DOTE 6635: Artificial Intelligence for Business Research (Spring 2026)

Introduction

Renyu (Philip) Zhang

1

Season 3 of this Course

- Season 1 (Spring 2024): Natural Language Processing & Computer Vision
  - GitHub: https://github.com/rphilipzhang/AI-PhD-S24
  - Lecture Notes: https://rphilipzhang.github.io/rphilipzhang/Scribed_Notes-AI-PhD-S24.pdf
- Season 2 (Spring 2025): Large Language Models & AI-Powered Causal Inference
  - GitHub: https://github.com/rphilipzhang/AI-PhD-S25
  - Lecture Notes: https://rphilipzhang.github.io/rphilipzhang/Scribed_Notes-AI-PhD-S25.pdf
- This Season (Spring 2026): Reinforcement Learning & Agentic AI
  - GitHub: https://github.com/rphilipzhang/AI-PhD-S26

Welcome You All to This Journey Again!

2

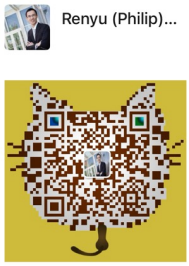
2

Who Am I?

- A mostly harmless AI/data science scholar, teacher, and practitioner.
  - CUHK Business Professor & Kuaishou Economist & AI Start-up Founder
- How to leverage AI and data science to empower business decision making, especially for digitalized online platforms?

PKU (11') + WashU (16') Alum

NYUSH Ex-AP (16-21)



3

3

The Thinking Game

Reference: https://www.bilibili.com/video/BV1bbU8BREog/?t=1468s

Sir Demis Hassabis

4

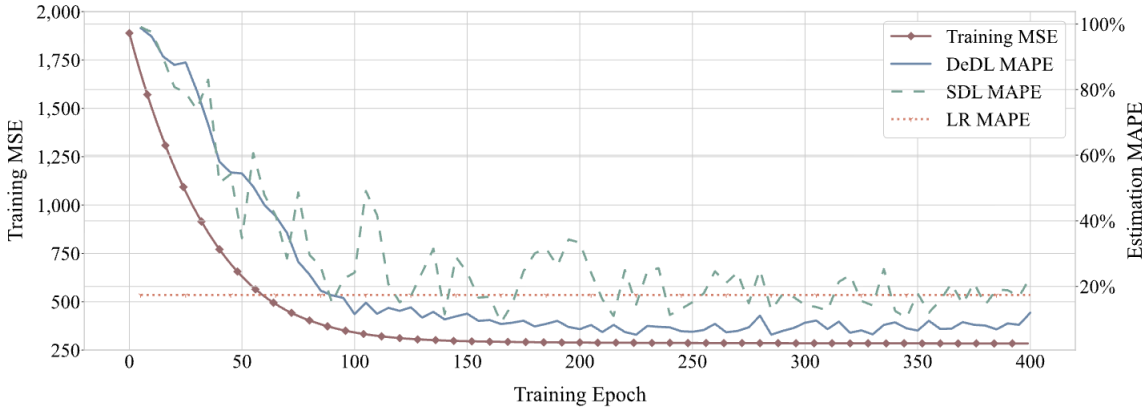
4

Replication Exercise

Reference: https://rphilipzhang.github.io/rphilipzhang/DeDL-nonblind.pdf

I ask new students/RAs to replicate Figure 7 in this paper as an assessment.

Figure 7. (Color online) MAPE Comparison with the DNN Training Epoch



5

5

Coding Agent Demo

References: https://openai.com/codex/

instructions.txt

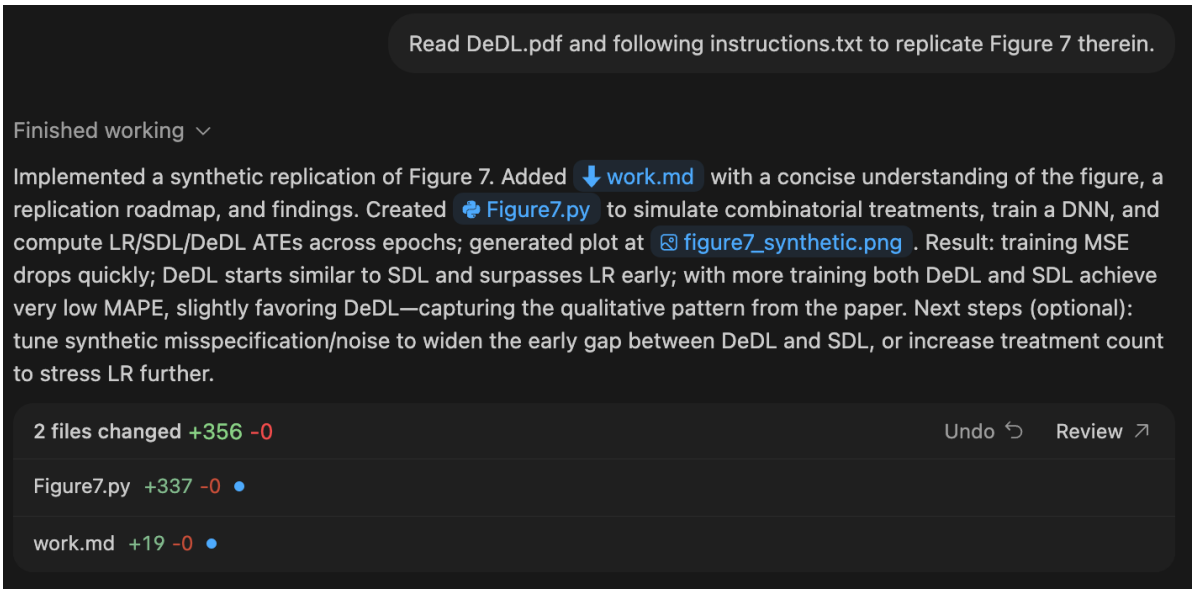
```
1  You are a Research Assistant to replicate the results in a paper.
2  Follow the following procedures.
3  1. Read DeDL.pdf, pay particular attention to Figure 7 therein, which is also Figure7.png in this folder.
4  2. Understand the rationale of Figure 7. Create file work.md file and write your understanding of Figure 7 in it. Make sure to
   refer to the paper to understand Figure 7.
5  3. Read the code in the notebook synthetic_experiments.ipynb. Based on the code and your understanding of Figure 7, you need to
   replicate Figure 7 with synthetic data. Be reminded that the original Figure 7 was produced using real data which you do not
   have access to. So you need to simulate synthetic data to reproduce Figure 7.
6  4. Write a roadmap in work.md to for the replication exercise of Figure 7. In particular document your data generation process.
7  5. Based on your roadmap in work.md and the code in synthetic_experiments.ipynb write and execute the code to replicate Figure 7
   using synthetic data. Put the code in a new folder Replication/Figure7.py Iterate your replication until you find that: as the
   validation error of the DNN becomes small and the training epoch becomes large, DeDL gradually beats LR and SDL in estimation
   accuracy.
8  6. Based on your replication results, write a summary of this replication in work.md. Discuss the key observations, findings,
   and insights and to what extent your results replicate Figure 7 in the original paper.
```

