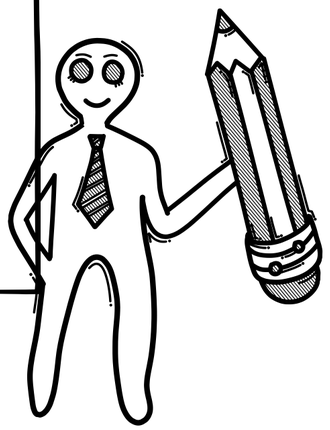
ScaleDown: Shrinking AI's Carbon Footprint, One Token at a Time

Archana

Information is power (literally)

In the race to make bigger models we implore you to look the other way



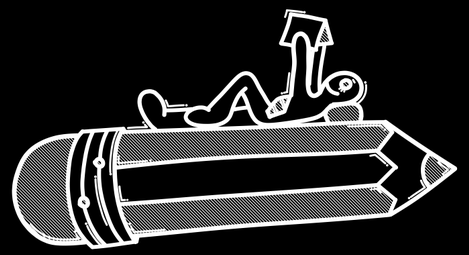
Introduction

Archana- Program Manager@Apart Research, Board Member at WIML, Fellow at Python Software Foundation

Background in Electrical and Electronics Engineering, TinyML and Masters in Sustainable Energy Solutions in Power Systems

Archana

extension.scaledown.ai



Today's Agenda

1. AI Inefficiencies Around Us
   Exploring everyday AI usage patterns

2. Understanding Carbon Footprint of AI: Factors affecting AI's carbon footprint

3. Prompt Templates and Python Package: Look into how the package was built

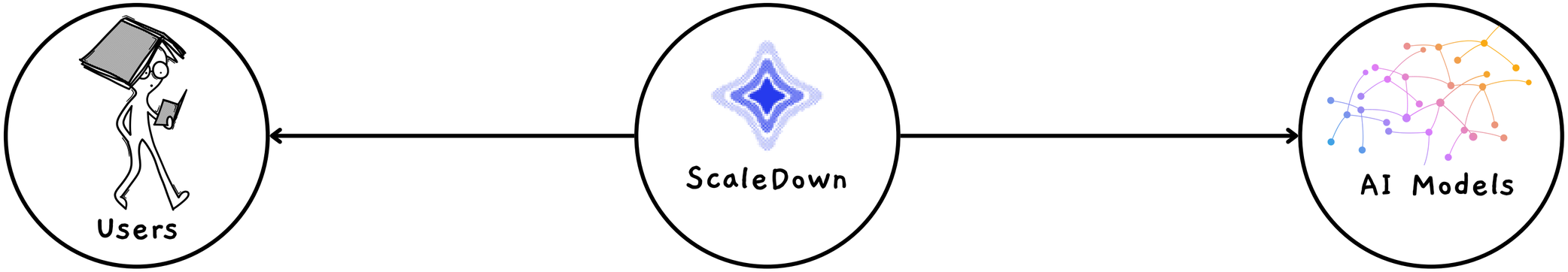
4. QnA and Ways to Contribute

"Information-->Action"

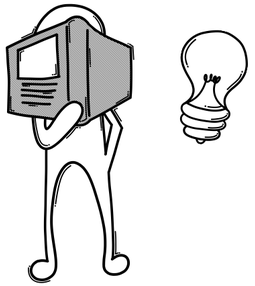
Where we want to take Carbon ScaleDown

extension.scaledown.ai

Track and Monitor AI Usage, Reduce Costs, Minimize Carbon Footprint



ScaleDown sits between users and AI model providers. Automatically tracking and optimizing your AI usage.



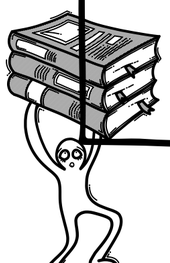
So what does Carbon ScaleDown Do?

extension.scaledown.ai

Write Prompts, Optimise tokens and Minimize Carbon Footprint

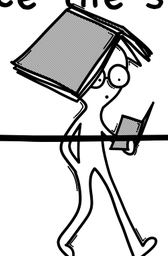
Write Prompts and Choose Styles

- Get domain-specific prompts for specialized fields
- Choose from different response styles (concise, detailed, creative)
- Access a library of community-contributed prompt templates



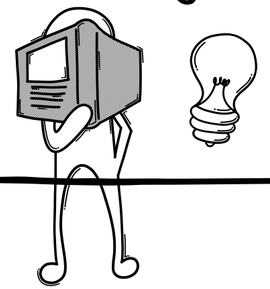
Optimise AI prompts and reduce Tokens

- Remove unnecessary phrases while preserving meaning
- Apply best practices for each AI model (Claude, GPT, Llama)
- Verify that optimized prompts produce the same results



Minimize Carbon Footprint

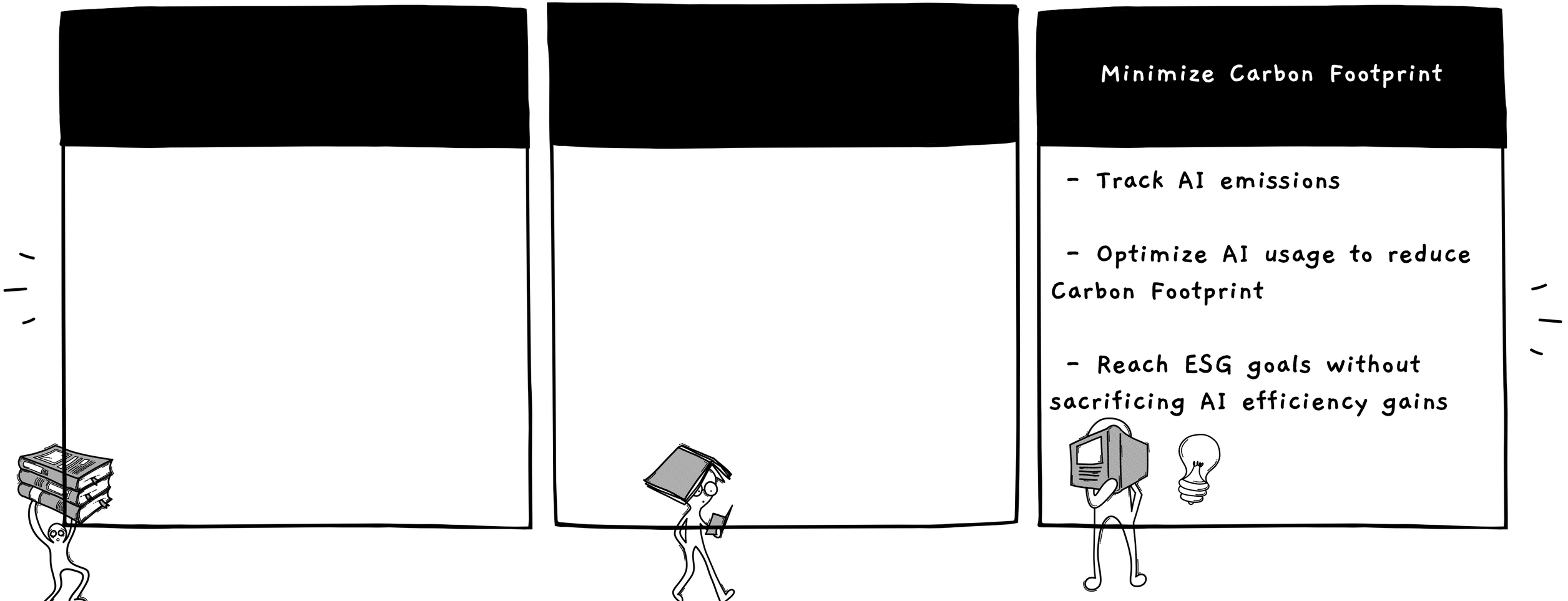
- Track AI emissions
- Optimize AI usage to reduce Carbon Footprint
- Reach ESG goals without sacrificing AI efficiency gains



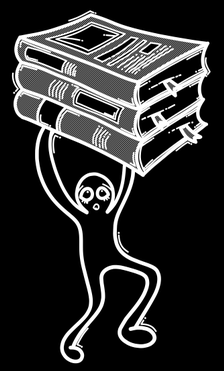
So what does Carbon ScaleDown Do?

extension.scaledown.ai

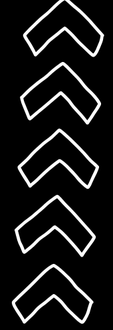
Write Prompts, Optimise tokens and Minimize Carbon Footprint

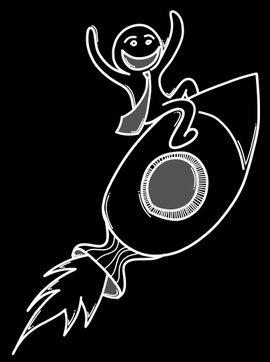


extension.scaledown.ai

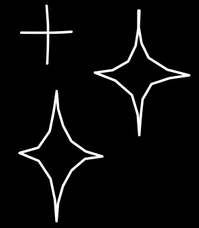


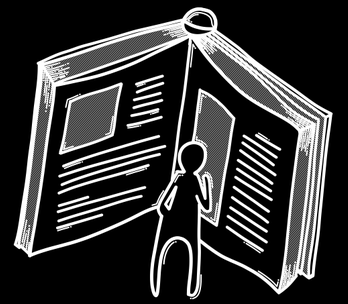
Understanding AI inefficiency through Quizzes





Are you ready?

