Green AI

Roy Schwartz

Hebrew University of Jerusalem

Workshop on AI and Environment, AI for Good Global Summit, Geneva
July 2025

THE HEBREW
UNIVERSITY
OF JERUSALEM

Scaling AI Models

Scaling AI Models

Gemini

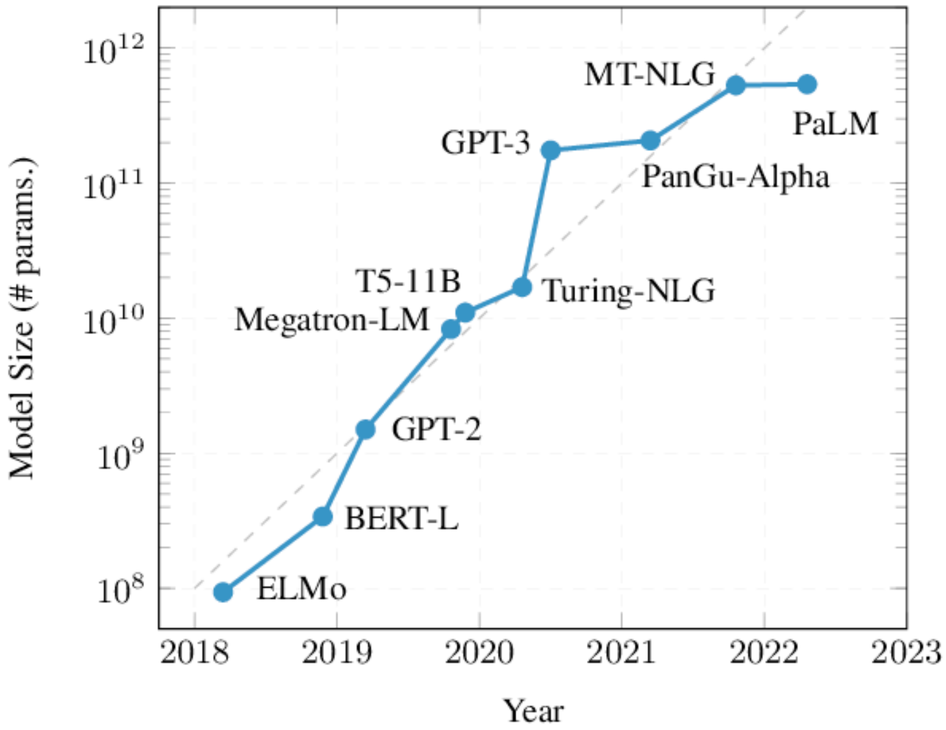
GPT - 4

Meet Claude

A next-generation AI assistant for your tasks, no matter the scale

Jamba

AI21's Hybrid SSM -Transformer Model



Taken from Lakim et al. (2022)

Scaling AI Models

Gemini

GPT - 4

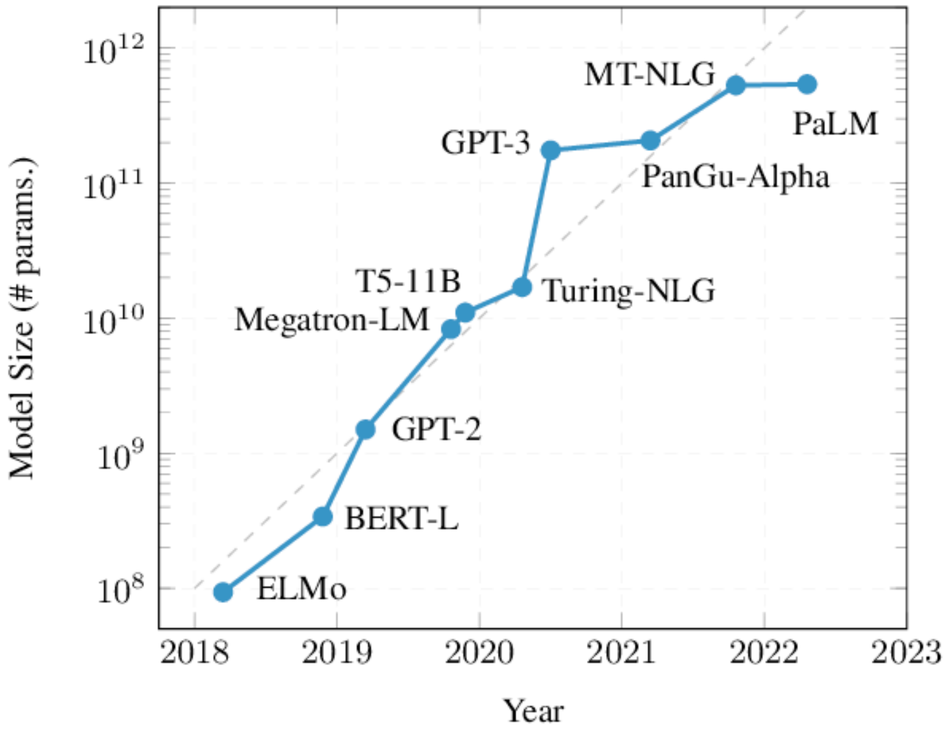
Meet Claude

A next-generation AI assistant for your tasks, no matter the scale

Jamba

AI21's Hybrid SSM -Transformer Model

5,000X in 4 years



Taken from Lakim et al. (2022)

Green AI

Schwartz*, Dodge*, Smith & Etzioni, CACM 2020

Green AI

Schwartz*, Dodge*, Smith & Etzioni, CACM 2020

- Red AI
  - Inclusiveness
  - Environment

Green AI

Schwartz*, Dodge*, Smith & Etzioni, CACM 2020

- Red AI
  - Inclusiveness
  - Environment
- Green AI
  - Enhanced reporting of computational budgets
  - Promote efficiency as a core evaluation for AI

Green AI

Schwartz*, Dodge*, Smith & Etzioni, CACM 2020

- Red AI
  - Inclusiveness
  - Environment
- Green AI
  - Enhanced reporting of computational budgets
  - Promote efficiency as a core evaluation for AI

Exclusive AI:

Training Costs

Exclusive AI:

Training Costs

- BERT (Devlin et al, 2019) was trained on **16** servers for **4** days

Exclusive AI:

Training Costs

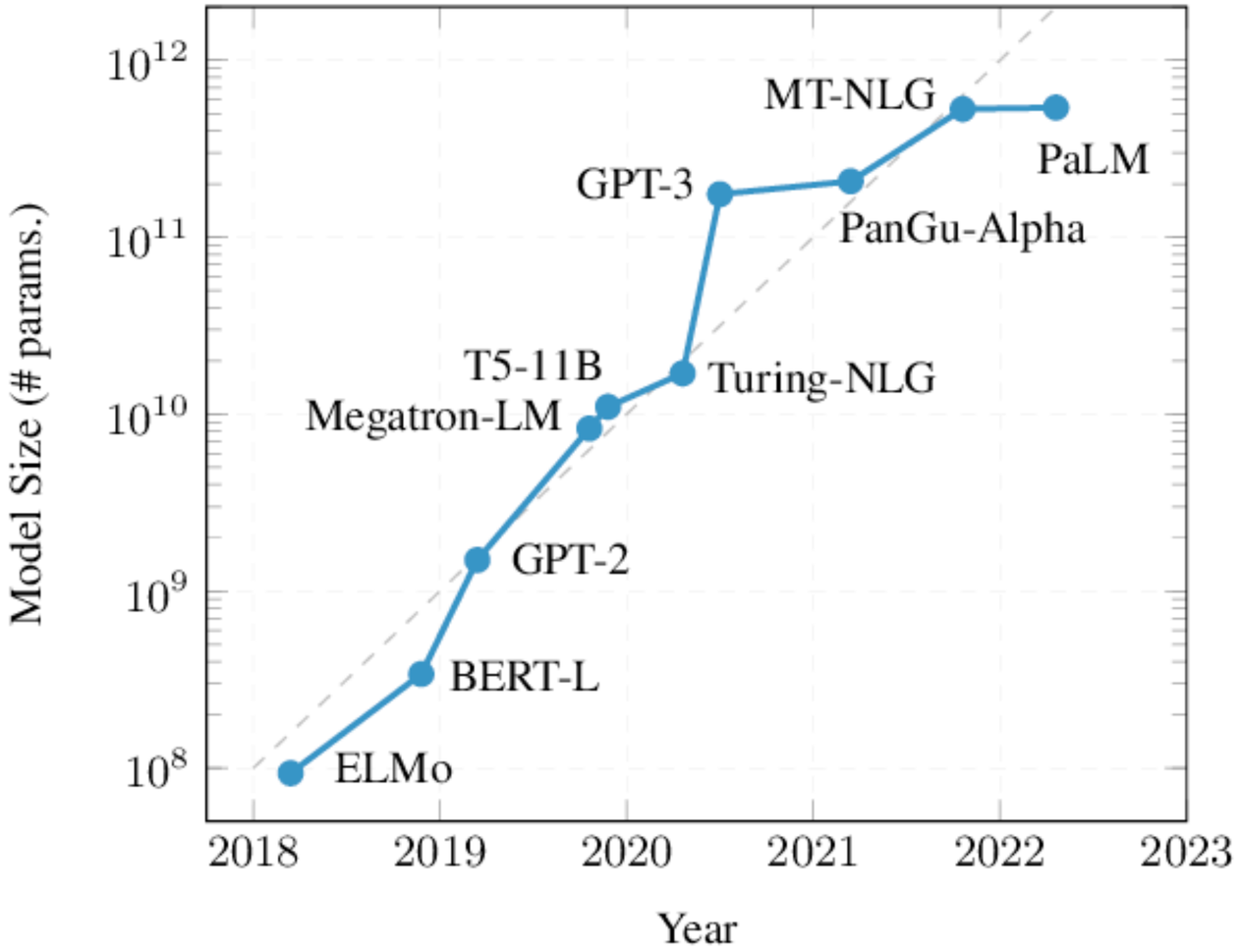
- BERT (Devlin et al, 2019) was trained on **16** servers for **4** days
- RoBERTa (Liu et al., 2019) was trained on **1024** servers for **1** day