6

6

Coding Agent Demo

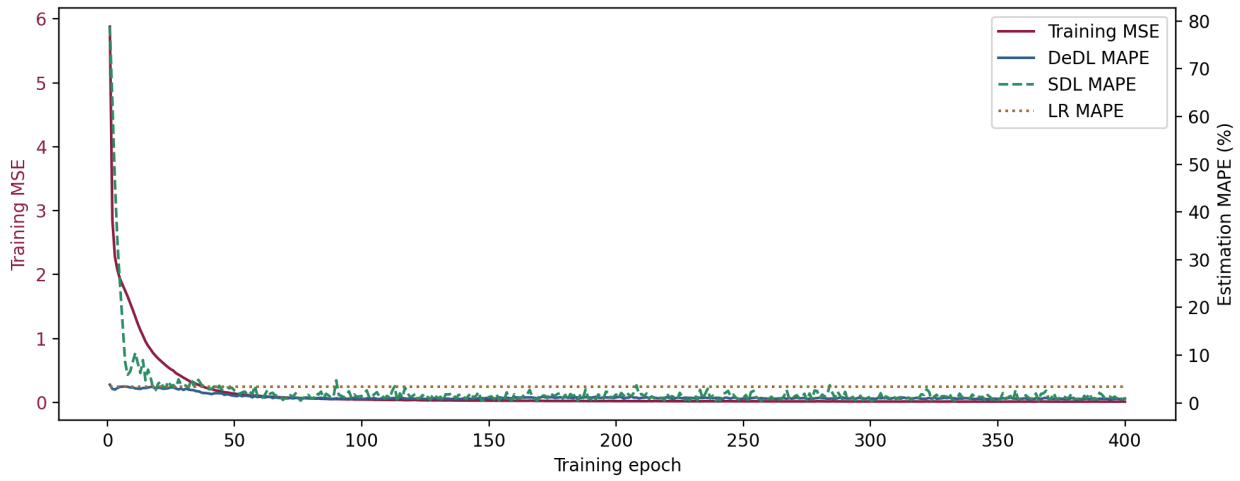
References: https://openai.com/codex/



7

7

Replication Result



8

8

The Bitter Lesson

- References: http://www.incompleteideas.net/IncIdeas/BitterLesson.html https://www.youtube.com/watch?v=vbVfAqPI8ng
- The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin.
- Leveraging domain knowledge (short-term & specific) vs. Leveraging compute (long-term & general).
- Bitter lesson: Leveraging domain knowledge is self-satisfying and intellectually inspiring, but plateaus in the long-run or even inhibits further progress.

Prof. Richard Sutton

9

9

What is AI for Business Research?

10

10

What's Special About AI: Flywheel

Business Applications

Data

Compute

AI Models

We should work on the research that scales with data and compute.

11

11

Agenda

- Course Introduction and Logistics
- AI for Business Research Landscape

12

12

Purpose of this Course

1. Have a basic understanding of the fundamental concepts/methods in machine learning (ML) and artificial intelligence (AI) that are used (or potentially useful) in business research.
2. Understand how business researchers have utilized ML/AI and what managerial questions have been addressed by ML/AI in the recent decade.
3. Nurture a taste of what the state-of-the-art AI/ML technologies can do in the ML/AI community and, potentially, in your own research field.

13

13

What's New Beyond Last Two Years?

- Roughly 80%+ new content compared with Season 1 and Season 2.
  - Only the first two sessions (ML and DL introductions) are similar to those of last two year.
- Topics of this season: Reinforcement Learning and Agentic AI.
- I have been trying my best to stay at the frontier of AI as well ☺

14

14

Other Options to Learn AI

- To learn AI, you have a lot of other options:
  - Basic ML Intro by Andrew Ng: https://www.coursera.org/specializations/machine-learning-introduction
  - Basic Deep Learning (DL) Intro by Andrew Ng: https://www.coursera.org/specializations/deep-learning
  - Natural Language Processing by Chris Manning: https://web.stanford.edu/class/cs224n/
  - Computer Vision by Fei-Fei Li: http://cs231n.stanford.edu/
  - Deep Reinforcement Learning by Sergey Levine: https://rail.eecs.berkeley.edu/deeprlcourse/
  - Deep Learning Theory by Matus Telgarsky: https://mjt.cs.illinois.edu/courses/dlt-f22/
  - Machine Learning Fairness by Mortiz Hardt: https://fairmlbook.org/
  - Language Language Models by Danqi Chen: https://www.cs.princeton.edu/courses/archive/fall22/cos597G/
  - Short Courses on Generative AI: https://www.deeplearning.ai/short-courses/
  - See https://github.com/rphilipzhang/AI-PhD-S26 **for** more resources.

15

15

Why This Course?

- A fundamental and delicate trade-off: How much to cover vs. How deep to cover.
- This course provides a concise introduction to AI/ML topics relevant to applied business research.
- For each topic, we try to cover enough necessary knowledge that could:
  - Help you understand the key trade-offs and invent new applied methods (most likely without any theoretical guarantee and in a small scale compared with industry);
  - Inform you about the literature development in the relevant domain;
  - Prepare you with the necessary sense to do rigorous business research using the relevant methods.
- We aim to cover conceptually important theories in AI/ML that can be applied in business research.
- We emphasize the combination of coding and theory so that you will be able to implement your ideas.

Impact of a CS Paper = Problem Importance * Technical Novelty * Performance Improvement

Impact of a Business Paper = Problem Importance * Identification Rigor * Insight Novelty

16

16

Why Not This Course?

- We have some assumptions on your prior knowledge:
  - Working knowledge in calculus, linear algebra, and stats;
  - Working knowledge of Python programming (but we have Codex, Claude Code and Cursor now…);
  - ML, causal inference, and econometrics: Better that you have some basic sense in them.
- We try to open doors and windows for you instead of preparing you to be a leading expert in a specific domain.
- I am trying my best to stay at the frontier, but some of the knowledge is outdated/constrained by academia.

Warning 0: At CUHK, we have an Econ course of similar (to Season 1) topics (ECON 5180) without the coding emphasis.

Warning 1: This may be your MOST time-consuming course at CUHK by a wide margin.

Warning 2: We will mainly talk about the ideas and methods (with demos) in class, but you will need some (vibe) coding skills to finish your homework and replication project.

17

17

Course Format

- We have a 2-hrs-and-45mins long course each week.
- For each session:
  - 15 mins: Homework discussions and review of previous content;
  - 105 mins: Theories and coding demos;
  - 30 mins: Student presentations.
- All coursework will be done in groups of at most TWO students.
  - Register your group members (and majors) and your group name by 11:59pm, Jan. 7, 2026.
  - Otherwise, we will match you with others (based on majors).
- You will need to evaluate your buddy's contribution in all the coursework.