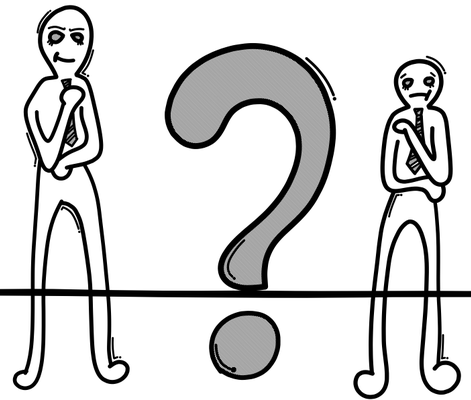




How many of you have used ChatGPT or similar AI tools in the last 24 hours?

For those who have, what was the primary purpose?

1. Work, 
2. personal research, or 
3. entertainment? 

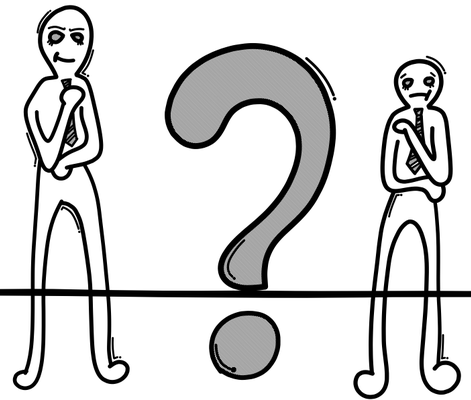


extension.scaledown.ai

How many of you have used ChatGPT or similar AI tools in the last 24 hours?

For those who have, what was the primary purpose?

1. Work, 
2. personal research, or 
3. entertainment? 



A recent study by AI Insights Today shows that 78% of regular AI users interact with these tools at least 3 times a day.

extension.scaledown.ai

How many of you have used ChatGPT or similar AI tools in the last 24 hours?

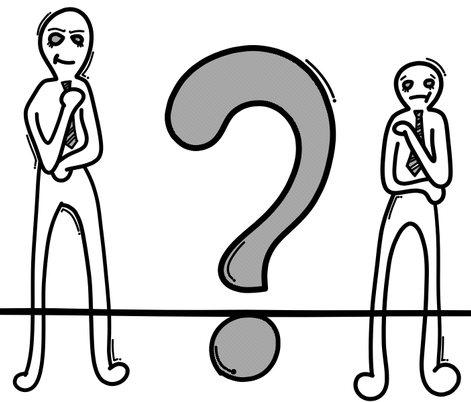
When using AI tools, how often do you find yourself rephrasing or refining your initial query?"

a) Rarely (less than 10% of the time)

b) Sometimes (10-30% of the time)

c) Often (30-50% of the time)

d) Very frequently (more than 50% of the time)



extension.scaledown.ai

How many of you have used ChatGPT or similar AI tools in the last 24 hours?

When using AI tools, how often do you find yourself rephrasing or refining your initial query?"

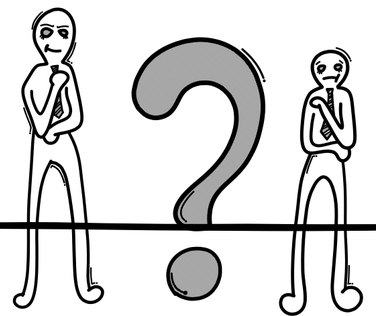
a) Rarely (less than 10% of the time)

b) Sometimes (10-30% of the time)

c) Often (30-50% of the time)

d) Very frequently (more than 50% of the time)

According to a 2023 survey by TechTrends, users spend an average of 18 minutes per day refining AI queries, which translates to nearly 110 hours per year



extension.scaledown.ai

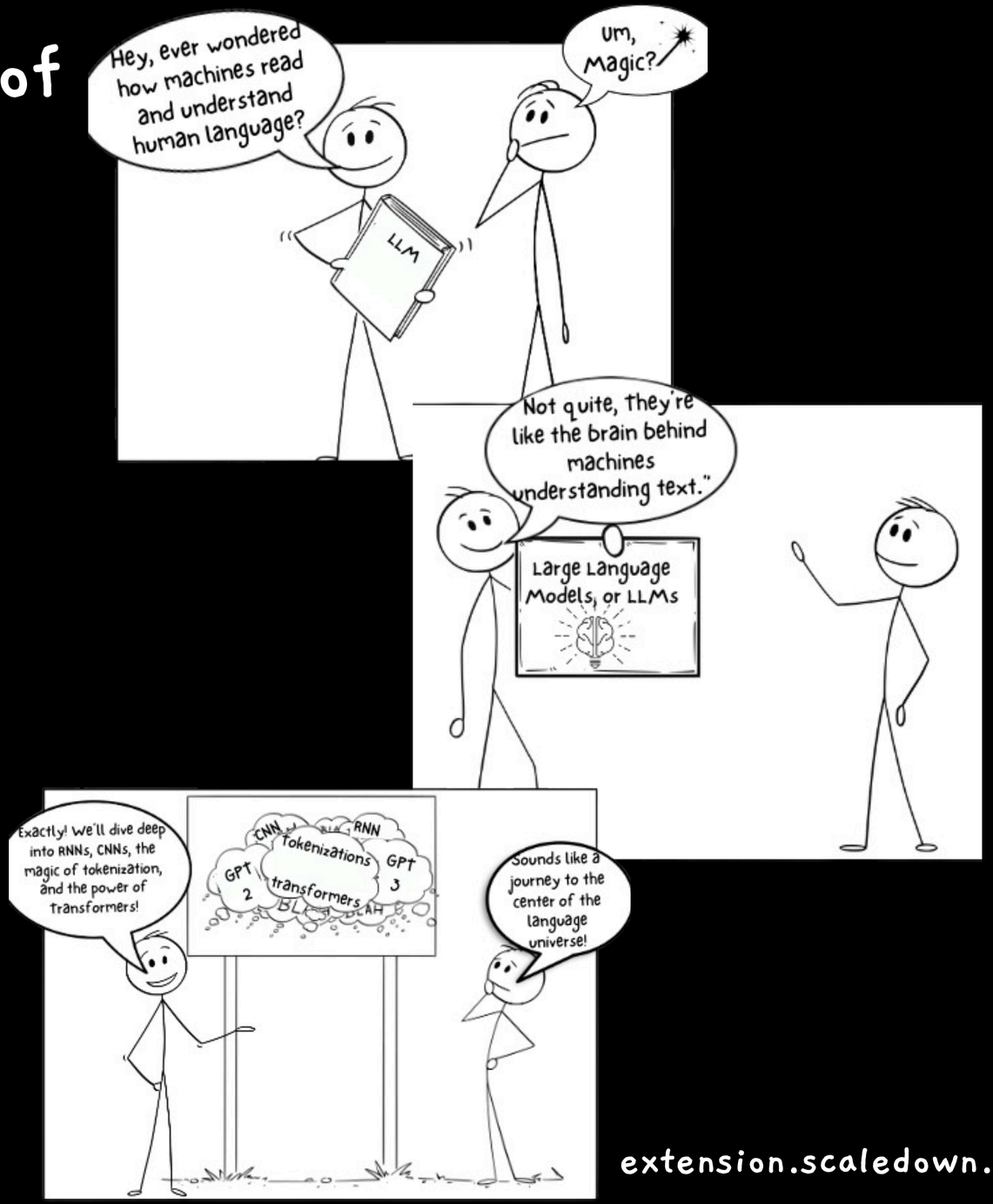
Tokens: The Building Blocks of AI Language Models

What are Tokens?

- Tokens are the basic units that AI language models process
- They can be words, parts of words, or even punctuation marks
- Example: "I love AI!" = 4 tokens (I / love / AI / !)
- More complex: "unprecedented" = 3 tokens (un / precede / nted)

Token Usage in AI Models:

- Every interaction with an AI model involves tokens
- Input (your query) and output (AI's response) are both measured in tokens
- Models have token limits (e.g., GPT-3 can handle up to 4,096 tokens per interaction)



extension.scaledown.ai

Tokens: The Building Blocks of AI Language Models

extension.scaledown.ai

How AI Models Process Tokens:

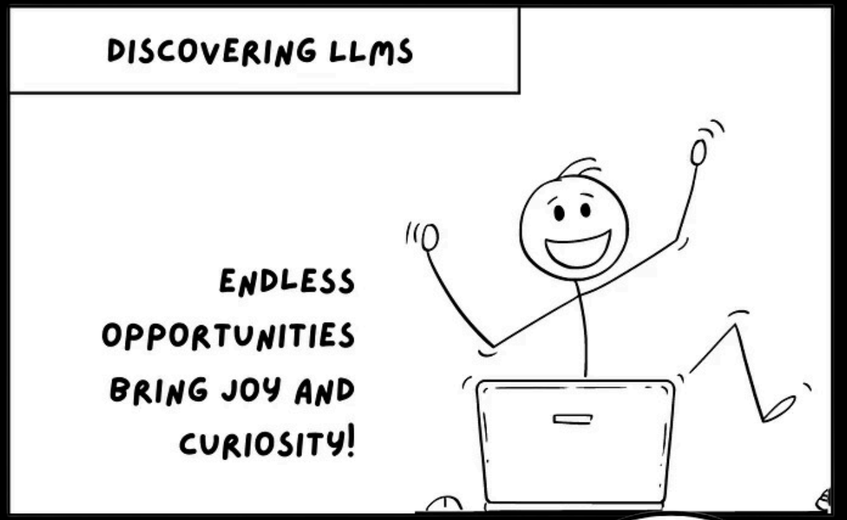
- Language models predict the next token based on the previous ones
- This process is repeated for each token in the sequence
- Example: "The cat sat on the..." (model predicts "mat" as the next likely token)

Token Usage and Model Performance:

- More tokens generally lead to better understanding and more detailed responses
- But also increase computational load and energy consumption

Optimizing Token Usage:

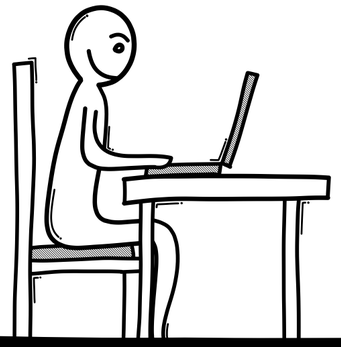
- Efficient queries use fewer tokens to get the same information
- Reducing unnecessary tokens saves energy and reduces carbon footprint



Let's do a quick experiment. Everyone think of a complex question you've asked an AI in the past week. Now, imagine you're explaining it to a 5-year-old. How would you simplify it?