Exclusive AI:

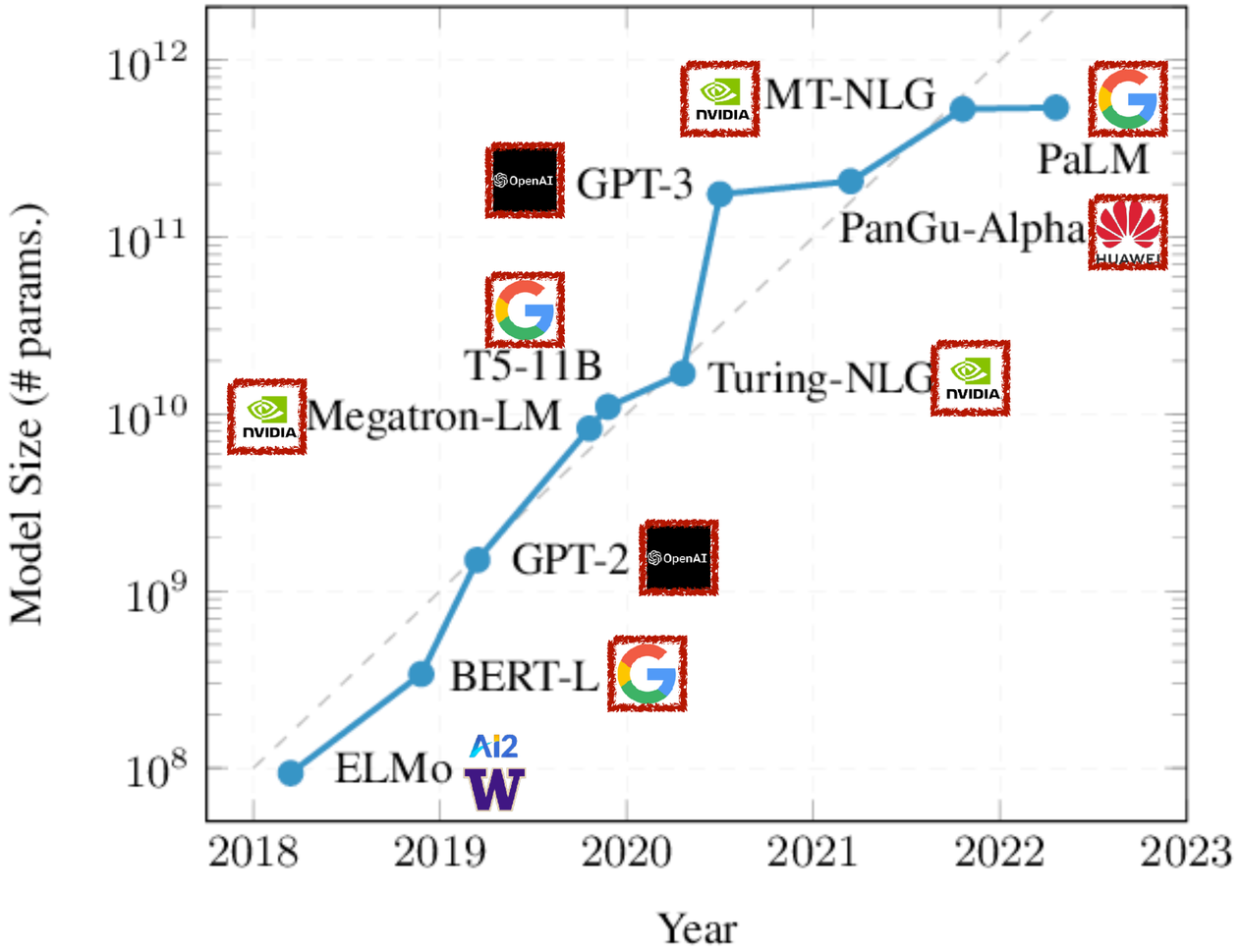
Training Costs

- BERT (Devlin et al, 2019) was trained on **16** servers for **4** days
- RoBERTa (Liu et al., 2019) was trained on **1024** servers for **1** day
- PaLM (Chowdhery et al., 2022) was trained on **10,000** servers for **50** days

It's a Rich Man's World



It's a Rich Man's World



Problems with Scaling

Environment

Consumption	CO_2e (lbs)
Air travel, 1 person, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experiments	78,468
Transformer (big)	192
w/ neural arch. search	626,155

Strubell et al. (2019)

Problems with Scaling

Environment

Consumption	CO_2e (lbs)
Air travel, 1 person, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000

Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experiments	78,468
Transformer (big)	192
w/ neural arch. search	626,155

Strubell et al. (2019)

Is AI really creating an environmental problem?

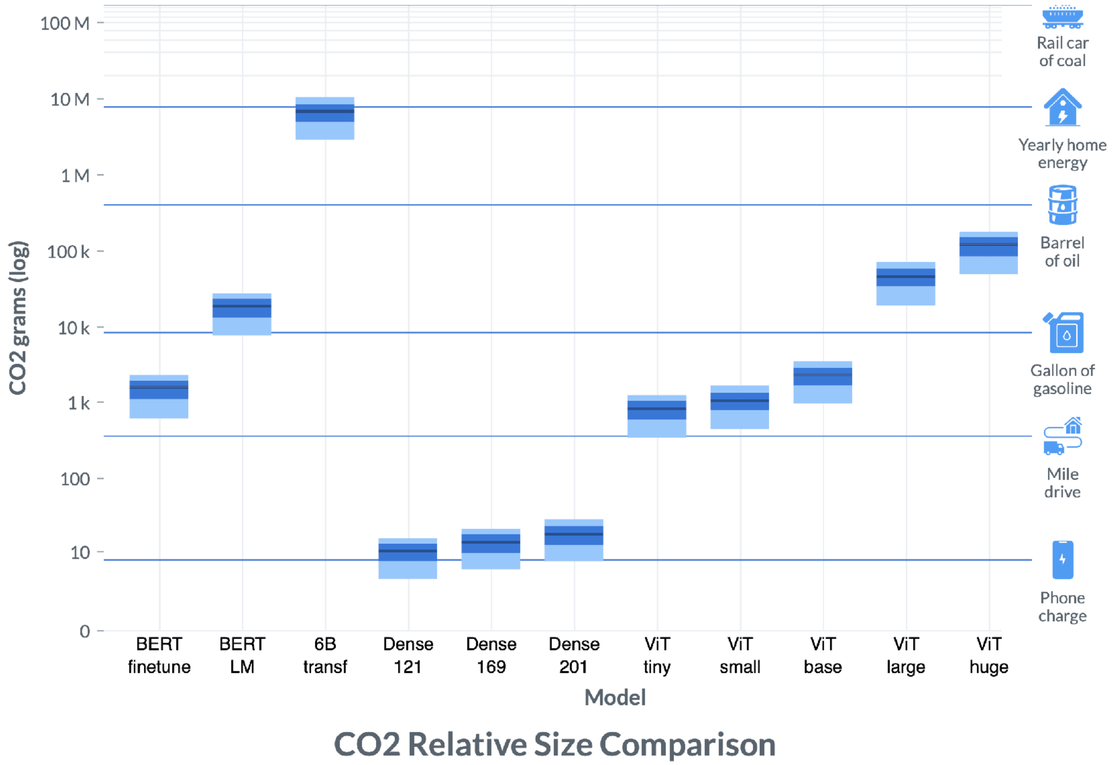
Is AI really creating an environmental problem?

We need better reporting!

Is AI Really Creating an Environmental Problem?