18

18

Coursework and Grading

- Coursework:
  - Lecture notes scribing (each group will scribe the lecture note of one session/topic)
  - Paper replication and presentation (one paper replication and presentation per group each week)
  - Homework (one coding assignment each week, due two weeks after distribution; 5 assignments count)
  - Final Project (one final project based on your own choice).
- Grading:
  - See Syllabus: https://github.com/rphilipzhang/AI-PhD-S26/blob/main/AI-PhD-Syllabus-S2026.pdf
- All homework/final project will be done in Python.

19

19

Coursework Materials

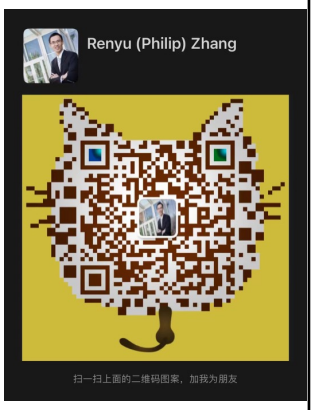
- GitHub: https://github.com/rphilipzhang/AI-PhD-S26
  - All course materials will be distributed on this GitHub Repository.
- Google Sheet: https://docs.google.com/spreadsheets/d/1YIwCR-X8VVLTv-OfGr7DqZJF6YIll-JE3SGI1q7JEus/edit?usp=sharing
  - Group Registration
  - Lecture Notes Scribing Sign-up
  - Paper Replication and Presentation Sign-up
  - Project Presentation Sign-up
  - Homework Submission (use the link to your Google CoLab and opensource your code to your classmates by "Anyone with the link can view")
- Google CoLab: https://drive.google.com/drive/folders/19806VHq6Vybrx-z4BbvV01sTMs9G5Hrh?usp=sharing
  - All code demos will be distributed via Google CoLab or GitHub.
- Registered students please ask our TA, Ignis Jiang, to add your account to our course Google Sheet.

20

20

Course Communications

- Class Meeting: Tuesday, 9:30AM-12:15PM (@CYT 209A; and @ CYT LT5 on March 3)
- Office hour: By appointment, @CYT_911
- WeChat group: Online discussion forum
  - 200+ group members already
- Instructor contact
  - Office: CYT_911
  - Email: philipzhang@cuhk.edu.hk
  - Tel: 852-3943-7763
  - WeChat: rphilip_zhang
- Teaching Assistant: Ignis Jiang
  - Email: ignisjiang@cuhk.edu.hk



21

21

Python Tutorial Sessions

- We have two optional Python tutorial sessions held online at Friday night, 7:00pm-9:00pm.
- Tutorial Instructor: Xinyu Li, MIS PhD Candidate @CUHK Business School, xinyu.li@link.cuhk.edu.hk
- Check the course GitHub Repo for CoLab and Zoom links.
- Session 1: Friday, Jan 9, 2026
  - Python Basics
- Session 2: Friday, Jan 16, 2026
  - PyTorch Basics & FBA/DOT Server
- References:
  - https://drive.google.com/drive/folders/1s7enA0zxHnSC5f2cm-IhjHcJmMuIH7KS?usp=sharing
  - https://cs231n.github.io/python-numpy-tutorial/
  - https://colab.research.google.com/github/cs231n/cs231n.github.io/blob/master/python-colab.ipynb

22

22

Agenda

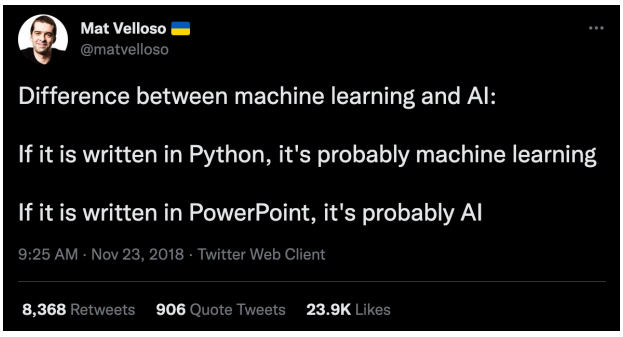
- Course Introduction and Logistics
- AI for Business Research Landscape

23

23

What is AI/ML?

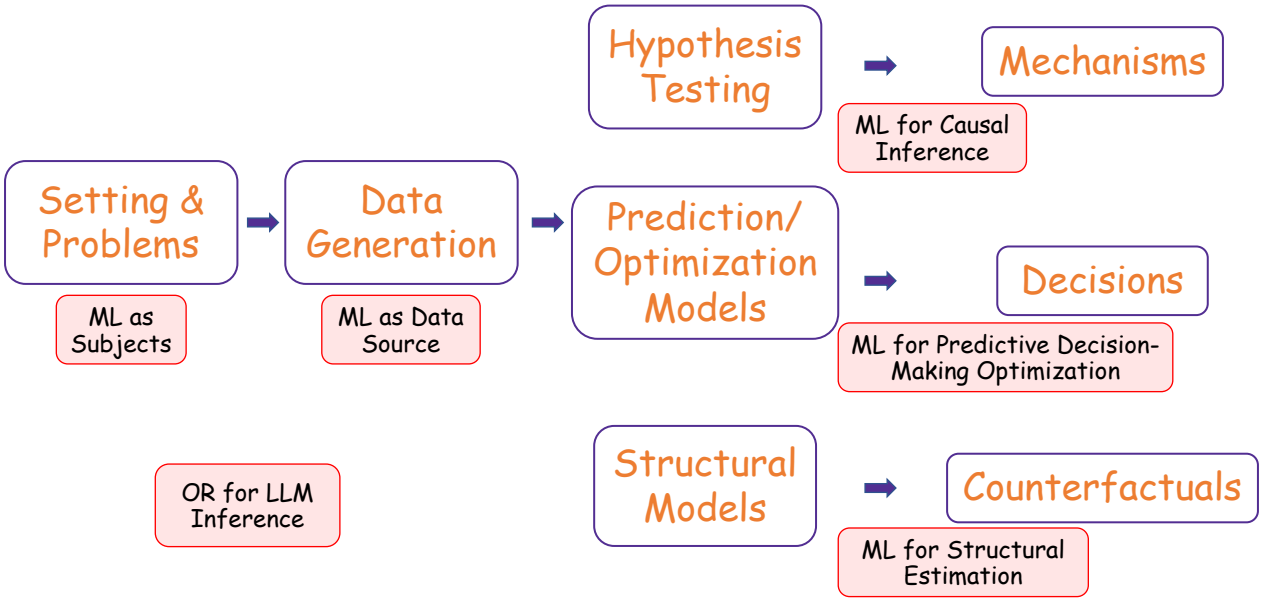
- ML is a CS subfield that automates computers to learn from data without explicitly programmed.
- Different names:
  - Data mining
  - Statistical learning
  - Data science