Original Prompt: I'm writing a comprehensive research paper on the impact of climate change on global agriculture. Can you provide a detailed analysis of how rising temperatures, changing precipitation patterns, and extreme weather events are affecting crop yields, soil health, and food security worldwide? Please include specific examples from different regions, potential future scenarios, and possible mitigation strategies for farmers and policymakers.

125 tokens



57 tokens

Summarize climate change effects on global agriculture. Include:

1. Impact on crop yields
2. Regional examples
3. Future scenarios
4. Mitigation strategies for farmers and policymakers

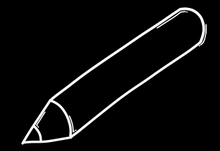
extension.scaledown.ai

The Need for Efficiency: Sarah's AI Adventure

Sarah, a freelance content creator, relies heavily on AI tools for research, writing, and idea generation. Let's follow her typical workday:

1

9:00 AM: Sarah starts her day by asking ChatGPT for content ideas. She spends 20 minutes refining her prompts to get the perfect pitch ideas.

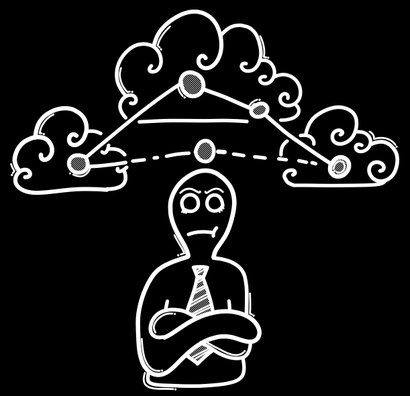


2

11:00 AM: While writing, she frequently consults the AI for fact-checking and synonyms. Each query takes about 30 seconds, and she makes about 40 queries in two hours.

3

2:00 PM: Sarah uses an AI image generator for creating visual content. She spends 45 minutes describing and refining her prompts to get the right image.



4

4:00 PM: For her podcast script, Sarah uses an AI tool to generate an outline. She spends 30 minutes tweaking the prompts to get a satisfactory structure.

extension.scaledown.ai

Think about your own workflow. How does it compare to Sarah's?

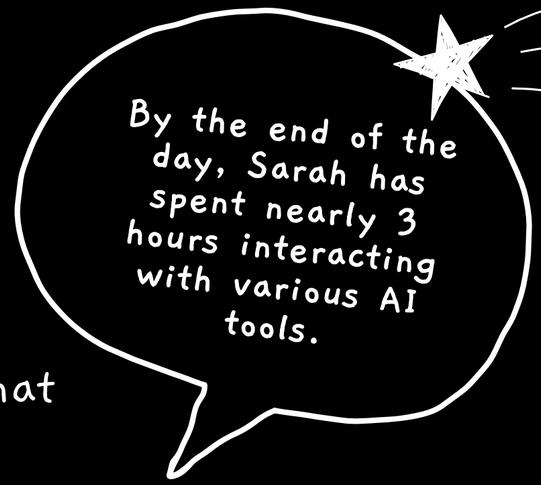
"If you could reduce your AI interaction time by 40% while achieving the same results, how would you use that saved time?"

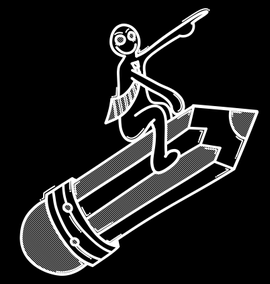
a) Take on more projects

b) Improve work-life balance

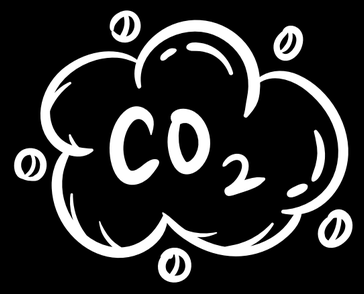
c) Learn new skills

d) Increase creativity in your work

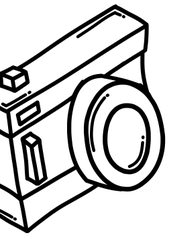




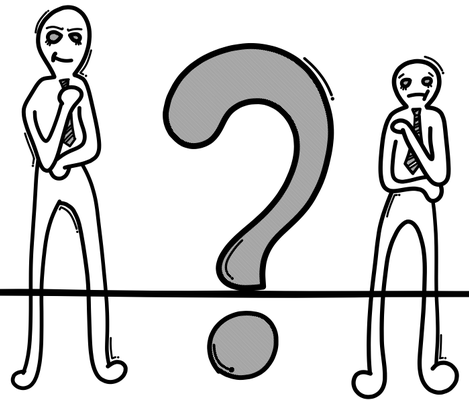
What is the Carbon Footprint of your AI Usage?

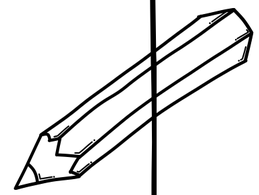


extension.scaledown.ai



How many of you have used an AI tool like ChatGPT or Claude instead of a Google Query?



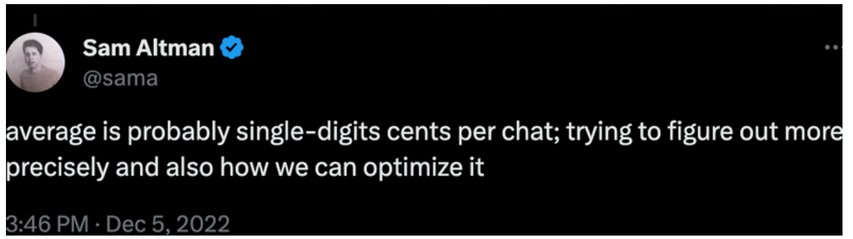


extension.scaledown.ai

GPT vs Google Search

How much energy consumption is involved in Chat GPT responses being generated?

Asked 10 months ago Modified 1 month ago Viewed 15k times

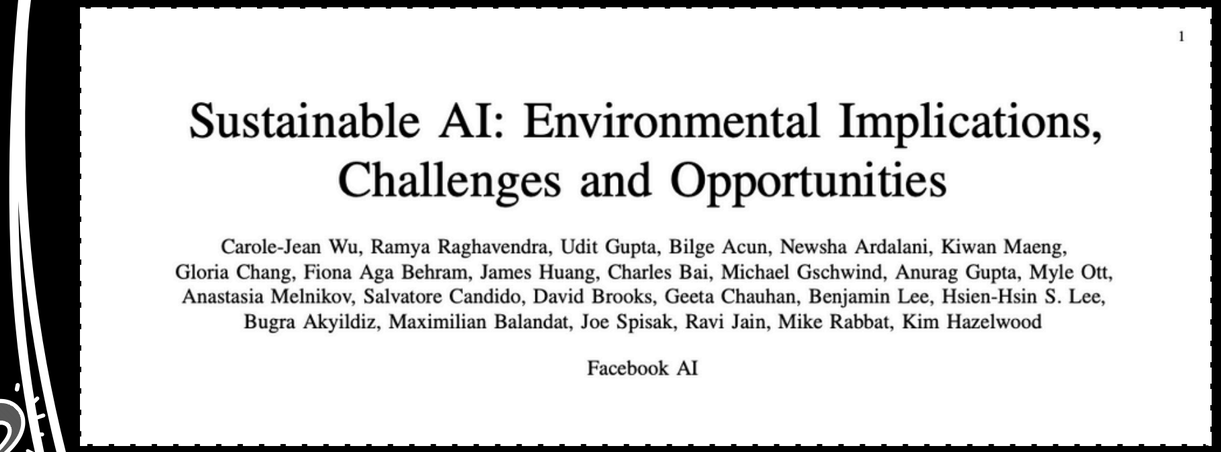


- With context to GPT most people are starting to use ChatGPT as a search engine, and energy consumption per query is could mean a 0,3 kWh per request, versus a 0.0003 kWh per Google search.So GPT-3's energy consumption is 1000x more than a simple Google search

- How did they come to this conclusion?: According to Sam Altman, a single prompt costs"probably single-digits cents" thus worst case 0,09€/request.
  - a least half the cost are energy at a cost of 0,15€/1kWh,
  - a request would cost 0,09€/request*50%/0,15€/1kW=0,3kWh/request = 300Wh per request.
  - 60 Smartphone charges of 5Wh per Charge

Source: Stack Exchange

Paper 1: Life cycle Analysis for LM



1

Sustainable AI: Environmental Implications, Challenges and Opportunities

Carole-Jean Wu, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, Fiona Aga Behram, James Huang, Charles Bai, Michael Gschwind, Anurag Gupta, Myle Ott, Anastasia Melnikov, Salvatore Candido, David Brooks, Geeta Chauhan, Benjamin Lee, Hsien-Hsin S. Lee, Bugra Akyildiz, Maximilian Balandat, Joe Spisak, Ravi Jain, Mike Rabbat, Kim Hazelwood