Measuring the Carbon Intensity of AI in Cloud Instances

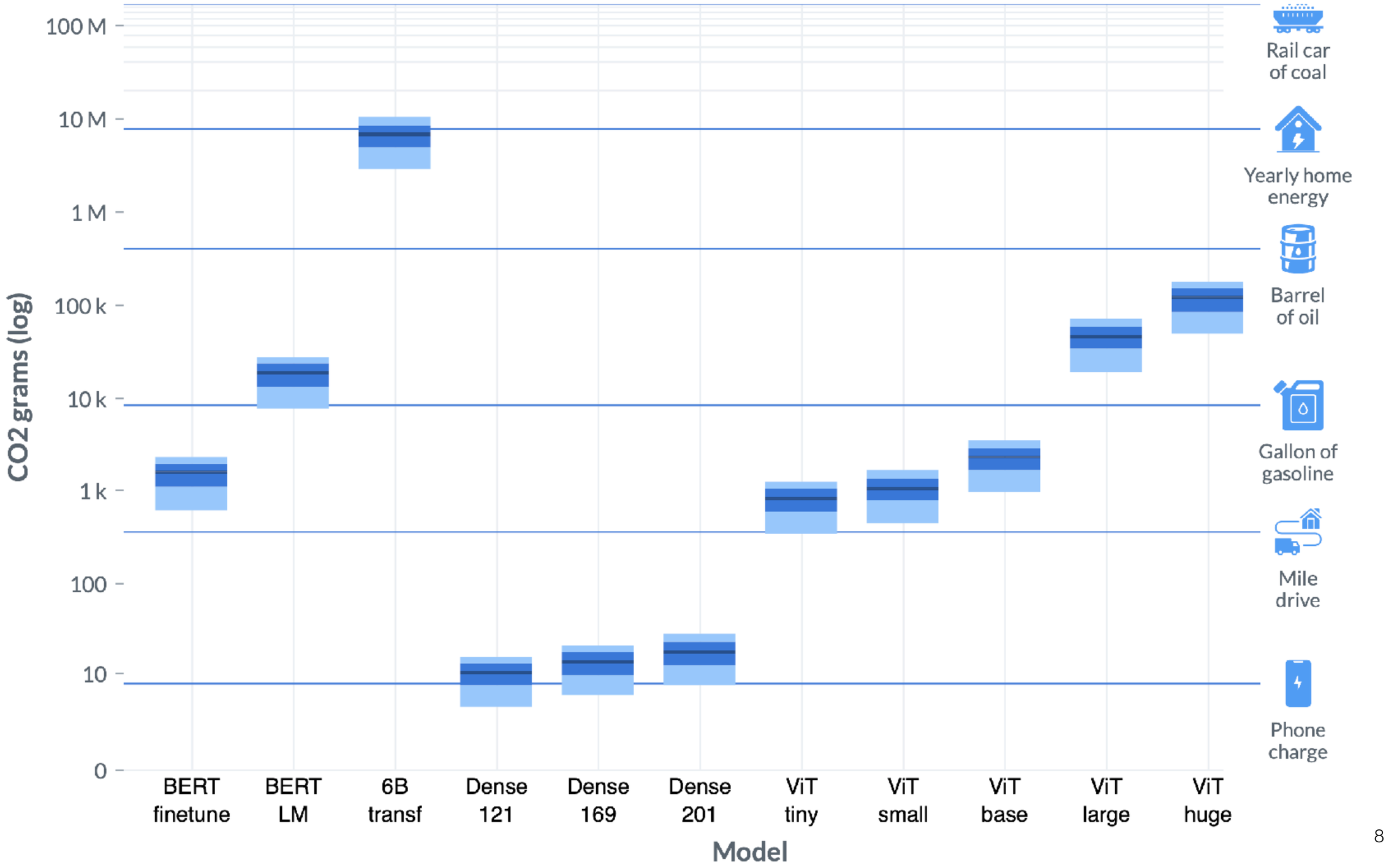
JESSE DODGE, Allen Institute for AI, USA
TAYLOR PREWITT, University of Washington, USA
REMI TACHET DES COMBES, Microsoft Research Montreal, USA
ERIKA ODMARK, Microsoft, USA
ROY SCHWARTZ, Hebrew University of Jerusalem, Israel
EMMA STRUBELL, Carnegie Mellon University, USA
ALEXANDRA SASHA LUCCIONI, Hugging Face, USA
NOAH A. SMITH, Allen Institute for AI and University of Washington, USA
NICOLE DECARIO, Allen Institute for AI, USA
WILL BUCHANAN, Microsoft, USA



CO2 Relative Size Comparison

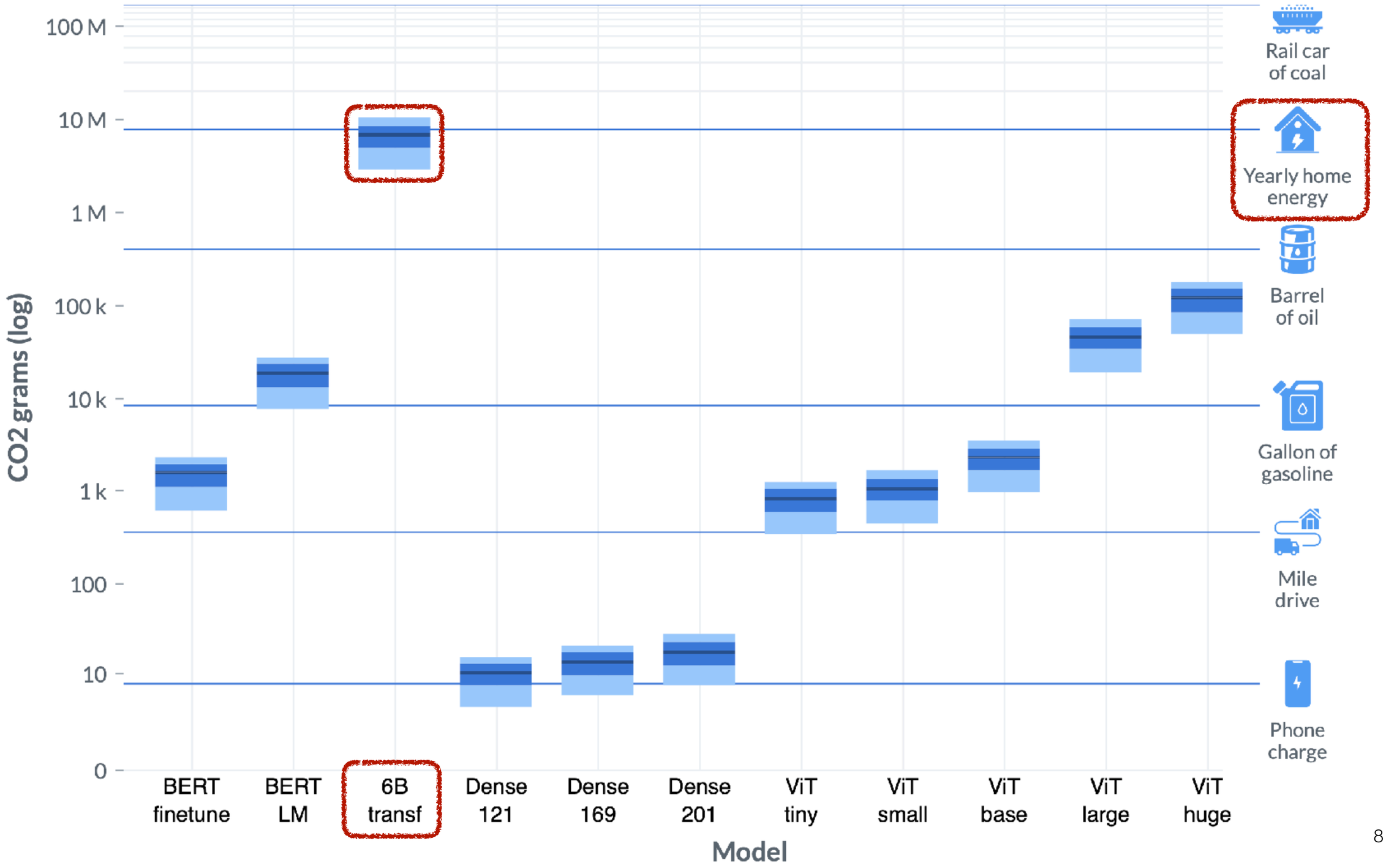
Measu

JESSE D
TAYLO
REMI T
ERIKA
ROY SC
EMMA
ALEXA
NOAH
NICOL
WILL B



Measu

JESSE D
TAYLO
REMI T
ERIKA
ROY SC
EMMA
ALEXA
NOAH
NICOL
WILL B



AI and the Environment

- Evidence around the **most expensive experiments**
  - More recent models consume 2-3 orders of magnitude more CO_2 (Luccioni et al., 2022)
  - These are typically run very few times

AI and the Environment

- Evidence around the **most expensive experiments**
  - More recent models consume 2-3 orders of magnitude more CO_2 (Luccioni et al., 2022)
  - These are typically run very few times
- What about “normal” experiments?
  - Much cheaper, but run hundreds / thousands of times a day?