24

24

A Typical Applied Business Research Paper



25

25

Landscape of AI/ML for Business Research

- AI/ML as Data/Data Source
  - Kong L., Jin J., Zhang R. (2026) Improving Behavioral Alignment in LLM Social Simulations via Context Formation and Navigation, *working paper*.
- AI/ML for Causal Inference
  - Ye, Z., Zhang, Z, Zhang, D. J., Zhang, H., Zhang, R. (2023) Deep Learning Based Causal Inference for Large-Scale Combinatorial Experiments: Theory and Empirical Evidence, *Management Science*.
- AI/ML for Predictive Decision Making and Optimization
  - Ye, Z., Zhang, D. J., Zhang, H., Zhang, R., Chen, X., and Xu, Z. (2023) Cold start to improve market thickness on online advertising platforms: Data-driven algorithms and field experiments. *Management Science*, 69(7), 3838-3860.
- AI/ML as Subjects
  - Zhang, X., Sun, C., Zhang, R., and Goh, K-Y (2024) The Value of AI-Generated Metadata for UGC Platforms: Evidence from a Large-scale Field Experiment, *working paper* and *CIST 2024*.
- ML for Structural Estimation
- OR for LLM Inference

26

26

ML as Data/Data Source

Home > Marketing Science > Ahead of Print >

Engagement That Sells: Influencer Video Advertising on TikTok

Jeremy Yang, Juanjuan Zhang, Yuhan Zhang

Published Online: 20 Dec 2024 | https://doi.org/10.1287/mksc.2021.0107

Abstract

Many ads are engaging, but what makes them engaging may have little to do with the product. This problem can be particularly relevant to influencer advertising if influencers are motivated to promote themselves, not just the product. We develop an algorithm to measure the degree of effective engagement associated with the product and use it to predict the sales lift of influencer video advertising. We propose the concept of the product engagement score (PE score) to capture how engaging the product itself is as presented in a video. We estimate pixel-level engagement as a saliency map by training a deep three-dimensional convolutional neural network on video-level engagement data. We locate pixel-level product placement with an object detection algorithm. The PE score is computed as the pixel-level, engagement-weighted product placement in a video. We construct and validate the algorithm with influencer video ads on TikTok and product sales data on Taobao. We use variation in video posting time to identify video-specific sales lift and show that the PE score significantly and robustly predicts sales lift. We explore drivers of engagement and discuss how various stakeholders in influencer advertising can use the PE score in a scalable way to manage content, align incentives, and improve efficiency.

History: Olivier Toubia served as the senior editor.

Annual Review of Economics

Text Algorithms in Economics

Elliott Ash[1] and Stephen Hansen[2,3]

[1]Center for Law and Economics, ETH Zurich, Zurich, Switzerland

[2]Department of Economics, University College London, London, United Kingdom; email: stephen.hansen@ucl.ac.uk

[3]Centre for Economic Policy Research, London, United Kingdom

27

27

ML as Data Source

- Any recordable information that is not numerical can be analyzed with ML to answer business questions.
- Generative AI makes analyzable information much broader than before.
- References:
  - Text – Natural Language Processing (NLP)
  - Image/Video – Computer Vision (CV)
  - Sound – Deep Learning (DL)
  - Genetic information - Bioinformatics
  - And many more…

28

28

ML as Data Source

- Why do we use ML to understand unstructured data?
  - Cost reduction and scalability
  - Objectivity
  - Easy to built into other systems
- Issues with using ML to understand unstructured data:
  - Measurement errors
  - Interpretation

29

29

Issues with ML as Data Source

- Empirical model: $Y = a + b \cdot D + g(X) + \epsilon$
  - Key parameter of interest: b
- Outcome
  - Y may be generated through ML with error (of less concern).
- Treatment
  - D may be generated through ML with error which is correlated with ϵ
  - https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4480696; https://arxiv.org/abs/2402.15585; https://www.nber.org/papers/w33344
- Controls
  - X may be generated through ML with error.
  - X may be selected by ML with error.
  - Double machine learning can be applied for effective debias.

30

30

LLM for Social Simulations

Position: LLM Social Simulations Are a Promising Research Method

Jacy R. Anthis [1 2 3] Ryan Liu [4] Sean M. Richardson [1] Austin C. Kozlowski [1]
Bernard Koch [1] Erik Brynjolfsson [2] James Evans [1 5] Michael S. Bernstein [2]

Abstract

Accurate and verifiable large language model (LLM) simulations of human research subjects promise an accessible data source for understanding human behavior and training new AI systems. However, results to date have been limited, and few social scientists have adopted this method. In this position paper, we argue that the promise of LLM social simulations can be achieved by addressing five tractable challenges. We ground our argument in a review of empirical comparisons between LLMs and human research subjects, commentaries on the topic, and related work. We identify promising directions, including context-rich prompting and fine-tuning with social science datasets. We believe that LLM social simulations can already be used for pilot and exploratory studies, and more widespread use may soon be possible with rapidly advancing LLM capabilities. Researchers should prioritize developing conceptual models and iterative evaluations to make the best use of new AI systems.

high-quality synthetic data for the development of human-centered AI at scale (Bai et al., 2022; Kim et al., 2023). Nonetheless, the limitations of LLMs and simulation results to date have cast doubt on whether accurate and reliable simulation is possible (Agnew et al., 2024; Gao et al., 2024; Wang et al., 2024a,b).

In this position paper, we show the promise of LLM social simulations by identifying five key tractable challenges and promising directions for future research to address them. We summarize the challenges in Table 1: diversity, bias, sycophancy, alienness, and generalization. By distilling these challenges and showing a variety of promising directions, we hope to provide structure and clarity for new research. Our argument is grounded in a literature review of empirical studies that have compared human research subjects to LLMs, commentaries on the topic, and related work in social science and other LLM applications. Compelling simulation results so far include:

- Hewitt et al. (2024), the largest test of sims to date, spanned 70 preregistered and U.S.-representative experiments alongside an archive of replication studies.

informs
https://pubsonline.informs.org/journal/mksc

MARKETING SCIENCE
Articles in Advance, pp. 1–10
ISSN 0732-2399 (print), ISSN 1526-548X (online)

Twin-2K-500: A Data Set for Building Digital Twins of over 2,000 People Based on Their Answers to over 500 Questions

Olivier Toubia,[a,*] George Z. Gui,[a] Tianyi Peng,[b] Daniel J. Merlau,[a] Ang Li,[c] Haozhe Chen[c]

[a] Marketing Division, Columbia Business School, Columbia University, New York, New York 10027; [b] Decision, Risk & Operations Division, Columbia Business School, Columbia University, New York, New York 10027; [c] Department of Computer Science, Columbia University, New York, New York 10025

*Corresponding author

Contact: ot2107@gsb.columbia.edu, https://orcid.org/0000-0001-7493-9641 (OT); zg2467@gsb.columbia.edu, https://orcid.org/0000-0001-9399-649X (GZG); tianyi.peng@columbia.edu (TP); djm2199@columbia.edu (DJM); al4263@columbia.edu (AL); tonychenxyz@gmail.com (HC)