Facebook AI

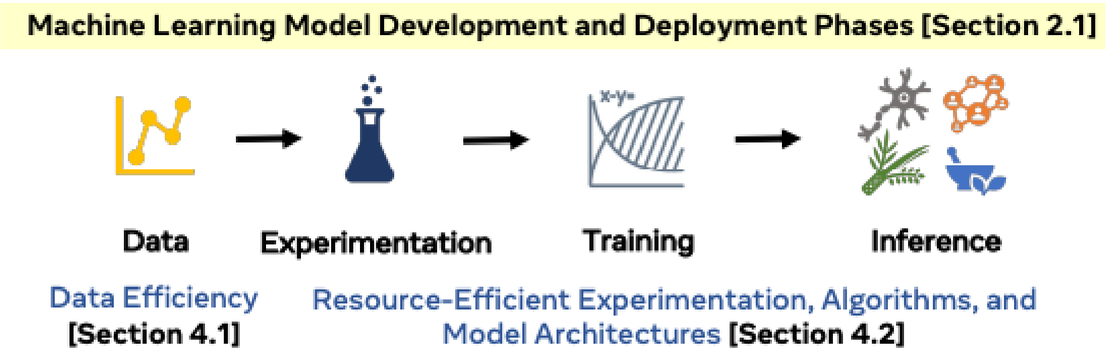
1. Lifecycle of a Typical Model in comparison to product lifecycle for estimating carbon footprint
2. Training vs Inference for Carbon Footprint
3. GPT vs Google Search

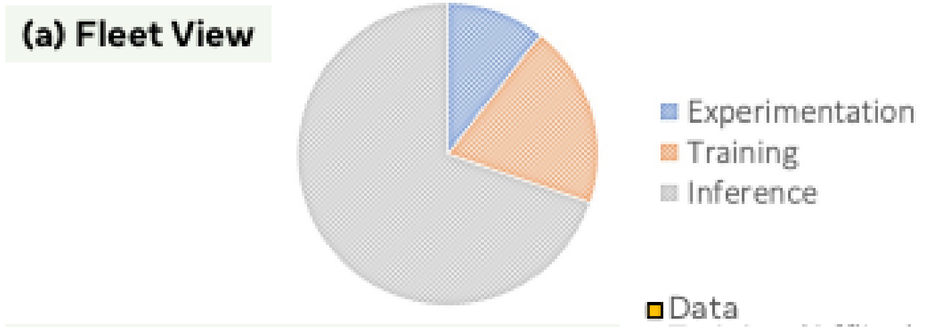
extension.scaledown.ai

- Life Cycle Analysis (LCA) is a common methodology to assess the carbon emissions over the product life cycle. There are four major phases: manufacturing, transport, product use,and recycling

- From the perspective of AI's carbon footprint analysis, manufacturing and product use are the focus

- At Facebook, power capacity breakdown of 10:20:70 for AI infrastructures devoted to the three key phases — Experimentation, Training, and Inference; Inference being the highest





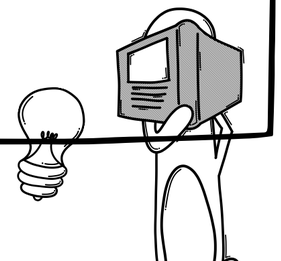
Training vs Inference

Training

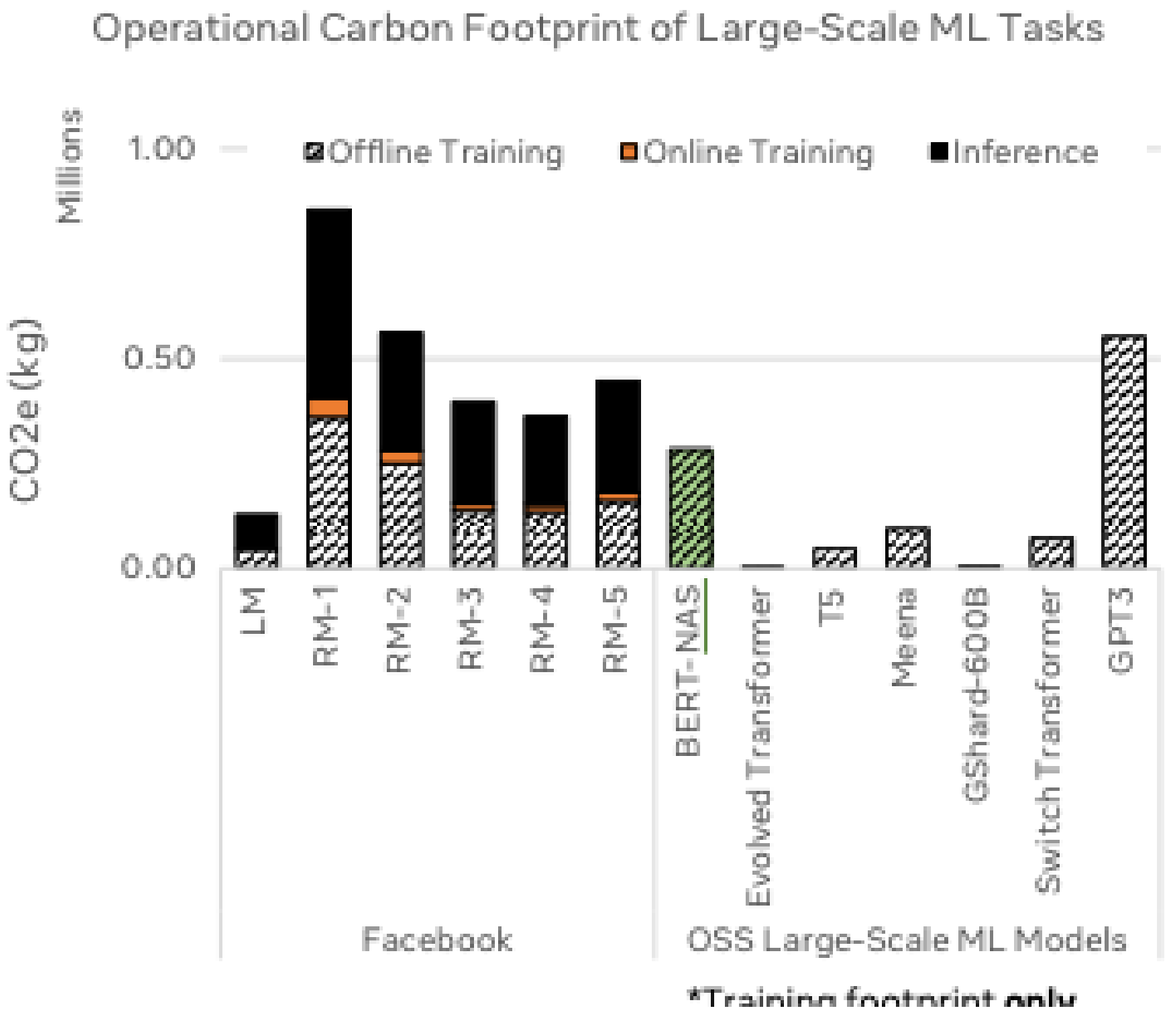
- Role of Data in Training: Training utilizes recent, extensive production data
- Training Frequency: Models are trained at varying frequencies depending on the use-case requirement.
- AI use cases at Facebook has driven 2.9× increase in AI training infrastructure capacity over the 1.5 years

Inference

- The best-performing model is deployed, producing trillions of daily predictions to serve billions of users worldwide.
- The total compute cycles for inference predictions are expected to exceed the corresponding training cycles for the deployed model.
- In addition, trillions of inference per day across Facebook's data centers—more than doubling in the past 3 years. The increase in inference demands has also led to an 2.5× increase in AI inference infrastructure capacity



Training vs Inference



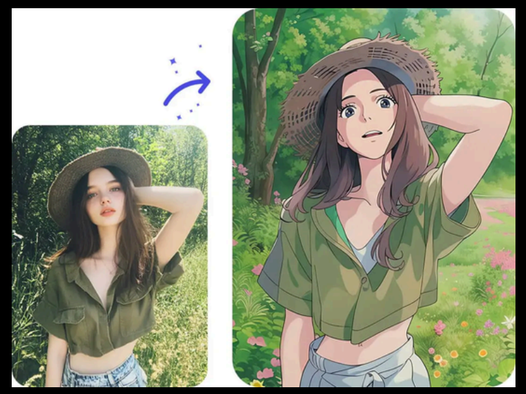
- The carbon footprint of the LM model is dominated by Inference whereas, for RM1 – RM5, the carbon footprint of Training versus Inference is roughly equal

- In contrast to the finite span of model training, inference operates continuously. The carbon emissions thus become a long-term factor, potentially outweighing the training phase emissions over the operational lifetime of the model.

Quick Question for Everyone!

In the past month, how many of you have...

No judgment here - we're all exploring these amazing new tools!