AI and the Environment

- Evidence around the **most expensive experiments**
  - More recent models consume 2-3 orders of magnitude more CO_2 (Luccioni et al., 2022)
  - These are typically run very few times
- What about “normal” experiments?
  - Much cheaper, but run hundreds / thousands of times a day?
- What about inference operations?
  - Very cheap (though increasingly more expensive; Luccioni et al., 2024)
  - Run billions of times a day?
  - 80-90% of AI computation is spent on inference

AI and the Environment

- Evidence around the **most expensive experiments**
  - More recent models consume 2-3 orders of magnitude more CO_2 (Luccioni et al., 2022)
  - These are typically run very few times
- What about “normal” experiments?
  - Much cheaper, but run hundreds / thousands of times a day?
- What about inference operations?
  - Very cheap (though increasingly more expensive; Luccioni et al., 2024)
  - Run billions of times a day?
  - 80-90% of AI computation is spent on inference

We need better **reporting!**

Overview

- Red AI
  - Inclusiveness
  - Environment
- Green AI
  - Enhanced reporting of computational budgets
  - Promote efficiency as a core evaluation for AI

Overview

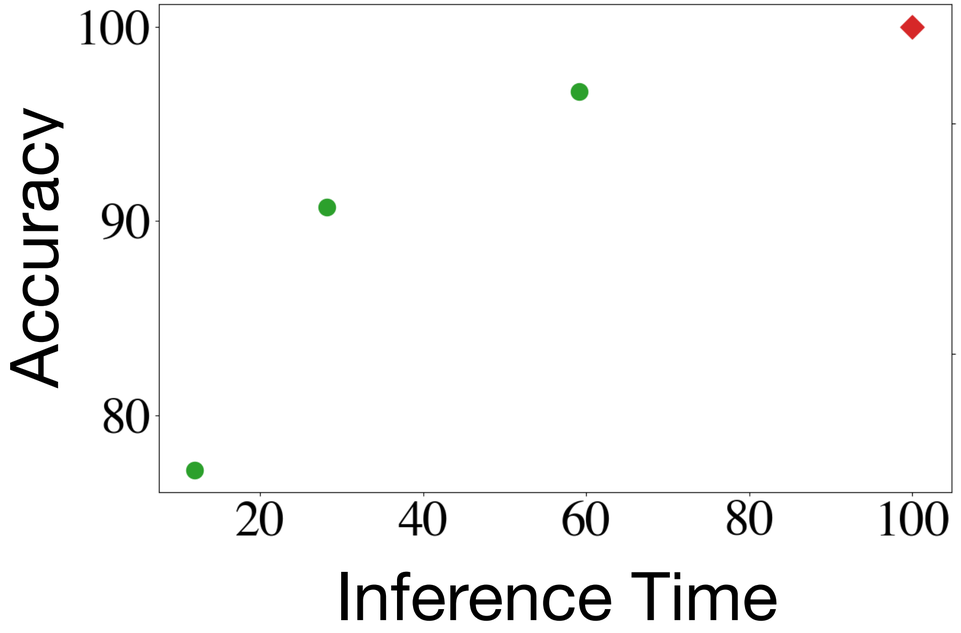
- Red AI
  - Inclusiveness
  - Environment
- Green AI
  - Enhanced reporting of computational budgets
  - Promote efficiency as a core evaluation for AI

Overview

- Red AI
  - Inclusiveness
  - Environment
- Green AI
  - Enhanced reporting of computational budgets
  - Promote efficiency as a core evaluation for AI