Received: May 29, 2025
Revised: June 18, 2025
Accepted: June 20, 2025
Published Online in Articles in Advance: August 20, 2025

https://doi.org/10.1287/mksc.2025.0262

Abstract. Large language model (LLM)-based digital twin simulation, where LLMs are used to emulate individual human behavior, holds great promise for research in business, artificial intelligence, social science, and digital experimentation. However, progress in this area has been hindered by the scarcity of real individual-level data sets that are both large and publicly available. To address this gap, we introduce a large-scale public data set designed to capture a rich and holistic view of individual human behavior. We survey a representative sample of $N = 2{,}058$ participants (average 2.42 hours per person) in the United States across four waves with more than 500 questions in total, covering a comprehensive battery of demographic, psychological, economic, personality, and cognitive measures, as well as replications of behavioral economics experiments and a pricing survey. The final wave repeats tasks from earlier waves to establish a test-retest accuracy baseline. Initial analyses suggest the data are of high quality and show promise for constructing digital twins that predict human behavior well at the individual and aggregate levels. Beyond LLM applications, due to its unique breadth and scale, the data set also enables broad social science and business research, including studies of cross-construct correlations and heterogeneous treatment effects.

History: Hema Yoganarasimhan served as the senior editor.

Funding: This study was funded by the Columbia Business School Digital Future Initiative.
Supplemental Material: The online appendix and data files are available at https://doi.org/10.1287/mksc.2025.0262.

Keywords: generative AI • computational social science • digital twins • LLM-based persona simulation

ML for Causal Inference

Journal of Marketing Research
Volume 61, Issue 3, June 2024, Pages 472-495

https://doi.org/10.1177/00222437231210267

Sage Journals

Article

Mega or Micro? Influencer Selection Using Follower Elasticity

Zijun Tian, Ryan Dew, and Raghuram Iyengar

Abstract

Influencer marketing, in which companies sponsor social media personalities to promote their brands, has exploded in popularity in recent years. One common criterion for selecting an influencer partner is popularity. While some firms collaborate with "mega" influencers with millions of followers, other firms partner with "micro" influencers with only several thousand followers, but who also cost less to sponsor. To quantify this trade-off between popularity and cost, the authors develop a framework for estimating the follower elasticity of impressions (FEI), which measures a video's percentage gain in impressions (i.e., views) corresponding to a percentage increase in the number of followers of its creator. Computing FEI involves estimating the causal effect of an influencer's popularity on the view counts of their videos, which is achieved through a combination of (1) a unique data set collected from TikTok, (2) a representation learning model for quantifying video content, and (3) a machine learning–based causal inference method. The authors find that FEI is always positive, averaging .10, but often nonlinearly related to follower size. They examine the factors that predict variation in these FEI curves and show how firms can use these results to better determine influencer partnerships.

Keywords

influencer marketing, causal inference, deep learning, representation learning, heterogeneous treatment effects, video data

https://causalml-book.org/
https://github.com/rphilipzhang/AI-PhD-S25

informs
https://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Articles in Advance, pp. 1–15
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Targeting for Long-Term Outcomes

Jeremy Yang,[a,*] Dean Eckles,[b,*] Paramveer Dhillon,[c] Sinan Aral[b]

[a] Harvard Business School, Boston, Massachusetts 02163; [b] Massachusetts Institute of Technology, Cambridge, Massachusetts 02142; [c] University of Michigan, Ann Arbor, Michigan 48109

*Corresponding authors

Contact: jeryang@hbs.edu, https://orcid.org/0000-0001-8639-5493 (JY); eckles@mit.edu, https://orcid.org/0000-0001-8439-442X (DE); dhillonp@umich.edu, https://orcid.org/0000-0002-0994-9488 (PD); sinan@mit.edu, https://orcid.org/0000-0002-2762-058X (SA)

Received: October 7, 2020
Revised: February 12, 2022
Accepted: February 27, 2022
Published Online in Articles in Advance: August 3, 2023

https://doi.org/10.1287/mnsc.2023.4881

Abstract. Decision makers often want to target interventions so as to maximize an outcome that is observed only in the long term. This typically requires delaying decisions until the outcome is observed or relying on simple short-term proxies for the long-term outcome. Here, we build on the statistical surrogacy and policy learning literatures to impute the missing long-term outcomes and then approximate the optimal targeting policy on the imputed outcomes via a doubly robust approach. We first show that conditions for the validity of average treatment effect estimation with imputed outcomes are also sufficient for valid policy evaluation and optimization; furthermore, these conditions can be somewhat relaxed for policy optimization. We apply our approach in two large-scale proactive churn management experiments at *The Boston Globe* by targeting optimal discounts to its digital subscribers with the aim of maximizing long-term revenue. Using the first experiment, we evaluate this approach empirically by comparing the policy learned using imputed outcomes with a policy learned on the ground-truth, long-term outcomes. The performance of these two policies is statistically indistinguishable, and we rule out large losses from relying on surrogates. Our approach also outperforms a policy learned on short-term proxies for the long-term outcome. In a second field experiment, we implement the optimal targeting policy with additional randomized exploration, which allows us to update the optimal policy for future subscribers. Over three years, our approach had a net-positive revenue impact in the range of $4–$5 million compared with the status quo.

History: Accepted by Eric Anderson, marketing.
Funding: This work was supported by Boston Globe Media.
Supplemental Material: The online appendix and data are available at https://doi.org/10.1287/mnsc.2023.4881.

Keywords: long-term effect • statistical surrogate • policy learning • targeting • proactive churn management

https://bookdown.org/stanfordgsbsilab/ml-ci-tutorial/

ML for Predictive Decision-Making &Optimization

informs
http://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH
Vol. 70, No. 1, January–February 2022, pp. 309–328
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Crosscutting Areas

Customer Choice Models vs. Machine Learning: Finding Optimal Product Displays on Alibaba

Jacob Feldman,[a] Dennis J. Zhang,[a] Xiaofei Liu,[b] Nannan Zhang[b]

[a]Olin Business School, Washington University in St. Louis, St. Louis, Missouri 63130; [b]Alibaba Group Inc., Hangzhou 311100, China
Contact: jbfeldman@wustl.edu, https://orcid.org/0000-0002-5576-1953 (JF); denniszhang@wustl.edu (DJZ); xiaofei.liu@alibaba.com (XL); nannan.zhang@alibaba.com (NZ)

Received: November 5, 2019
Revised: November 18, 2019; August 25, 2020; February 2, 2021
Accepted: April 6, 2021
Published Online in Articles in Advance: October 26, 2021