01 Created an AI image in the Ghibli style? 🖼️

02 Used AI to write or edit text? ✍️

03 Done both? 🖼️✍️

04 Neither yet? 🤔

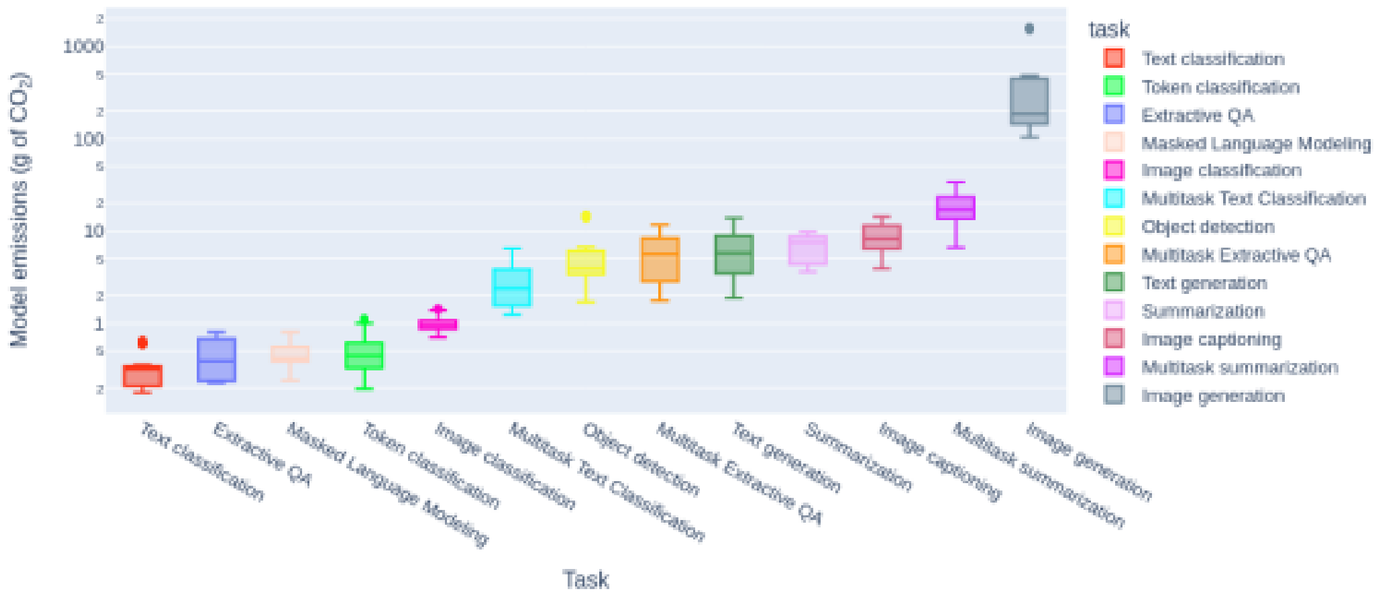
Paper that looks at Inference part of the ML Lifecycle

Power Hungry Processing: ⚡ Watts ⚡ Driving the Cost of AI Deployment?

ALEXANDRA SASHA LUCCIONI and YACINE JERNITE, Hugging Face, Canada/USA
EMMA STRUBELL, Carnegie Mellon University, Allen Institute for AI, USA

- **Experiment Setup**: The study involved analyzing 88 models across different tasks and datasets, focusing on the inference stage of ML models.
- **Methodology**: The research analyzed the impact of various factors such as end task modality, model size, architecture, and learning paradigm on the energy efficiency of these models.
- **Results**: It found significant differences in energy requirements per inference, depending on the model and task. The study particularly noted the higher energy costs of multi-purpose systems.
- **Implications**: The findings illuminate critical aspects of energy use in ML, particularly the balance between the functionality of multi-purpose models and their environmental impact.

Image vs Text Tasks



	inference energy (kWh)	
task	**mean**	**std**
text classification	0.002	0.001
extractive QA	0.003	0.001
masked language modeling	0.003	0.001
token classification	0.004	0.002
image classification	0.007	0.001
object detection	0.038	0.02
text generation	0.047	0.03
summarization	0.049	0.01
image captioning	0.063	0.02
image generation	2.907	3.31

Carbon Footprint of GPT-4

- Energy Consumption in 1 hour:

  No. of hardware * average hardware power draw * TDP * PUE

- Total Tokens in 1 hour:

  (DAU * Average Daily Queries * Avg Tokens/Query) / 24

- Carbon Footprint:

  (Tokens * Energy/token * gCO2e/KWh)

- 28,936 Nvidia A100 GPUs
- TDP of 6.5 kW/server
- PUE 1.2
- 13 million DAU
- 15 daily queries of 2k tokens
- 20 tokens/sec
- 240.6 gCO2e/KWh for Microsoft Azure US West

0.3 gCo2e for 1k tokens

What about DALL-E 3?

DALL-E 2 was estimated at 2.2g CO2e/image

We assume DALL-E 3 must be at least 4 gCO2e/image

Source: Semianalysis
Microsoft PUE
DALL-E 2 gCO2e

Realizing Carbon Footprint:

We discovered the massive environmental impact of AI usage that most users were completely unaware of

Building a community-powered ecosystem of efficient AI usage that makes sustainability the default, not the exception.

extension.scaledown.ai

Realising Carbon Footprint behind every query

Users didn't engage with tools that simply made them feel bad about their AI usage

Beyond Carbon Guilt

Users didn't engage with tools that simply made them feel bad about their AI usage. Negative messaging created awareness but didn't drive sustainable behavior change.

Building a community-powered ecosystem of efficient AI usage that makes sustainability the default, not the exception.

extension.scaledown.ai

Realising Carbon Footprint behind every query

Users didn't engage with tools that simply made them feel bad about their AI usage

Beyond Carbon Guilt

Users didn't engage with tools that simply made them feel bad about their AI usage. Negative messaging created awareness but didn't drive sustainable behavior change.

Pivot to Prompt Optimization

We shifted to developing a prompt optimization tool that could reduce token usage while maintaining output quality, making it easy for users to see immediate benefits.

Building a community-powered ecosystem of efficient AI usage that makes sustainability the default, not the exception.

extension.scaledown.ai

Realising Carbon Footprint behind every query

Users didn't engage with tools that simply made them feel bad about their AI usage

Beyond Carbon Guilt

Users didn't engage with tools that simply made them feel bad about their AI usage. Negative messaging created awareness but didn't drive sustainable behavior change.

Pivot to Prompt Optimization

We shifted to developing a prompt optimization tool that could reduce token usage while maintaining output quality, making it easy for users to see immediate benefits.

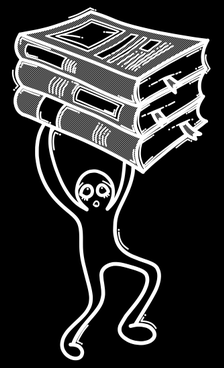
The Current Approach

We realized technical optimization alone was creating the behavioral change needed. By adding community-driven prompt templates, we're helping users structure their queries effectively while transparently showing the carbon impact of their choices.

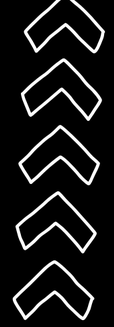
Building a community-powered ecosystem of efficient AI usage that makes sustainability the default, not the exception.

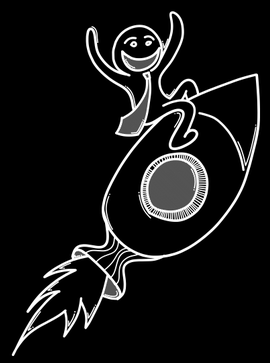
extension.scaledown.ai

extension.scaledown.ai

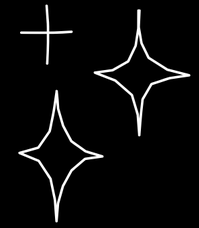


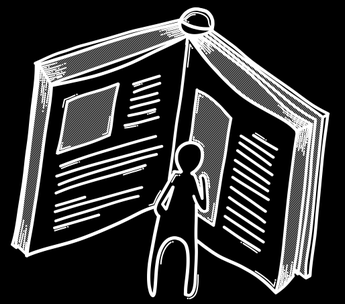
Demo Time





Are you ready?





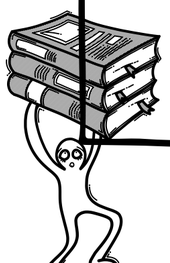
So what does Carbon ScaleDown Do?

extension.scaledown.ai

Write Prompts, Optimise tokens and Minimize Carbon Footprint

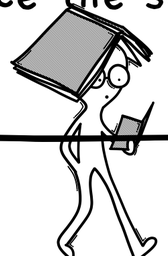
Write Prompts and Choose Styles

- Get domain-specific prompts for specialized fields
- Choose from different response styles (concise, detailed, creative)
- Access a library of community-contributed prompt templates



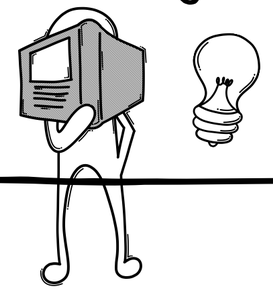
Optimise AI prompts and reduce Tokens

- Remove unnecessary phrases while preserving meaning
- Apply best practices for each AI model (Claude, GPT, Llama)
- Verify that optimized prompts produce the same results



Minimize Carbon Footprint

- Track AI emissions
- Optimize AI usage to reduce Carbon Footprint
- Reach ESG goals without sacrificing AI efficiency gains

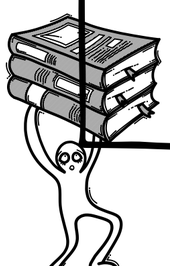


So what does Carbon ScaleDown Do?

extension.scaledown.ai

Write Prompts, Optimise tokens and Minimize Carbon Footprint

Write Prompts and Choose Styles		
• Get domain-specific prompts for specialized fields • Choose from different response styles (concise, detailed, creative) • Access a library of community-contributed prompt templates		



Prompt Templates

Template Library: By the Community, For the Community

Current Template Categories:

- Writing: Academic essays, blog posts, creative stories, product descriptions
- Technical: Code documentation, bug reports, technical explanations
- Business: Meeting agendas, project proposals, marketing copy, emails
- Creative: Character descriptions, story starters, world-building

Prompt Templates

Template Library: By the Community, For the Community

Current Template Categories:

- Writing: Academic essays, blog posts, creative stories, product descriptions
- Technical: Code documentation, bug reports, technical explanations
- Business: Meeting agendas, project proposals, marketing copy, emails
- Creative: Character descriptions, story starters, world-building

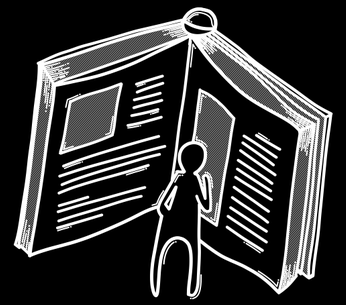
Contribute Your Templates:

- Share your most effective prompt patterns
- Help others in specific domains or use cases
- Get community feedback and improvements
- Build your reputation as an AI prompt engineer

docs.scaledown.ai

Ready to Contribute?

