Massive Multitask Language Understanding[C]//International Conference on Learning Representations. 2020.
[5] Huang Y, Bai Y, Zhu Z, et al. C-Eval: A Multi-Level Multi-Discipline Chinese Evaluation Suite for Foundation Models[J]. arXiv e-prints, 2023: arXiv: 2305.08322.
[6] Cobbe K, Kosaraju V, Bavarian M, et al. Training verifiers to solve math word problems[J]. arXiv preprint arXiv:2110.14168, 2021.

LREC-COLING 2024

Thanks for watching !

LatEval: An Interactive LLMs Evaluation Benchmark with Incomplete Information from Lateral Thinking Puzzles

Shulin Huang*, Shirong Ma*, Yinghui Li*, Mengzuo Huang, Wuhe Zou,

Weidong Zhang and Hai-Tao Zheng

sl-huang21@mails.tsinghua.edu