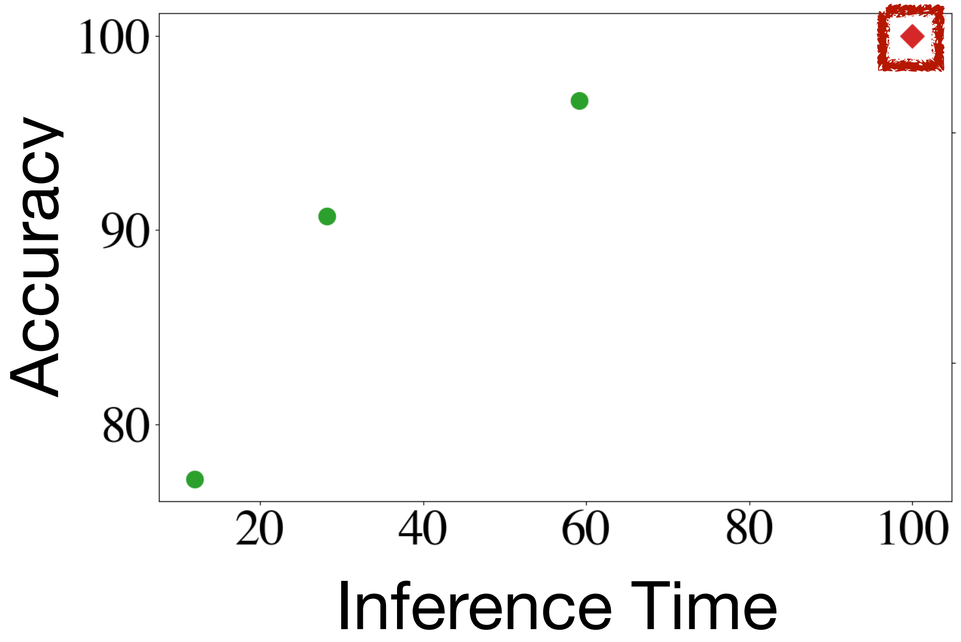
Matching Model and Instance Complexity

S. et al. (2020)



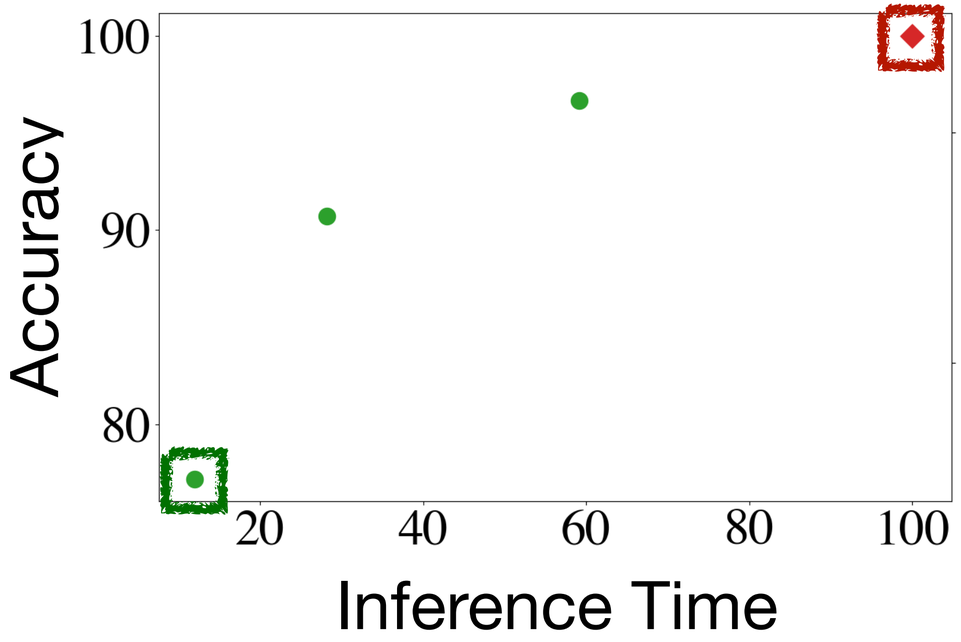
Matching Model and Instance Complexity

S. et al. (2020)



Matching Model and Instance Complexity

S. et al. (2020)

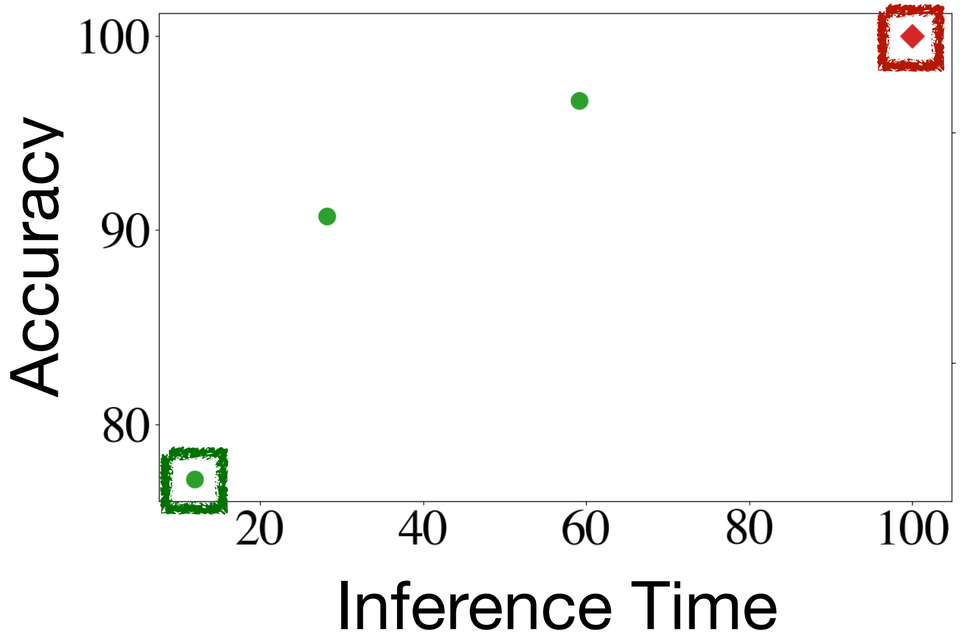


Matching Model and Instance Complexity

S. et al. (2020)

Run an efficient model on "easy" instances,

and an expensive model on "hard" instances

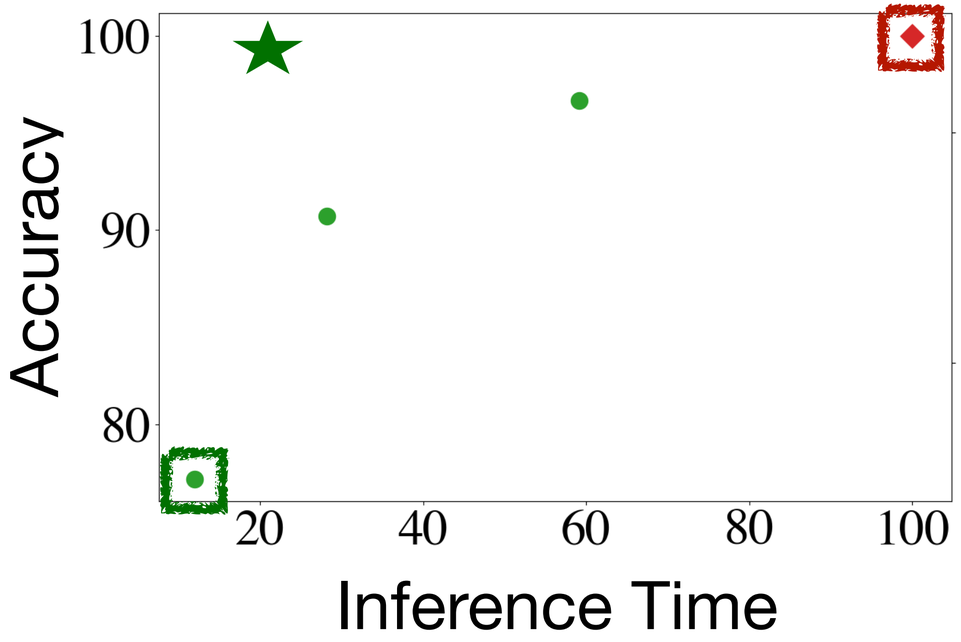


Matching Model and Instance Complexity

S. et al. (2020)

Run an efficient model on "easy" instances,

and an expensive model on "hard" instances

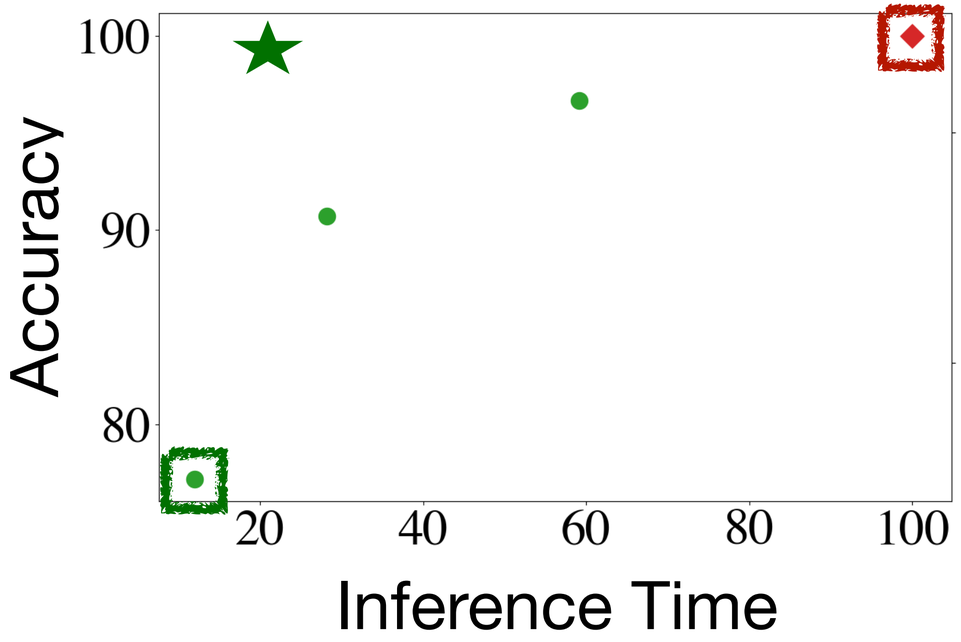


Matching Model and Instance Complexity

S. et al. (2020)

Run an efficient model on "easy" instances,

and an expensive model on "hard" instances



Improved Budget Allocation

Hassid et al. (2024)

Improved Budget Allocation

Hassid et al. (2024)

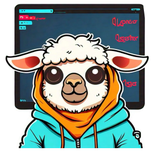
x1
Model Size

Improved Budget Allocation

Hassid et al. (2024)