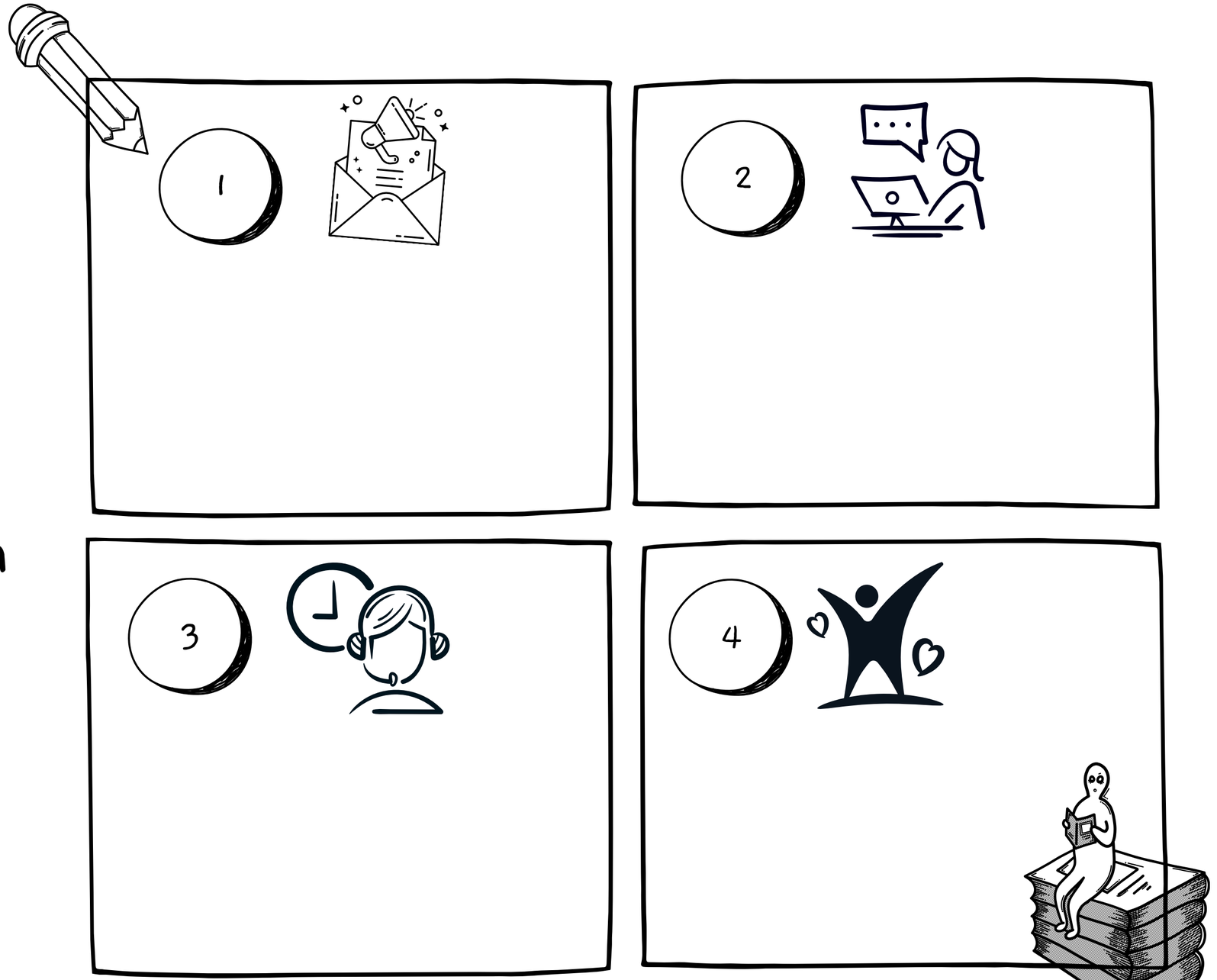
Lets check the Templates



Even small efficiency improvements, when adopted widely, can have massive impact on global AI energy usage!

Quiz Time

Why use templates instead of writing prompts from scratch each time?

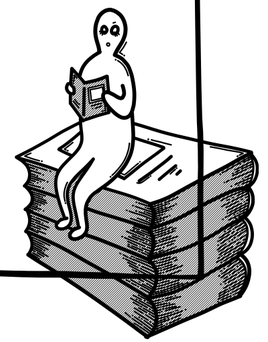


Quiz Time

Why use templates instead of writing prompts from scratch each time?

1

Templates are always shorter than custom prompts



Quiz Time

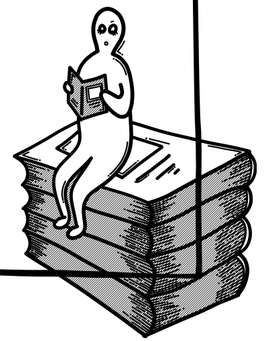
Why use templates instead of writing prompts from scratch each time?

1

Templates are always shorter than custom prompts

2

Templates guarantee perfect AI responses every time



Quiz Time

Why use templates instead of writing prompts from scratch each time?

1

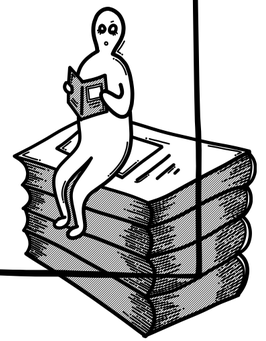
Templates are always shorter than custom prompts

2

Templates guarantee perfect AI responses every time

3

Templates provide structure and save time while maintaining consistency



Quiz Time

Why use templates instead of writing prompts from scratch each time?

1

Templates are always shorter than custom prompts

2

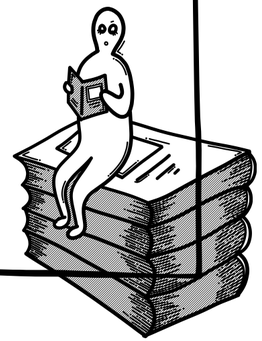
Templates guarantee perfect AI responses every time

3

Templates provide structure and save time while maintaining consistency

4

Templates bypass token limits on AI models



So what does Carbon ScaleDown Do?

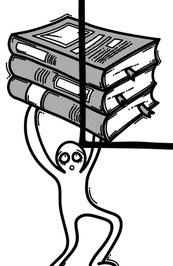
Write Prompts, Optimise tokens and Minimize Carbon Footprint

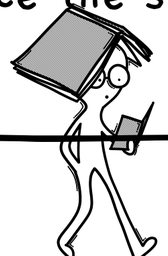
Optimise AI prompts and reduce Tokens

- Remove unnecessary phrases while preserving meaning

- Apply best practices for each AI model (Claude, GPT, Llama)

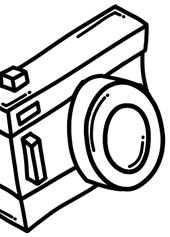
- Verify that optimized prompts produce the same results





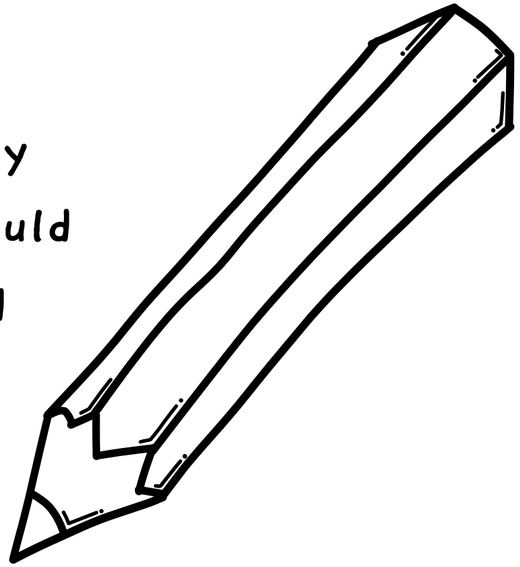
blog.scaledown.ai

https://tinyurl.com/agentcon-workshop



How much could the following prompt be reduced by using ScaleDown's optimization techniques?

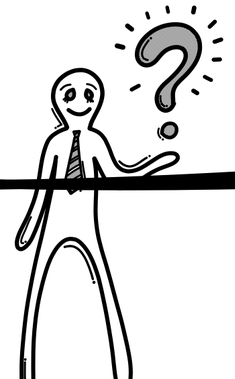
Could you please kindly help me write a comprehensive analysis of the quarterly financial results for my company if you don't mind? I would really appreciate it if you could include all the important metrics and insights that would be helpful for the executive team to understand.



Token reduction

Could you please kindly help me write a comprehensive analysis of the quarterly financial results for my company if you don't mind? I would really appreciate it if you could include all the important metrics and insights that would be helpful for the executive team to understand.

