Workshop on the Scaling Behavior of Large Language Models

Antonio Valerio Miceli-Barone
amiceli@ed.ac.uk

Elena Voita
lenavoita@meta.com

Fazl Barez
fazl@robots.ox.ac.uk

Ulrich Germann
ugermann@ed.ac.uk

Shay B. Cohen
scohen@inf.ed.ac.uk

Michal Lukasik
mlukasik@google.com

Why?

- LLMs tend to become better with increasing size

Why?

- LLMs tend to become better with increasing size (Kaplan's power law scaling, "Chinchilla" scaling, Emergent abilities, Sutton's Bitter Lesson)

Why?

- LLMs tend to become better with increasing size (Kaplan's power law scaling, "Chinchilla" scaling, Emergent abilities, Sutton's Bitter Lesson)
- But not for all tasks and scenarios

Why?

- LLMs tend to become better with increasing size (Kaplan's power law scaling, "Chinchilla" scaling, Emergent abilities, Sutton's Bitter Lesson)
- But not for all tasks and scenarios: **Inverse Scaling**
  - Performance decreases as model size increases

Why?

- LLMs tend to become better with increasing size (Kaplan's power law scaling, "Chinchilla" scaling, Emergent abilities, Sutton's Bitter Lesson)
- But not for all tasks and scenarios: **Inverse Scaling**
  - Performance decreases as model size increases
    - Social biases
    - Unwanted memorization
    - Incorrect reasoning on OOD Python code
    - Compositional generalization failures
    - ...and so on

Why?

- More non-monotonic scaling trends discovered

Why?

- More non-monotonic scaling trends discovered (U-shaped, inverse U-shaped)

Why?

- More non-monotonic scaling trends discovered (U-shaped, inverse U-shaped)
- Inverse Scaling Prize
  - Shared task with strict submission format and automatic evaluation
  - Discovered many interesting inverse and non-monotonic scaling tasks
  - Ian McKenzie will tell you all about it

Why?

- More non-monotonic scaling trends discovered (U-shaped, inverse U-shaped)
- Inverse Scaling Prize
  - Shared task with strict submission format and automatic evaluation
  - Discovered many interesting inverse and non-monotonic scaling tasks
  - Ian McKenzie will tell you all about it
- Going beyond the Inverse Scaling Prize
  - Submissions are academic papers

Why?

- More non-monotonic scaling trends discovered (U-shaped, inverse U-shaped)
- Inverse Scaling Prize
  - Shared task with strict submission format and automatic evaluation
  - Discovered many interesting inverse and non-monotonic scaling tasks
  - Ian McKenzie will tell you all about it
- Going beyond the Inverse Scaling Prize
  - Submissions are academic papers
  - Scale parameter: not just model size
    - E.g. number of languages, number of domains
  - Performance measure: not just accuracy
    - E.g. calibration, uncertainty, internal characteristics
  - Prompting strategy: not just direct zero-shot
    - E.g. number of in-context examples, number of chain-of-thought "reasoning" steps

Submissions

- 9 submissions

Submissions

- 9 submissions
- 14 reviewers

Submissions

- 9 submissions
- 14 reviewers
- 4 accepted

Submissions

- 9 submissions
- 14 reviewers
- 4 accepted+ a Findings of EACL paper

Submissions

- 9 submissions
- 14 reviewers
- 4 accepted+ a Findings of EACL paper
- 1 best paper

Keynote talks

Ian McKenzie

Keynote talks

Ian McKenzie

Lead organizer of the Inverse Scaling Prize and first author of the associated paper, currently he is a contracting Research Engineer on OpenAI's Dangerous Capability Evaluations project.

Keynote talks

Ian McKenzie

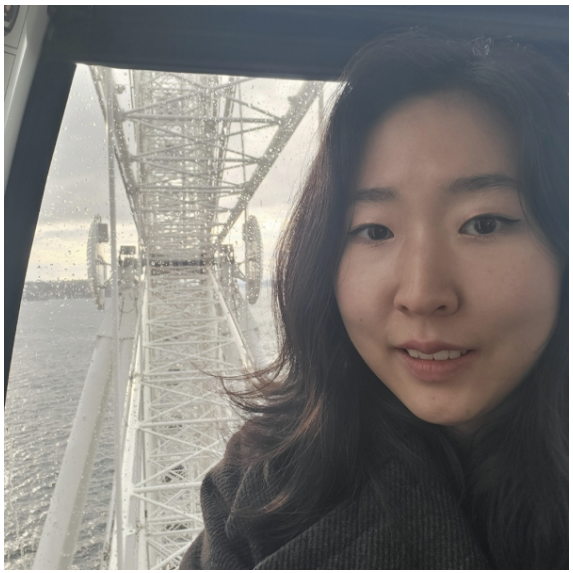
Lead organizer of the Inverse Scaling Prize and first author of the associated paper, currently he is a contracting Research Engineer on OpenAI's Dangerous Capability Evaluations project.

