SCALE-LLM 2024

# First edition of the Workshop on the Scaling Behavior of Large Language Models (SCALE-LLM 2024)

## Proceedings of the Workshop

March 22, 2024

The SCALE-LLM organizers gratefully acknowledge the support from the following sponsors.

**Platinum**

Google

**Silver**

Meta

**Organizers' personal sponsors**

UKRI Research Node on TAS Governance & Regulation

Apart

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 979-8-89176-077-6

# Introduction

We are excited to welcome you to SCALE-LLM 2024, the First Workshop on the Scaling Behavior of Large Language Models. SCALE-LLM 2024 is being held in Malta on 22 March 2024, co-located with EACL 2024.

The purpose of this workshop is to provide a venue to share and discuss results of investigations into the scaling behavior of Large Language Models (LLMs). We are particularly interested in results displaying interestingscaling curves (e.g., inverse, u-shaped, or inverse u-shaped scaling curves) for a variety of tasks. These results, where the performance of the LLMs decreases with increasing model size or follows a non-monotonic trend, deviating from the expected the bigger, the betterpositive scaling laws, are of great scientific interest as they can reveal intrinsic limitations of current LLM architectures and training paradigms and they provide novel research directions towards a better understanding of these models and of possible approaches to improve them.

Recently, there has been an increasing interest in these phenomena from the research community, culminating in the Inverse Scaling Prize, which solicited tasks to be systematically evaluated according to a standardized protocol in order to perform a systematic study. The SCALE-LLM Workshop will expand these efforts. In contrast to the Inverse Scaling Prize, which focused on zero-shot tasks with a fixed format, we are also interested in, for example, few-shot and alternate prompting strategies (e.g. Chain-of-Thoughts), multi-step interactions (e.g. Tree-of-Thoughts, self-critique), hardening against prompt injection attacks (e.g. user input escaping, canary tokens), etc.

The program includes two keynote talks, three oral presentations, a discussion panel and a poster session.

We extend special thanks to our Program Committee members, our Keynote speakers Najoung Kim and Ian McKenzie, the EACL workshop chairs Nafise Moosavi and Zeerak Talat, the publication chairs Danilo Croce and Goezde Guel Sahin and all the EACL organizers.

We thank our Platinum sponsor Google Research, our Silver Sponsor Meta and our organizers personal sponsors UKRI Research Node on Trustworthy Autonomous Systems Governance and Regulation (for Antonio Valerio Miceli-Barone) and Apart Research (for Fazl Barez). Thanks to the generosity of our Platinum sponsor Google Research we are able to provide a monetary prize for our best paper award and provide financial aid to student presenters.

The SCALE-LLM 2024 organizers

Antonio Valerio Miceli-Barone, Fazl Barez, Shay B. Cohen, Elena Voita, Ulrich Germann, Michal Lukasik

# Organizing Committee

**Workshop Organizers**

Antonio Valerio Miceli-Barone, University of Edinburgh
Fazl Barez, University of Oxford
Shay B. Cohen, University of Edinburgh
Elena Voita, Meta
Ulrich Germann, University of Edinburgh
Michal Lukasik, Google Research

# Program Committee

**Reviewers**

Ernest Davis

Marcio Fonseca

Ulrich Germann, Adam Grycner, Sarang Gupta

Barry Haddow, Jacob Hilton

Najoung Kim

Shihao Liang

Ian R. McKenzie, Antonio Valerio Miceli-Barone

Alicia Parrish

Parth Sarthi

Lucas Torroba Hennigen

Yftah Ziser

## Keynote Talk

# Inverse Scaling: When Bigger isn't Better

**Ian McKenzie**
OpenAI (contractor)
**2024-03-22 09:45:00** – Room: **Fortress 1**

**Abstract:** Work on scaling laws has found that large language models (LMs) show predictable improvements to overall loss with increased scale (model size, training data, and compute). I'll discuss the phenomenon of inverse scaling": that LMs may show worse task performance with increased scale, e.g., due to flaws in the training objective and data. We collected empirical evidence of inverse scaling on 11 datasets collected by running a public contest, the Inverse Scaling Prize. Through analysis of the datasets, along with other examples found in the literature, we identified four potential causes of inverse scaling: (i) preference to repeat memorized sequences over following in-context instructions, (ii) imitation of undesirable patterns in the training data, (iii) tasks containing an easy distractor task which LMs could focus on, rather than the harder real task, and (iv) correct but misleading few-shot demonstrations of the task. Our tasks have helped drive the discovery of U-shaped and inverted-U scaling trends, where an initial trend reverses, suggesting that scaling trends are not always monotonic and that existing scaling laws less reliable at predicting the behavior of larger-scale models than previously understood. Our results suggest that there are tasks for which increased model scale alone may not lead to improved performance, and that more careful thought needs to go into the data and objectives for training language models.

**Bio:** Ian McKenzie is the main 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 Talk

**Najoung Kim**
Boston University / Google
**2024-03-22 14:30:00** – Room: **Fortress 1**

**Abstract:** to be decided

**Bio:** Najoun Kim is an 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.

# Table of Contents

# Program

**Friday, March 22, 2024**

09:00 - 09:15 *Opening Remarks*

09:15 - 09:45 *Invited Talk 1 - Ian McKenzie*

09:45 - 10:30 *Oral presentations*

*Scaling Behavior of Machine Translation with Large Language Models under Prompt Injection Attacks*
Zhifan Sun and Antonio Valerio Miceli-Barone

*InstructEval: Towards Holistic Evaluation of Instruction-Tuned Large Language Models*
Yew Ken Chia, Pengfei Hong, Lidong Bing and Soujanya Poria

*When do Generative Query and Document Expansions Fail? A Comprehensive Study Across Methods, Retrievers, and Datasets*
Orion Weller, Kyle Lo, David Wadden, Dawn J Lawrie, Benjamin Van Durme, Arman Cohan and Luca Soldaini

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*

*A Proposal for Scaling the Scaling Laws*
Wout Schellaert, Ronan Hamon, Fernando Martínez-Plumed and Jose Hernandez-Orallo

*Scaling Behavior of Machine Translation with Large Language Models under Prompt Injection Attacks*
Zhifan Sun and Antonio Valerio Miceli-Barone

*Can Large Language Models Reason About Goal-Oriented Tasks?*
Filippos Bellos, Yayuan Li, Wuao Liu and Jason J Corso

*InstructEval: Towards Holistic Evaluation of Instruction-Tuned Large Language Models*
Yew Ken Chia, Pengfei Hong, Lidong Bing and Soujanya Poria

*Detecting Mode Collapse in Language Models via Narration*
Sil Hamilton

*When do Generative Query and Document Expansions Fail? A Comprehensive Study Across Methods, Retrievers, and Datasets*
Orion Weller, Kyle Lo, David Wadden, Dawn J Lawrie, Benjamin Van Durme, Arman Cohan and Luca Soldaini