Check the Mentimeter



1. 20-30% fewer tokens
2. 40-50% fewer tokens
3. 60-70% fewer tokens
4. 70-80% fewer tokens

docs.scaledown.ai

ScaleDown Optimization Pipeline

Input Parser & Analyzer

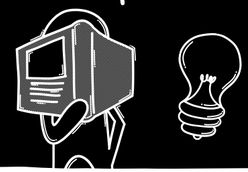
- Parses user input and analyzes content
- Identifies key components of the prompt
- Detects redundancies, filler phrases, and politeness markers

Input Parser & Analyze

- Applies semantic optimization rules
- Removes unnecessary phrases and terms
- Restructures content for maximum efficiency
- Uses model-specific prompt engineering techniques

Output Creator

- Generates the optimized prompt
- Maintains semantic meaning while reducing token count
- Formats the final output according to model requirements



docs.scaledown.ai

ScaleDown

Review and Evaluate your Prompts

Reduce Token Usage by up to 80%
Without sacrificing output quality

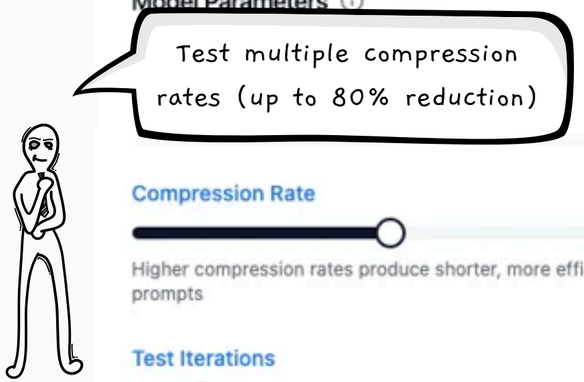
Scaledown API Key

Enter your API key

Set Key

Your API key is required to run consistency tests and is stored only in your browser's session.

Model Parameters



Compression Rate 50%

Higher compression rates produce shorter, more efficient prompts

Test Iterations 10

More iterations provide more accurate consistency results

Prompt Consistency Test

Run the test to see results

Generate Single Response

Original Prompt

Sample Templates

Create or select a prompt template to compress

Enter your prompt

Enter your policy template here...

Original Response

Standard response generated from your prompt

No response generated yet

Compressed Response

Optimized response from compressed prompt

No response generated yet

Scaledown API Key

Enter your API key

Set Key

Your API key is required to run consistency tests and is stored only in your browser's session.

Model Parameters

AI Model

GPT-4o

Compression Rate 50%

Higher compression rates produce shorter, more efficient prompts

Test Iterations 10

More iterations provide more accurate consistency results

Prompt Consistency Test

Run the test to see results

Generate Single Response

Run Consistency Test

Original Prompt

Sample Templates

Create or select a prompt template to compress

Enter your prompt

Enter your policy template here...

Original Response

Standard response generated from your prompt

No response generated yet

Compressed Response

Optimized response from compressed prompt

No response generated yet

Run consistency tests across multiple iterations

Scaledown API Key

Enter your API key

Set Key

Your API key is required to run consistency tests and is stored only in your browser's session.

Model Parameters

AI Model

GPT-4o

Compression Rate 50%

Higher compression rates produce shorter, more efficient prompts

Test Iterations 10

More iterations provide more accurate consistency results

Prompt Consistency Test

Run the test to see results

Generate Single Response

Run Consistency Test

Original Prompt

Create or select a prompt template to compress

Sample Templates

Choose Different Templates

Enter your prompt

Enter your policy template here...

Original Response

Standard response generated from your prompt

No response generated yet

Compressed Response

Optimized response from compressed prompt

No response generated yet

Scaledown API Key

Enter your API key

Set Key

Your API key is required to run consistency tests and is stored only in your browser's session.

Model Parameters

AI Model

GPT-4o

Choose Specific Models

Compression Rate 50%

Higher compression rates produce shorter, more efficient prompts

Test Iterations 10

More iterations provide more accurate consistency results

Prompt Consistency Test

Run the test to see results

Generate Single Response

Run Consistency Test

Original Prompt

Sample Templates

Create or select a prompt template to compress

Enter your prompt

Enter your policy template here...

Original Response

Standard response generated from your prompt

No response generated yet

Compressed Response

Optimized response from compressed prompt

No response generated yet

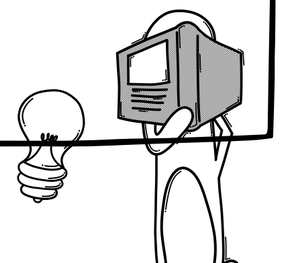
How to contribute?

Beginner Level: Consistency Challenge

- Features an interactive leaderboard for people to compete on optimizing prompts while maintaining consistency
- Includes weekly challenges across different domains
- Encourages users to submit their results and learn from the community

Advanced Level: Code Contribution

- Enhancing the semantic optimizer patterns in semantic_optimizer.py
- Adding optimization guides for new models in prompting_guides.py
- Improving the consistency testing framework in the GuideBasedOptimizer class
- Developing specialized templates for different industries in default_templates.py



scaledown-package Public

Edit Pins | Unwatch 1 | Fork 0 | Star 0

main | 1 Branch | 0 Tags | Go to file t | Add file | Code

Open Source

varchanaiyer example file for llamaindex — d1f3016 · 15 hours ago — 9 Commits

examples	example file for llamaindex	15 hours ago
src/scaledown	changes to compression api	15 hours ago
tests	changes to compression api	15 hours ago
.gitignore	Create .gitignore	2 days ago
LICENSE	Initial commit	2 days ago
README.md	changes to compression api	15 hours ago
pyproject.toml	first package push	2 days ago

README | GPL-3.0 license

ScaleDown

SCALEDOWN | AI PROMPT OPTIMIZATION

pypi v0.1.0 | python 3.9+ | License MIT | docs latest

ScaleDown is a powerful Python library for optimizing prompts for large language models. It helps reduce token usage while preserving semantic meaning, saving costs and improving response quality when working with AI

About

No description, website, or topics provided.

Readme
GPL-3.0 license
Activity
Custom properties
0 stars
1 watching
0 forks
Report repository

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Contributors 2

varchanaiyer Archana Vaidheeswaran
soham96 Soham Chatterjee

scaledown-package Public

Edit Pins | Unwatch 1 | Fork 0 | Star 0

main | 1 Branch | 0 Tags | Go to file t | Add file | Code

Open Source

Easy integration into your existing workflows

varchanaiyer example file for llamaindex — d1f3016 · 15 hours ago — 9 Commits

examples	example file for llamaindex	15 hours ago
src/scaledown	changes to compression api	15 hours ago
tests	changes to compression api	15 hours ago
.gitignore	Create .gitignore	2 days ago
LICENSE	Initial commit	2 days ago
README.md	changes to compression api	15 hours ago
pyproject.toml	first package push	2 days ago

README | GPL-3.0 license

ScaleDown

SCALEDOWN | AI PROMPT OPTIMIZATION

pypi v0.1.0 | python 3.9+ | License MIT | docs latest

ScaleDown is a powerful Python library for optimizing prompts for large language models. It helps reduce token usage while preserving semantic meaning, saving costs and improving response quality when working with AI

About

No description, website, or topics provided.

- Readme
- GPL-3.0 license
- Activity
- Custom properties
- 0 stars
- 1 watching
- 0 forks

Report repository

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Contributors 2

varchanaiyer Archana Vaidheeswaran

soham96 Soham Chatterjee

scaledown-package Public

Edit Pins | Unwatch 1 | Fork 0 | Star 0

main | 1 Branch | 0 Tags | Go to file t | Add file | Code

Open Source

Easy integration into your existing workflows

Interactive Demos, to customise and easily plug it into current flows

varchanaiyer example file for llamaindex — d1f3016 · 15 hours ago — 9 Commits

examples	example file for llamaindex	15 hours ago
src/scaledown	changes to compression api	15 hours ago
tests	changes to compression api	15 hours ago
.gitignore	Create .gitignore	2 days ago
LICENSE	Initial commit	2 days ago
README.md	changes to compression api	15 hours ago
pyproject.toml	first package push	2 days ago

README | GPL-3.0 license

ScaleDown

SCALEDOWN | AI PROMPT OPTIMIZATION

pypi v0.1.0 | python 3.9+ | License MIT | docs latest

ScaleDown is a powerful Python library for optimizing prompts for large language models. It helps reduce token usage while preserving semantic meaning, saving costs and improving response quality when working with AI

About

No description, website, or topics provided.

Readme
GPL-3.0 license
Activity
Custom properties
0 stars
1 watching
0 forks
Report repository

Releases

No releases published
Create a new release

Packages

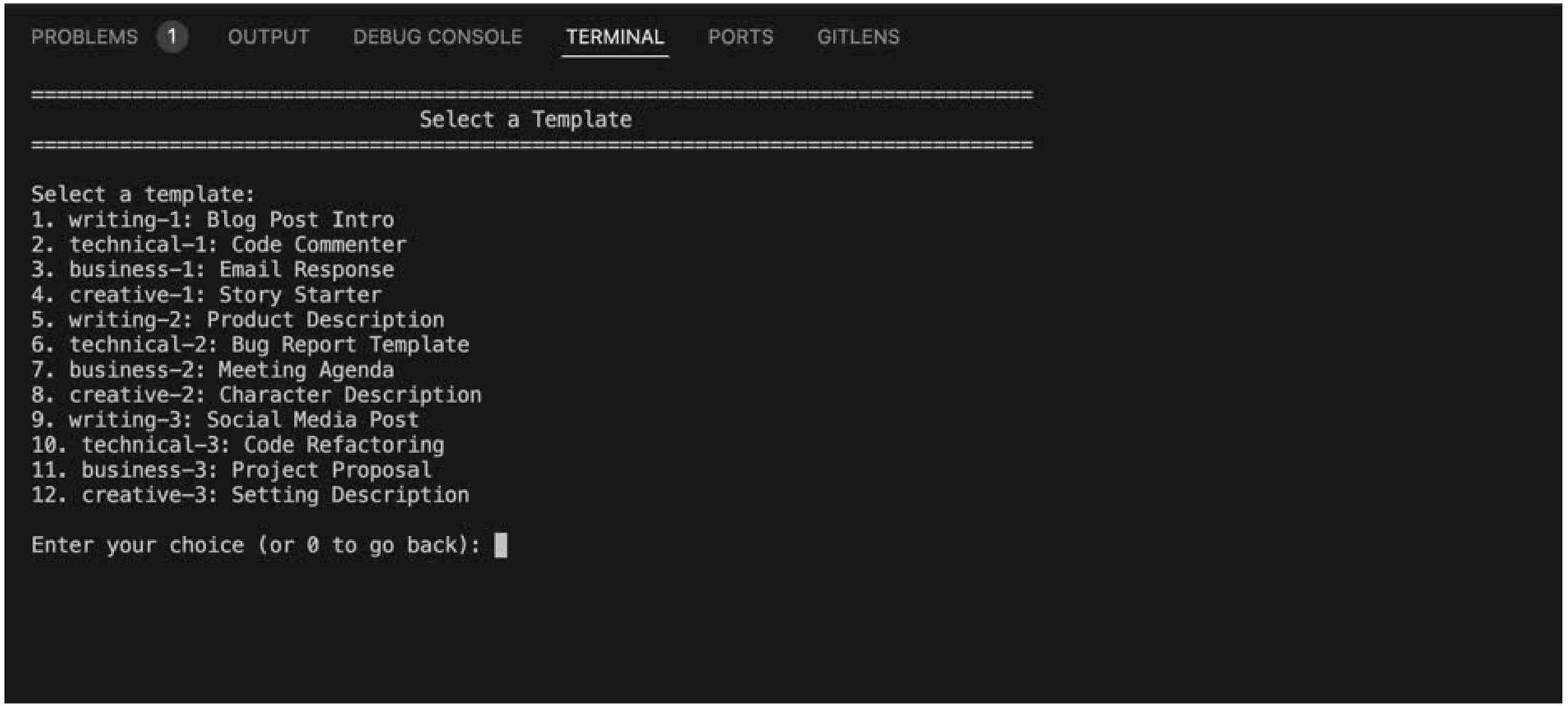
No packages published
Publish your first package