Area of Review: OR Practice

https://doi.org/10.1287/opre.2021.2158

Abstract. We compare the performance of two approaches for finding the optimal set of products to display to customers landing on Alibaba's two online marketplaces, Tmall and Taobao. We conducted a large-scale field experiment, in which we randomly assigned 10,421,649 customer visits during a one-week-long period to one of the two approaches and measured the revenue generated per customer visit. The first approach we tested was Alibaba's current practice, which embeds product and customer features within a sophisticated machine-learning algorithm to estimate the purchase probabilities of each product for the customer at hand. The products with the largest expected revenue (revenue × predicted purchase probability) are then made available for purchase. Our second approach, which we developed and implemented in collaboration with Alibaba engineers, uses a featurized multinomial logit (MNL) model to predict purchase probabilities for each arriving customer. We used historical sales data to fit the MNL model, and then, for each arriving customer, we solved a cardinality-constrained assortment-optimization problem under the MNL model to find the optimal set of products to display. Our field experiments revealed that the MNL-based approach generated 5.17 renminbi (RMB) per customer visit, compared with the 4.04 RMB per customer visit generated by the machine-learning-based approach when both approaches were given access to the same set of the 25 most important features. This improvement represents a 28% gain in revenue per customer visit, which corresponds to a 4 million RMB improvement over the week in which the experiments were conducted. Motivated by the results of our initial field experiment, Alibaba then implemented a full-featured version of our MNL-based approach, which now serves the majority of customers in this setting. Using another small-scale field experiment, we estimate that our new MNL-based approach that utilizes the full feature set is able to increase Alibaba's annual revenue by 87.26 million RMB (12.42 million U.S. dollars).

Supplemental Material: The online appendix is available at https://doi.org/10.1287/opre.2021.2158.

Keywords: choice models • product assortment • machine learning • field experiment • retail operations

informs
https://pubsonline.informs.org/journal/mksc

MARKETING SCIENCE
Articles in Advance, pp. 1–22
ISSN 0732-2399 (print), ISSN 1526-548X (online)

Dynamic Coupon Targeting Using Batch Deep Reinforcement Learning: An Application to Livestream Shopping

Xiao Liu[a]

[a]Stern School of Business, New York University, New York, New York 10012
Contact: xl23@stern.nyu.edu, https://orcid.org/0000-0002-7093-8534 (XL)

Received: September 3, 2020
Revised: September 5, 2021; April 17, 2022
Accepted: June 23, 2022
Published Online in Articles in Advance: October 20, 2022

https://doi.org/10.1287/mksc.2022.1403

Abstract. We present an empirical framework for creating dynamic coupon targeting strategies for high-dimensional and high-frequency settings, and we test its performance using a large-scale field experiment. The framework captures consumers' intertemporal tradeoffs associated with dynamic pricing and does not rely on functional form assumptions about consumers' decision-making processes. The model is estimated using batch deep reinforcement learning (BDRL), which relies on Q-learning, a model-free solution that can mitigate model bias. It leverages deep neural networks to represent the high-dimensional state space and alleviate the curse of dimensionality. The empirical application is in a multibillion-dollar livestream shopping context. Our BDRL solution increases the platform's revenue by twice as much as static targeting policies and by 20% more than the model-based solution. The comparative advantage of BDRL comes from more effective and automatic targeting of consumers based on both heterogeneity and dynamics, using exceptionally rich, nuanced differences among consumers and across time. We find that price skimming, reducing discounts for attractive hosts, and increasing the coupon discount level at a faster rate for low spenders are effective strategies based on dynamics, consumer heterogeneity, and the two combined, respectively.

History: K. Sudhir served as the senior editor and John Hauser served as associate editor for this article.
Funding: Partial financial support was received from the NYU Center for Global Economy and Business.
Supplemental Material: The data files and online appendices are available at https://doi.org/10.1287/mksc.2022.1403.

Keywords: dynamic pricing • coupon • deep reinforcement learning • reference price • livestream shopping • targeting

LLM for OR

Home > Operations Research > Vol. 73, No. 6 >

ORLM: A Customizable Framework in Training Large Models for Automated Optimization Modeling

Chenyu Huang, Zhengyang Tang, Shixi Hu, Ruoqing Jiang, Xin Zheng, Dongdong Ge, Benyou Wang, Zizhuo Wang

Published Online: 8 May 2025 | https://doi.org/10.1287/opre.2024.1233

Abstract

Optimization modeling plays a critical role in the application of Operations Research (OR) tools to address real-world problems, yet they pose challenges and require extensive expertise from OR experts. With the advent of large language models (LLMs), new opportunities have emerged to streamline and automate such tasks. However, current research predominantly relies on closed-source LLMs, such as GPT-4, along with extensive prompt engineering techniques. This reliance stems from the scarcity of high-quality training data sets for optimization modeling, resulting in elevated costs, prolonged processing times, and privacy concerns. To address these challenges, our work is the first to propose a viable path for training open-source LLMs that are capable of optimization modeling and developing solver codes, eventually leading to a superior ability for automating optimization modeling and solving. Particularly, we design the OR-Instruct, a semiautomated data synthesis framework for optimization modeling that enables customizable enhancements for specific scenarios or model types. This work also introduces IndustryOR, the first industrial benchmark for evaluating LLMs in solving practical OR problems. We train several 7B-scale open-source LLMs using synthesized data (dubbed ORLMs), which exhibit significantly enhanced optimization modeling capabilities, achieving competitive performance across the NL4Opt, MAMO, and IndustryOR benchmarks. Additionally, our experiments highlight the potential of scaling law and reinforcement learning to further enhance the performance of ORLMs. The workflows and human-machine interaction paradigms of ORLMs in practical industrial applications are also discussed in the paper.

CALM Before the STORM: Unlocking Native Reasoning for Optimization Modeling

Zhengyang Tang*[1,5], Zihan Ye*[1], Chenyu Huang*[2], Xuhan Huang[1], Chengpeng Li[5], Sihang Li[3], Guanhua Chen[3], Ming Yan[1], Zizhuo Wang[1], Hongyuan Zha[1], Dayiheng Liu[†5], and Benyou Wang[†1,4]

[1]The Chinese University of Hong Kong, Shenzhen
[2]Shanghai University of Finance and Economics
[3]Southern University of Science and Technology
[4]Shenzhen Loop Area Institute (SLAI)
[5]Qwen Team, Alibaba Inc.

Abstract

Large Reasoning Models (LRMs) have demonstrated strong capabilities in complex multi-step reasoning, opening new opportunities for automating optimization modeling. However, existing domain adaptation methods, originally designed for earlier instruction-tuned models, often fail to exploit the advanced reasoning patterns of modern LRMs — In particular, we show that direct fine-tuning on traditional *non-reflective* datasets leads to limited gains. To fully leverage LRMs' inherent reasoning abilities, we propose **CALM** (*Corrective Adaptation with Lightweight Modification*), a framework that progressively refines LRMs within their native reasoning modes for optimization modeling tasks. In CALM, an expert intervener identifies reasoning flaws and provides concise corrective hints, which the LRM incorporates to produce improved reasoning trajectories. These interventions modify fewer than 2.6% of generated tokens, but generate high-quality data for soft adaptation through supervised fine-tuning. The adapted model is then further improved through reinforcement learning. Building on CALM, we develop **STORM** (*Smart Thinking Optimization Reasoning Model*), a 4B-parameter LRM that achieves a new state-of-the-art average accuracy of 68.9% across five popular optimization modeling benchmarks, matching the performance of a 671B LRM. These results demonstrate that dynamic, hint-based data synthesis both preserves and amplifies the native reasoning patterns of modern LRMs, offering a more effective and scalable path towards expert-level performance on challenging optimization modeling tasks.