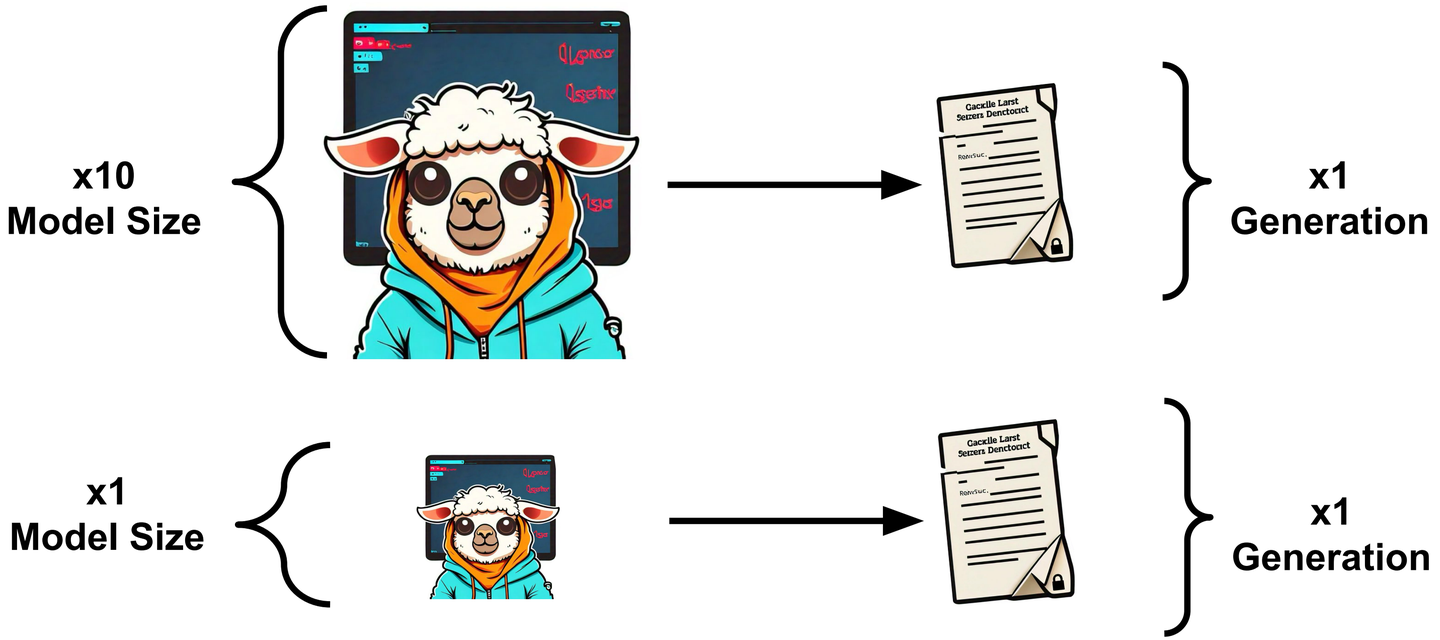
x10
Model Size

x1
Model Size



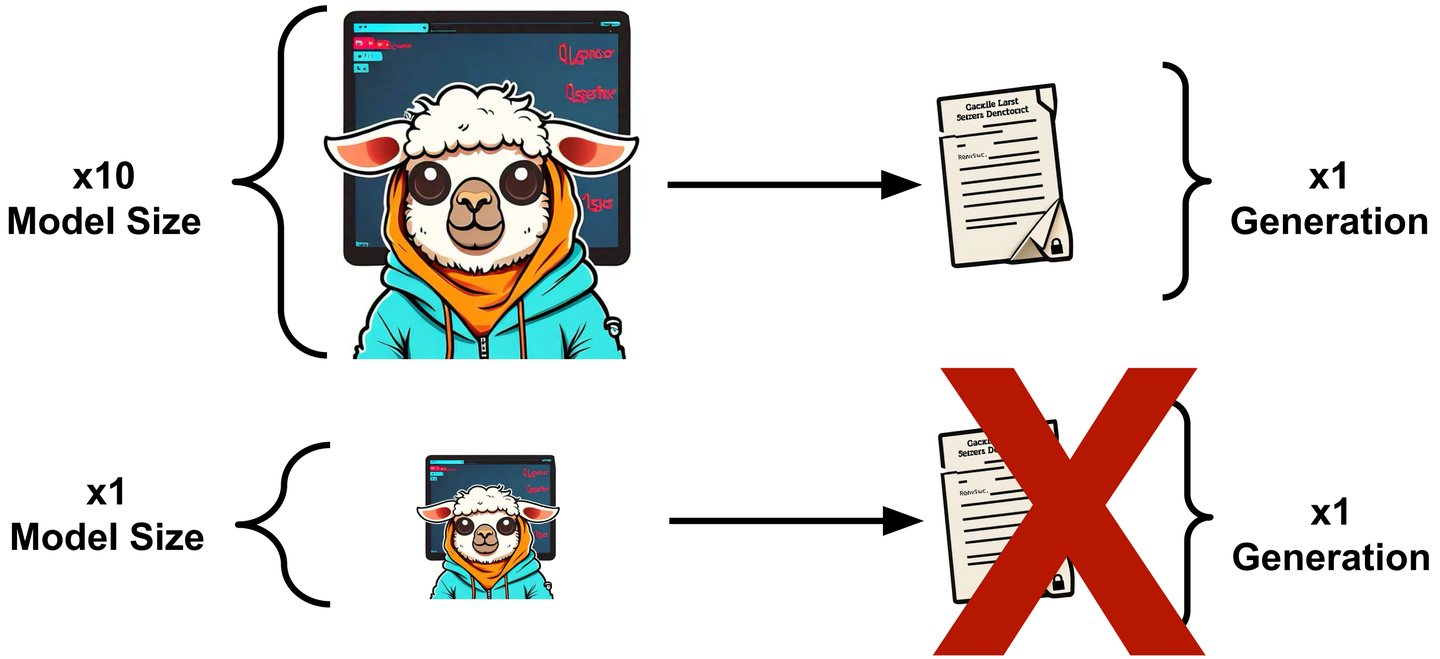
Improved Budget Allocation

Hassid et al. (2024)



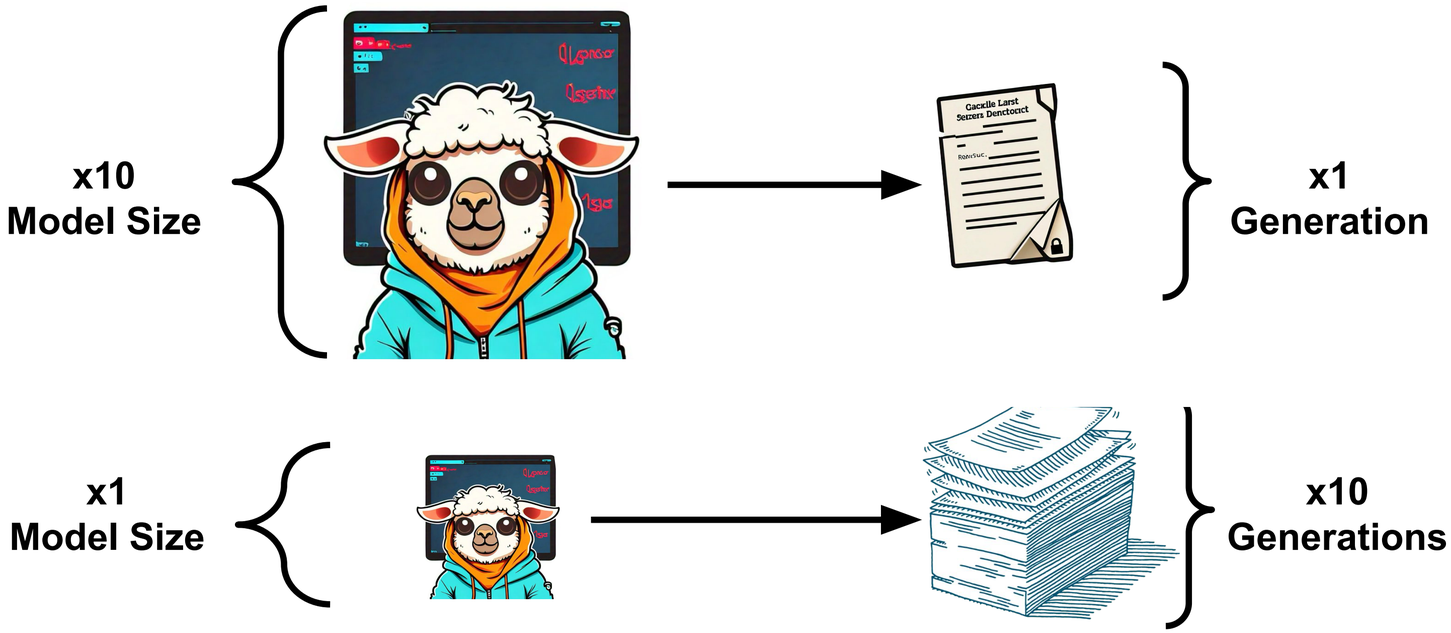
Improved Budget Allocation

Hassid et al. (2024)



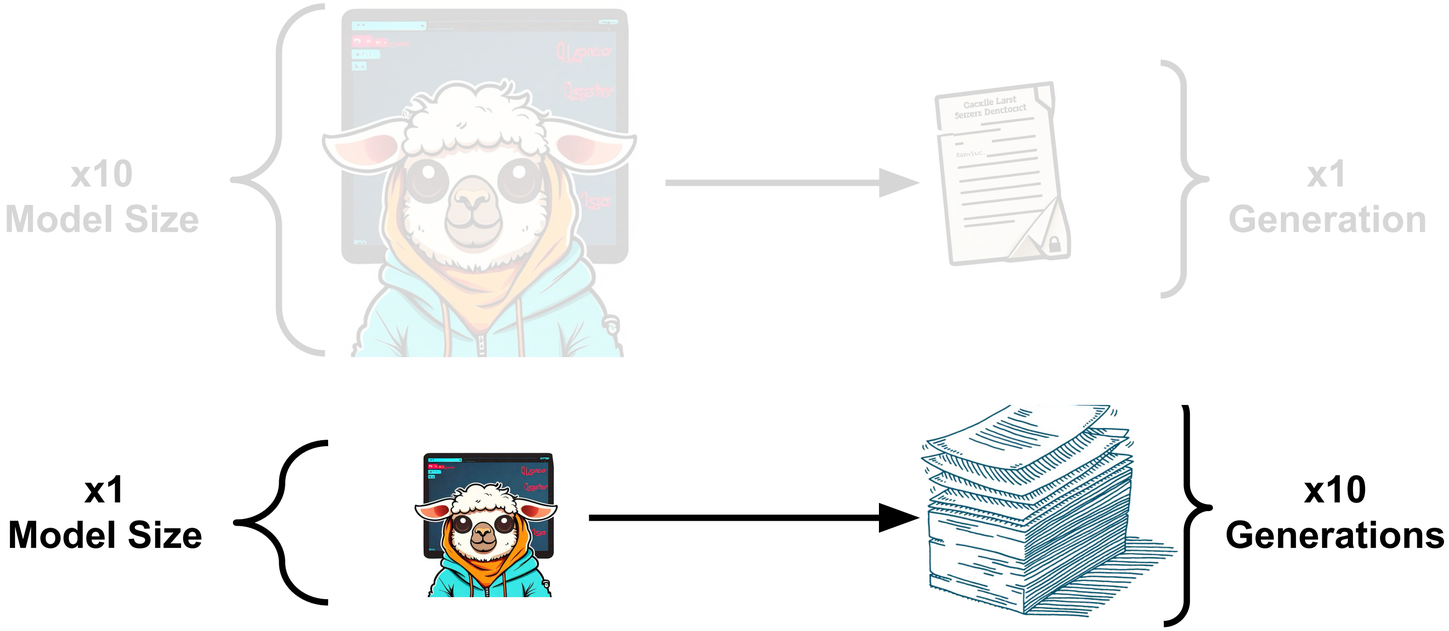
Improved Budget Allocation

Hassid et al. (2024)



Improved Budget Allocation

Hassid et al. (2024)



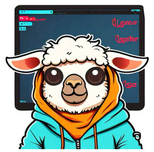
Improved Budget Allocation

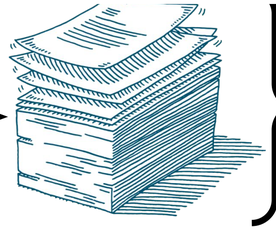
Hassid et al. (2024)

x10
Model Size

x1
Generation

x1
Model Size





x10
Generations

Don’t Overthink it!