Inverse Scaling: When Bigger isn't Better

Keynote talks

Najoung Kim

Keynote talks



Najoung Kim

Assistant Professor at Boston University and a researcher at Google. She is also one of the authors of the Inverse Scaling Prize paper as well as other foundational works in this field.

Keynote talks

Najoung Kim

Assistant Professor at Boston University and a researcher at Google. She is also one of the authors of the Inverse Scaling Prize paper as well as other foundational works in this field.

Inverse scaling: mitigation strategies and open questions

Platinum sponsor

Platinum sponsor

- Best Paper Award
- Student financial support

Silver sponsor

Organizer personal sponsors

UKRI Research Node on Trustworthy Autonomous Systems Governance and Regulation
for Antonio Valerio Miceli-Barone

Apart Research
for Fazl Barez

Schedule

- 09:00 - 09:15 Opening Remarks
- 09:15 - 09:45 Invited Talk 1 - Ian McKenzie
- 09:45 - 10:30 Oral presentations
- 10:30 - 14:00 Break
- 14:00 - 14:30 Invited talk 2 - Najoung Kim
- 14:30 - 15:15 Panel discussion
- 15:15 - 15:30 Best paper announcement and closing remarks
- 15:30 - 17:30 Poster session

Schedule

- 09:00 - 09:15 Opening Remarks
- 09:15 - 09:45 Invited Talk 1 - Ian McKenzie
- 09:45 - 10:30 Oral presentations
- 10:30 - 14:00 Break
- 14:00 - 14:30 Invited talk 2 - Najoung Kim
- 14:30 - 15:15 Panel discussion
- 15:15 - 15:30 Best paper announcement and closing remarks
- 15:30 - 17:30 Poster session

`https://scale-llm-24.pages.dev/`

Underline

Best paper award

Best paper award

InstructEval: Towards Holistic Evaluation of Instruction-Tuned Large Language Models

Yew Ken Chia, Pengfei Hong, Lidong Bing, Soujanya Poria

Best paper award

InstructEval: Towards Holistic Evaluation of Instruction-Tuned Large Language Models

Yew Ken Chia, Pengfei Hong, Lidong Bing, Soujanya Poria

"... our evaluation involves a rigorous assessment of models based on problem-solving, writing ability, and alignment to human values. We take a holistic approach to analyze various factors affecting model performance, including the pretraining foundation, instruction-tuning data, and training methods. Our findings reveal that the quality of instruction data is a crucial factor in scaling model performance. "

Best paper award

InstructEval: Towards Holistic Evaluation of Instruction-Tuned Large Language Models

Yew Ken Chia, Pengfei Hong, Lidong Bing, Soujanya Poria

"... our evaluation involves a rigorous assessment of models based on problem-solving, writing ability, and alignment to human values. We take a holistic approach to analyze various factors affecting model performance, including the pretraining foundation, instruction-tuning data, and training methods. Our findings reveal that the quality of instruction data is a crucial factor in scaling model performance. "

Best paper award is sponsored by

Google

Thanks

- Thanks to our authors

Thanks

- Thanks to our authors, our reviewers

Thanks

- Thanks to our authors, our reviewers
- Thanks to our keynote speakers

Thanks

- Thanks to our authors, our reviewers
- Thanks to our keynote speakers
- Thanks to our panelists

Thanks

- Thanks to our authors, our reviewers
- Thanks to our keynote speakers
- Thanks to our panelists
- Thanks to our sponsors

Thank you!

Antonio Valerio Miceli-Barone

Fazl Barez

Shay B. Cohen

Elena Voita

Ulrich Germann

Michal Lukasik

Workshop on the Scaling Behavior of Large Language Models

Poster session

`https://scale-llm-24.pages.dev/`

Underline