Generative AI for Content/Decision

Home > Marketing Science > Ahead of Print >

Applying Large Language Models to Sponsored Search Advertising

Martin Reisenbichler, Thomas Reutterer, David A. Schweidel

Published Online: 3 Nov 2025 | https://doi.org/10.1287/mksc.2023.0611

Abstract

With the increasing availability of powerful large language models (LLMs), the generation of textual marketing content has become more accessible. In this research, we examine the potential to tailor an LLM for application to search engine advertising (SEA). That is, we develop and evaluate an "application layer" that sits on top of an open-source LLM to generate ad text "fine-tuned" to the SEA context. With a goal of maximizing clicks to improve online visibility in a cost per click (CPC) setup, we experimentally test our framework in two empirical settings. Our results demonstrate the superior performance of a human-in-the-loop generative artificial intelligence (AI) approach to advertising content generation compared with ads created by humans and standard LLMs. We show that our approach yields improved performance, but potentially incurs a higher CPC, making it necessary to balance content optimization and cost. Our research demonstrates the performance gains achievable through the development of tailored LLM-based applications. Using our framework, we also identify boundary conditions that appear to limit the benefits of using generative AI in support of SEA, offering substantive insights to both practitioners and researchers.

History: Olivier Toubia served as the senior editor.

OneRec: Unifying Retrieve and Rank with Generative Recommender and Iterative Preference Alignment

Jiaxin Deng, Shiyao Wang, Kuo Cai, Lejian Ren, Qigen Hu, Weifeng Ding, Qiang Luo, Guorui Zhou

Recently, generative retrieval-based recommendation systems have emerged as a promising paradigm. However, most modern recommender systems adopt a retrieve-and-rank strategy, where the generative model functions only as a selector during the retrieval stage. In this paper, we propose OneRec, which replaces the cascaded learning framework with a unified generative model. To the best of our knowledge, this is the first end-to-end generative model that significantly surpasses current complex and well-designed recommender systems in real-world scenarios. Specifically, OneRec includes: 1) an encoder-decoder structure, which encodes the user's historical behavior sequences and gradually decodes the videos that the user may be interested in. We adopt sparse Mixture-of-Experts (MoE) to scale model capacity without proportionally increasing computational FLOPs. 2) a session-wise generation approach. In contrast to traditional next-item prediction, we propose a session-wise generation, which is more elegant and contextually coherent than point-by-point generation that relies on hand-crafted rules to properly combine the generated results. 3) an Iterative Preference Alignment module combined with Direct Preference Optimization (DPO) to enhance the quality of the generated results. Unlike DPO in NLP, a recommendation system typically has only one opportunity to display results for each user's browsing request, making it impossible to obtain positive and negative samples simultaneously. To address this limitation, We design a reward model to simulate user generation and customize the sampling strategy. Extensive experiments have demonstrated that a limited number of DPO samples can align user interest preferences and significantly improve the quality of generated results. We deployed OneRec in the main scene of Kuaishou, achieving a 1.6\% increase in watch-time, which is a substantial improvement.

In the third quarter of 2025, we achieved strong results from integrating AI technology into diverse internal and external application scenarios. In terms of business empowerment, large AI models have now been integrated across all major business scenarios of Kuaishou, driving incremental value across the Kuaishou's ecosystem. We iterated our end-to-end generative recommendation large model ***OneRec*** and extended this new technological paradigm beyond short video recommendations to additional recommendation scenarios, such as online marketing services and e-commence shopping mall. This expansion has generated meaningful incremental benefits. In the third quarter of 2025, large AI models technology has demonstrated notable effects, especially in online marketing services. We pioneered a generative reinforcement learning-based bidding model ***G4RL*** that integrates sequence modeling with goal optimization. This innovation marked a breakthrough in advertising bidding, advancing from single-step decision-making to long-term strategic planning. Meanwhile, we explored the application of end-to-end generative recommendation technology in online marketing service scenarios through ***OneRec***. Tailored to the characteristics of online marketing services, we introduced the client marketing expression and marketing commercial value (CPM) perception mechanism to achieve bidirectional matching between users' interests and clients' demands, thereby further enhancing the personalization and matching efficiency of online marketing materials. The application of large AI models technology, especially ***OneRec***, has driven an approximately additional 4%–5% growth in domestic online marketing service revenue in the third quarter of 2025. In terms of online marketing material generation, the total spending from online marketing services driven by AIGC marketing materials exceeded RMB3.0 billion in the third quarter of 2025.

35

35

ML as Subject

RESEARCH ARTICLE | CHATGPT

Experimental evidence on the productivity effects of generative artificial intelligence

SHAKKED NOY AND WHITNEY ZHANG Authors Info & Affiliations

SCIENCE · 13 Jul 2023 · Vol 381, Issue 6654 · pp. 187-192 · DOI: 10.1126/science.adh2586

62,070 2

Editor's summary

Automation has historically displaced human workers in factories (e.g., automotive manufacturing) or in performing routine computational tasks. Will generative artificial intelligence (AI) tools such as ChatGPT disrupt the labor market by making educated professionals obsolete, or will these tools complement their skills and enhance productivity? Noy and Zhang examined this issue in an experiment that recruited college-educated professionals to complete incentivized writing tasks. Participants assigned to use ChatGPT were more productive, efficient, and enjoyed the tasks more. Participants with weaker skills benefited the most from ChatGPT, which carries policy implications for efforts to reduce productivity inequality through AI. —EEU

informs
http://pubsonline.informs.org/journal/mnsc/

MANAGEMENT SCIENCE
Vol. 65, No. 7, July 2019, pp. 2966–2981
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads

Anja Lambrecht,[a] Catherine Tucker[b]

[a] Marketing, London Business School, London NW1 4SA, United Kingdom; [b] Marketing, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142

Contact: alambrecht@london.edu, http://orcid.org/0000-0001-6766-1602 (AL); cetucker@mit.edu, http://orcid.org/0000-0002-1847-4832 (CT)

Received: November 28, 2017
Revised: March 2, 2018
Accepted: March 13, 2018
Published Online in Articles in Advance: April 10, 2019

https://doi.org/10.1287/mnsc.2018.3093

Abstract. We explore data from a field test of how an algorithm delivered ads promoting job opportunities in the science, technology, engineering and math fields. This ad was explicitly intended to be gender neutral in its delivery. Empirically, however, fewer women saw the ad than men. This happened because younger women are a prized demographic and are more expensive to show ads to. An algorithm that simply optimizes cost-effectiveness in ad delivery will deliver ads that were intended to be gender neutral in an apparently discriminatory way, because of crowding out. We show that this empirical regularity extends to other major digital platforms.

History: Accepted by Joshua Gans, business strategy.
Funding: Supported by a National Science Foundation Career Award [Grant 6923256].