Hassid et al. (2025)

Q: Find the sum of all positive integers n such that n+2 divides the product 3(n+3)(n2+9)

Don’t Overthink it!

Hassid et al. (2025)

Q: Find the sum of all positive integers n such that n+2 divides the product 3(n+3)(n2+9)

<think> ..... </think> So the answer is **52**

Don’t Overthink it!

Hassid et al. (2025)

Q: Find the sum of all positive integers n such that n+2 divides the product 3(n+3)(n2+9)

majority@k

<think> ..... </think> So the answer is **52**

<think> ..... </think> So the answer is **49**

<think> ..... </think> So the answer is **33**

<think> ..... </think> So the answer is **52**

Final answer:
52

Don’t Overthink it!

Hassid et al. (2025)

Q: Find the sum of all positive integers n such that n+2 divides the product 3(n+3)(n2+9)

majority@k

<think> ..... </think> So the answer is **52**

<think> ..... </think> So the answer is **49**

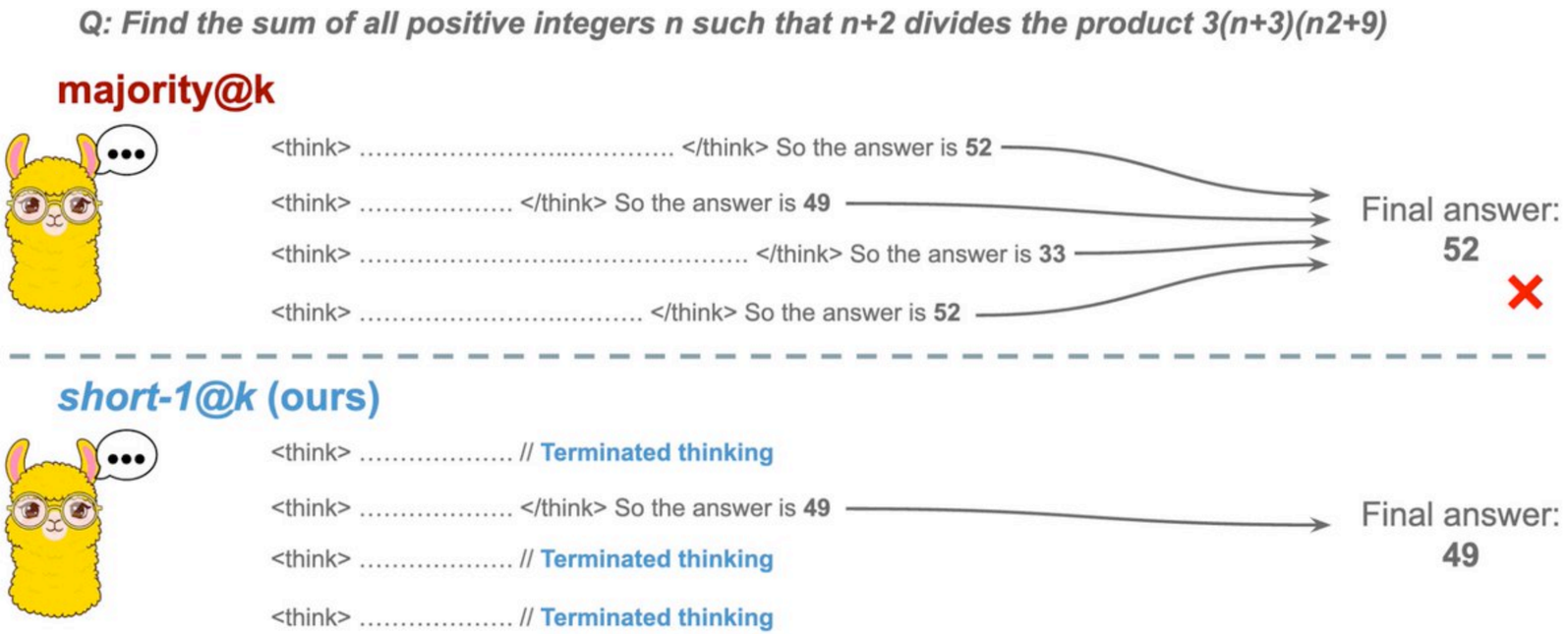
<think> ..... </think> So the answer is **33**

<think> ..... </think> So the answer is **52**

Final answer:
52

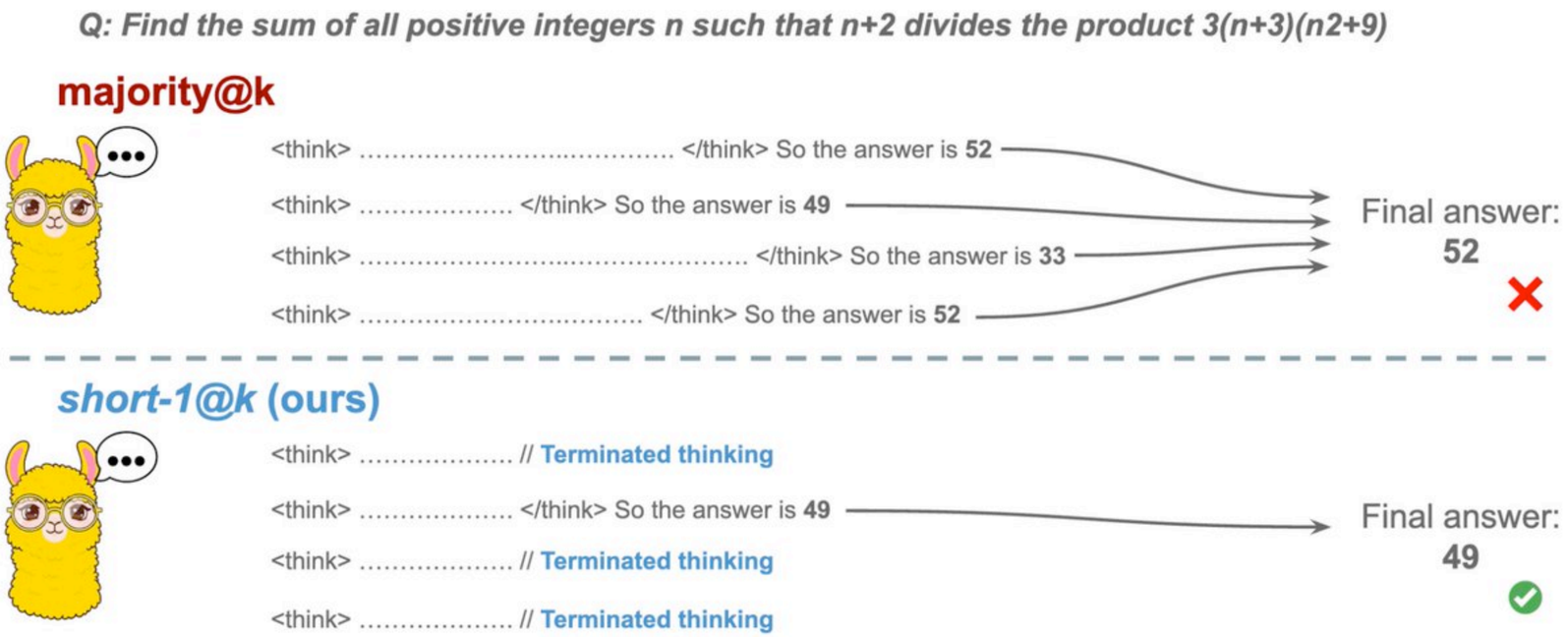
Don’t Overthink it!