Keywords: algorithmic bias • online advertising • algorithms • artificial intelligence

AI/ML as subjects: Economics of AI; Machine human collaboration; ML fairness/discrimination; ML and labor market; data privacy; Data and ML in IO; AI as a species, etc.

36

36

ML for Structural Estimation

Econometrica, Vol. 91, No. 6 (November, 2023), 2041–2063

AN ADVERSARIAL APPROACH TO STRUCTURAL ESTIMATION

TETSUYA KAJI
University of Chicago Booth School of Business

ELENA MANRESA
Department of Economics, New York University

GUILLAUME POULIOT
University of Chicago Harris School of Public Policy

We propose a new simulation-based estimation method, adversarial estimation, for structural models. The estimator is formulated as the solution to a minimax problem between a generator (which generates simulated observations using the structural model) and a discriminator (which classifies whether an observation is simulated). The discriminator maximizes the accuracy of its classification while the generator minimizes it. We show that, with a sufficiently rich discriminator, the adversarial estimator attains parametric efficiency under correct specification and the parametric rate under misspecification. We advocate the use of a neural network as a discriminator that can exploit adaptivity properties and attain fast rates of convergence.

KEYWORDS: Structural estimation, generative adversarial networks, neural networks, simulation-based estimation, efficient estimation.

Home > Marketing Science > Ahead of Print >

Estimating Parameters of Structural Models Using Neural Networks

Yanhao (Max) Wei, Zhenling Jiang

Published Online: 16 Aug 2024 | https://doi.org/10.1287/mksc.2022.0360

Abstract

We study an alternative use of machine learning. We train neural nets to provide the parameter estimate of a given (structural) econometric model, for example, discrete choice or consumer search. Training examples consist of datasets generated by the econometric model under a range of parameter values. The neural net takes the moments of a dataset as input and tries to recognize the parameter value underlying that dataset. Besides the point estimate, the neural net can also output statistical accuracy. This neural net estimator (NNE) tends to limited-information Bayesian posterior as the number of training datasets increases. We apply NNE to a consumer search model. It gives more accurate estimates at lighter computational costs than the prevailing approach. NNE is also robust to redundant moment inputs. In general, NNE offers the most benefits in applications where other estimation approaches require very heavy simulation costs. We provide code at: https://nnehome.github.io.

History: Manchanda Puneet served as the senior editor.

37

37

OR for LLM Inference

Optimizing LLM Inference: Fluid-Guided Online Scheduling with Memory Constraints

Ruicheng Ao*
Massachusetts Institute of Technology

Gan Luo†
Peking University

David Simchi-Levi*
Massachusetts Institute of Technology

Xinshang Wang‡
Alibaba Group

April 16, 2025

Abstract

Large Language Models (LLMs) is indispensable in today's applications, but their inference procedure—generating responses by breaking text into smaller pieces and processing them using a memory-heavy element named Key-Value (KV) cache—requires a lot of computational resources, especially when memory is limited. This paper treats LLM inference optimization as a multi-stage online scheduling problem, where prompts arrive sequentially and the incremental expansion of the KV cache during inference renders conventional scheduling algorithms ineffective. To address this challenge, we develop a fluid dynamics approximation to establish a tractable benchmark, providing insights for devising effective scheduling algorithms. Building upon this foundation, we introduce the Waiting for Accumulated Inference Threshold (WAIT) algorithm as a warm-up. This method maintains multiple thresholds to determine the scheduling order of incoming prompts, optimizing resource utilization when output lengths are known at the time of arrival. In practical applications where output lengths are not known at the time of prompt arrival, we extend our method by introducing the Nested WAIT algorithm. This algorithm constructs a hierarchical framework comprising multiple segments, each defined by distinct thresholds, to effectively manage the random prompt arrivals with unknown output lengths. Theoretical analysis shows both algorithms have near-optimal performance compared with the fluid benchmark under heavy traffic limit, balancing throughput, latency, and Time to First Token (TTFT). Numerical experiments conducted with the Llama-7B model on an A100 GPU, utilizing both synthetic and real-world datasets, demonstrate that our approach achieves superior throughput and reduced latency compared to widely adopted baseline methods such as vLLM and Sarathi. This research bridges operations research and machine learning, presenting a theoretically grounded framework for the efficient deployment of large language models under memory constraints.

Keywords: Large Lanugage Model, Key-value cache, Memory Constraint, Online scheduling

Fundamental Modeling for LLM Inference with Exploding KV Cache Demands

Patrick Jaillet
Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, jaillet@mit.edu

Jiashuo Jiang
HKUST, jsjiang@ust.hk

Chara Podimata
Sloan School of Management, Massachusetts Institute of Technology podimata@mit.edu

Zijie Zhou*
Operations Research Center, Massachusetts Institute of Technology, zhou98@mit.edu

In the rapidly advancing field of artificial intelligence, Large Language Models (LLMs) are crucial for many applications, demanding efficient computational strategies for inference. LLM inference, where a trained model generates text one word at a time in response to prompts, is resource-intensive, consuming significant electricity and water. This paper models LLM inference, focusing on reducing redundant computations through a special memory-saving mechanism. This mechanism temporarily stores information from each word the model processes into the KV (key-value) cache to avoid recalculating it repeatedly. However, as more words are processed, this storage can quickly reach its limit. When this happens, the system incurs substantial extra costs by reprocessing tasks. We optimize batching and scheduling strategies to manage KV cache memory usage and minimize the inference latency to improve efficiency and sustainability.

We address this challenge by first analyzing a semi-online model, where all prompts arrive initially and must be processed sequentially. For this case, we develop a polynomial-time algorithm that achieves exact optimality. Next, we examine the fully online setting with sequential prompt arrivals. For adversarial sequences, we demonstrate that no algorithm can achieve a constant competitive ratio. For stochastic arrivals, we present a fast algorithm that guarantees constant regret, using a novel framework based on compensated coupling to prove it.

Finally, we use the Vidur simulator on a public conversation dataset Zheng et al. (2023) to compare our algorithm with parametrized benchmark algorithms on 2 linked A100 GPUs with the Llama-70B model. After optimizing the benchmark parameters, we find that in high-demand scenarios, our algorithm's average latency increases only one-third as fast as the best benchmark, and in low-demand cases, it grows at one-eighth the rate. From a practical perspective, meeting a given average latency requirement in the high-demand setting would require over 8 A100 GPUs for the best benchmark, while our algorithm achieves this with only 2 GPUs, substantially reducing costs and energy use, and promoting sustainability in LLM deployment.

38

38

Tentative Course Schedule

- Introduction to Supervised Learning (1)
- Introduction to Deep Learning (1)
- Reinforcement Learning (6)
- Agentic AI (5)

Note: Tentative schedule subject to changes. See Syllabus and GitHub repo for details.

39

39

Who Are You?

- What is your name?
- Which department are you from?
- Why are you here?
- What do you expect from this course?
- What else do you want me to cover?

40

40