Hassid et al. (2025)



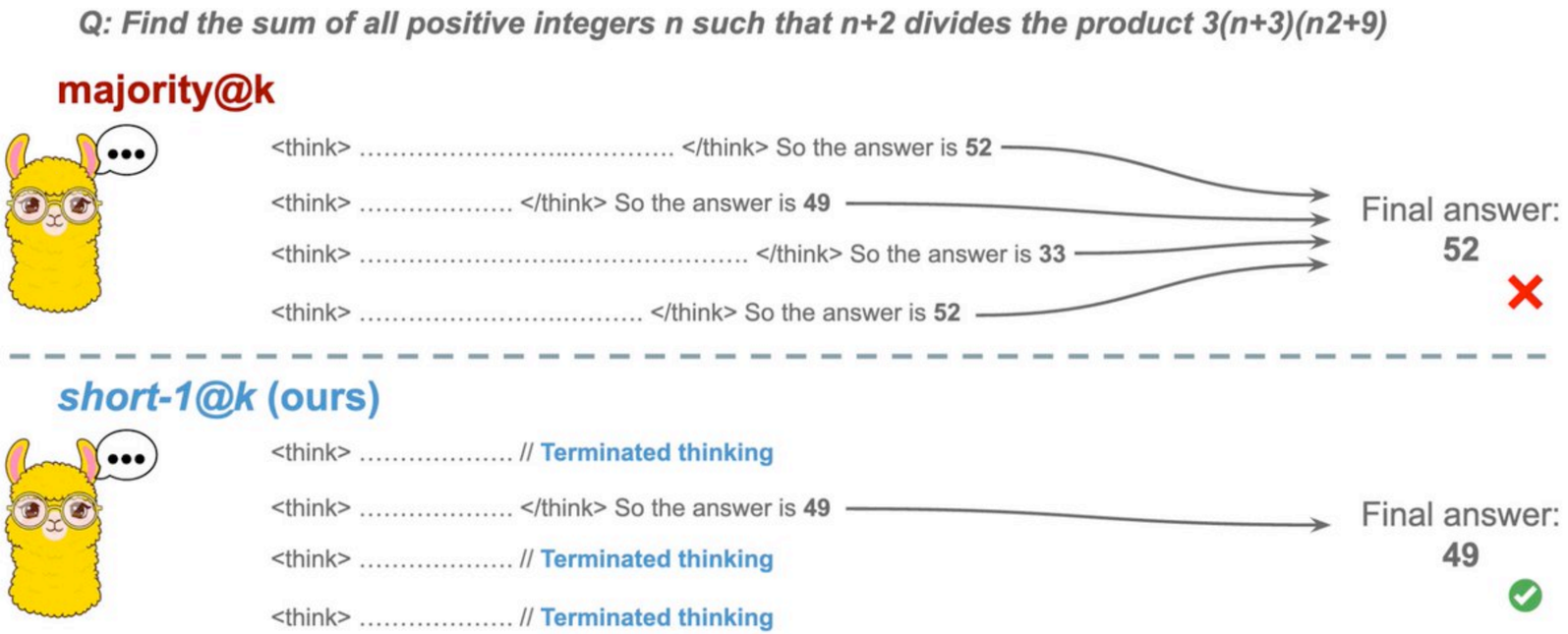
Don’t Overthink it!

Hassid et al. (2025)



Don’t Overthink it!

Hassid et al. (2025)

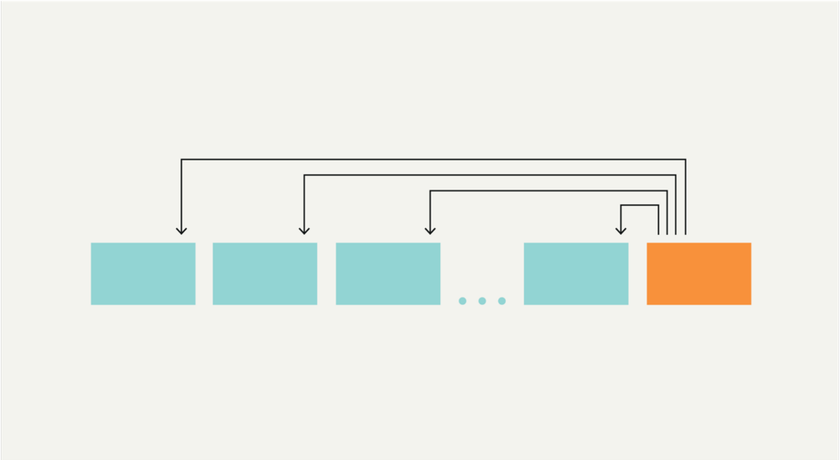


Which Words do we Store in Memory?

Oren et al. (2024)

Which Words do we Store in Memory?

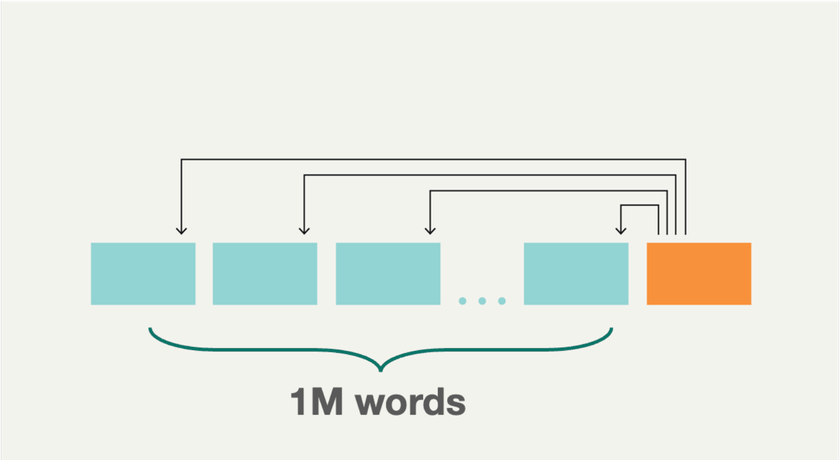
Oren et al. (2024)



Transformer-based LLMs

Which Words do we Store in Memory?

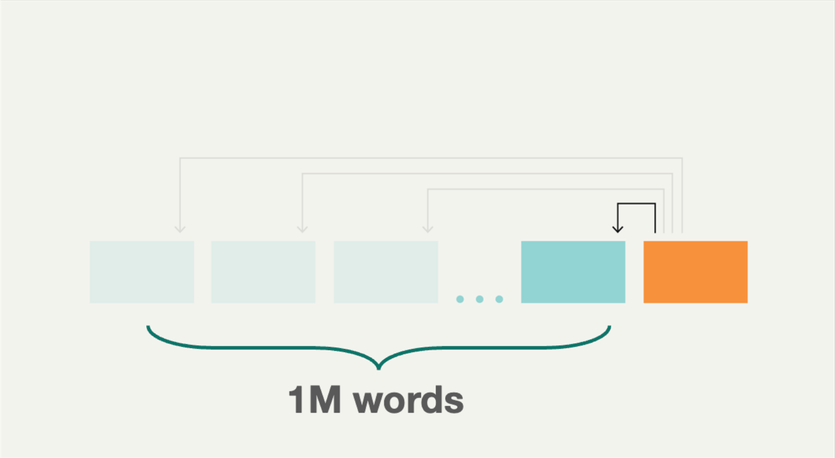
Oren et al. (2024)



Transformer-based LLMs

Which Words do we Store in Memory?

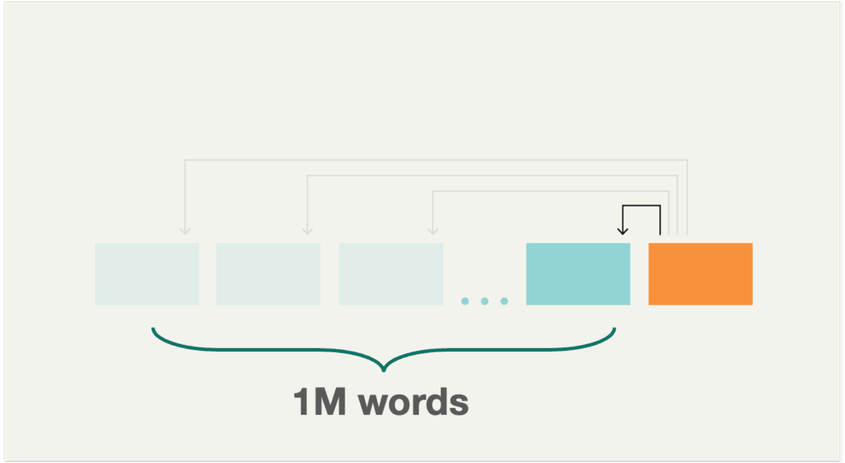
Oren et al. (2024)



Transformer-based LLMs

Which Words do we Store in Memory?

Oren et al. (2024)



Transformer-based LLMs

Summary

THE HEBREW UNIVERSITY OF JERUSALEM

- Red AI
  - Inclusiveness
  - Environment
- Green AI
  - Enhanced reporting of computational budgets
  - Promote efficiency as a core evaluation for AI

Summary

THE HEBREW UNIVERSITY OF JERUSALEM

- Red AI
  - Inclusiveness
  - Environment
- Green AI
  - Enhanced reporting of computational budgets
  - Promote efficiency as a core evaluation for AI