Contributors 2

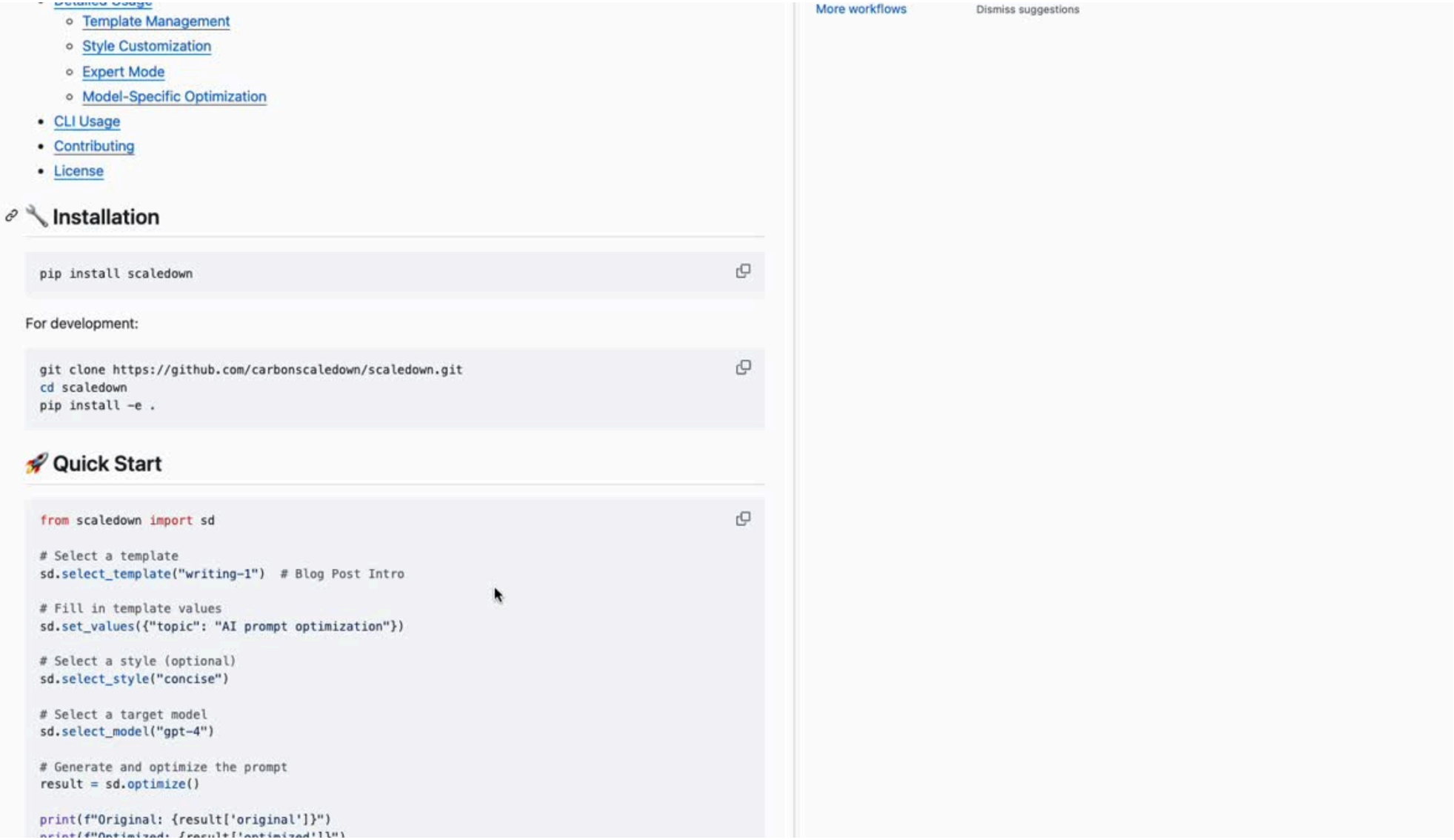
varchanaiyer Archana Vaidheeswaran

soham96 Soham Chatterjee

Check the interactive test file to start on your local device



Run the Example Files to understand the workflow



- Template Management
- Style Customization
- Expert Mode
- Model-Specific Optimization

- CLI Usage
- Contributing
- License

🔧 Installation

```
pip install scaledown
```

For development:

```
git clone https://github.com/carbonscaledown/scaledown.git
cd scaledown
pip install -e .
```

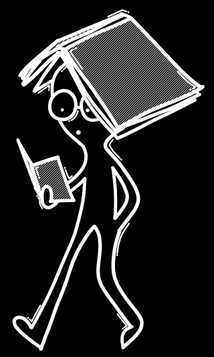
🚀 Quick Start

```
from scaledown import sd

# Select a template
sd.select_template("writing-1")  # Blog Post Intro

# Fill in template values
sd.set_values({"topic": "AI prompt optimization"})

# Select a style (optional)
sd.select_style("concise")

# Select a target model
sd.select_model("gpt-4")

# Generate and optimize the prompt
result = sd.optimize()

print(f"Original: {result['original']}")
```

More workflows

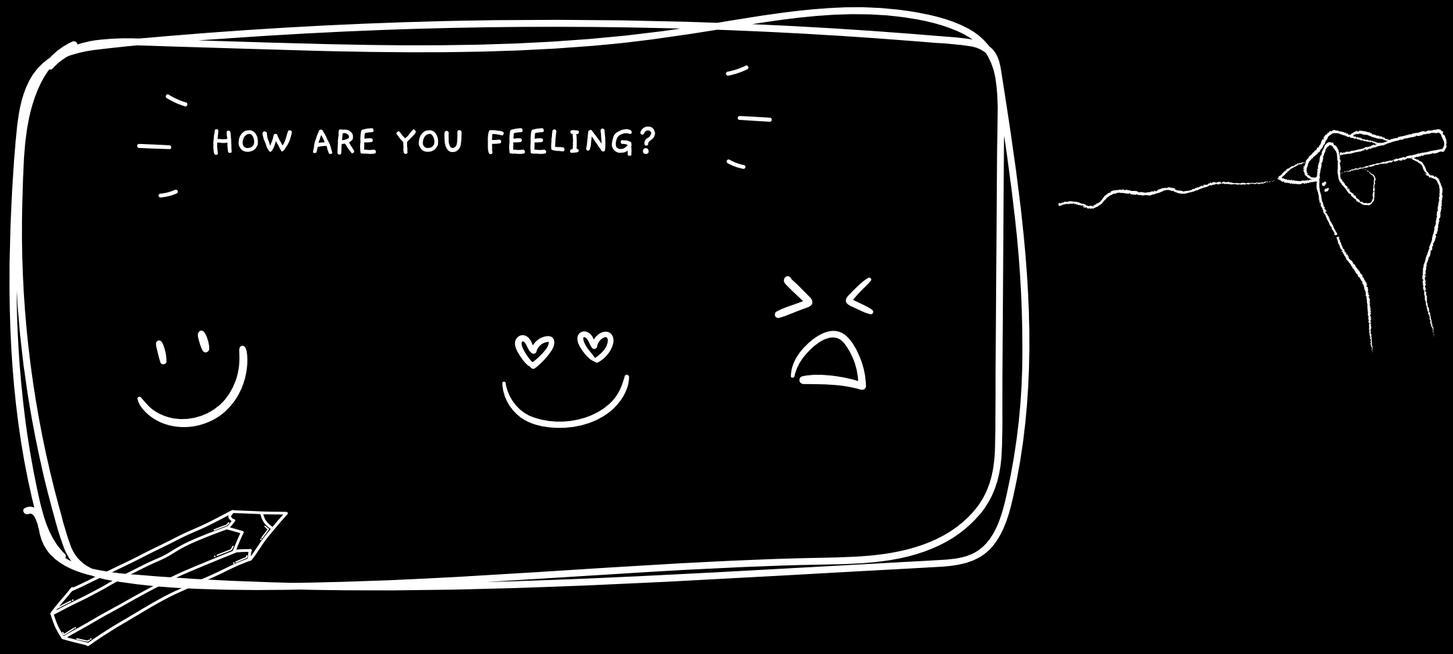
Dismiss suggestions

Five Ways to Make an Impact:



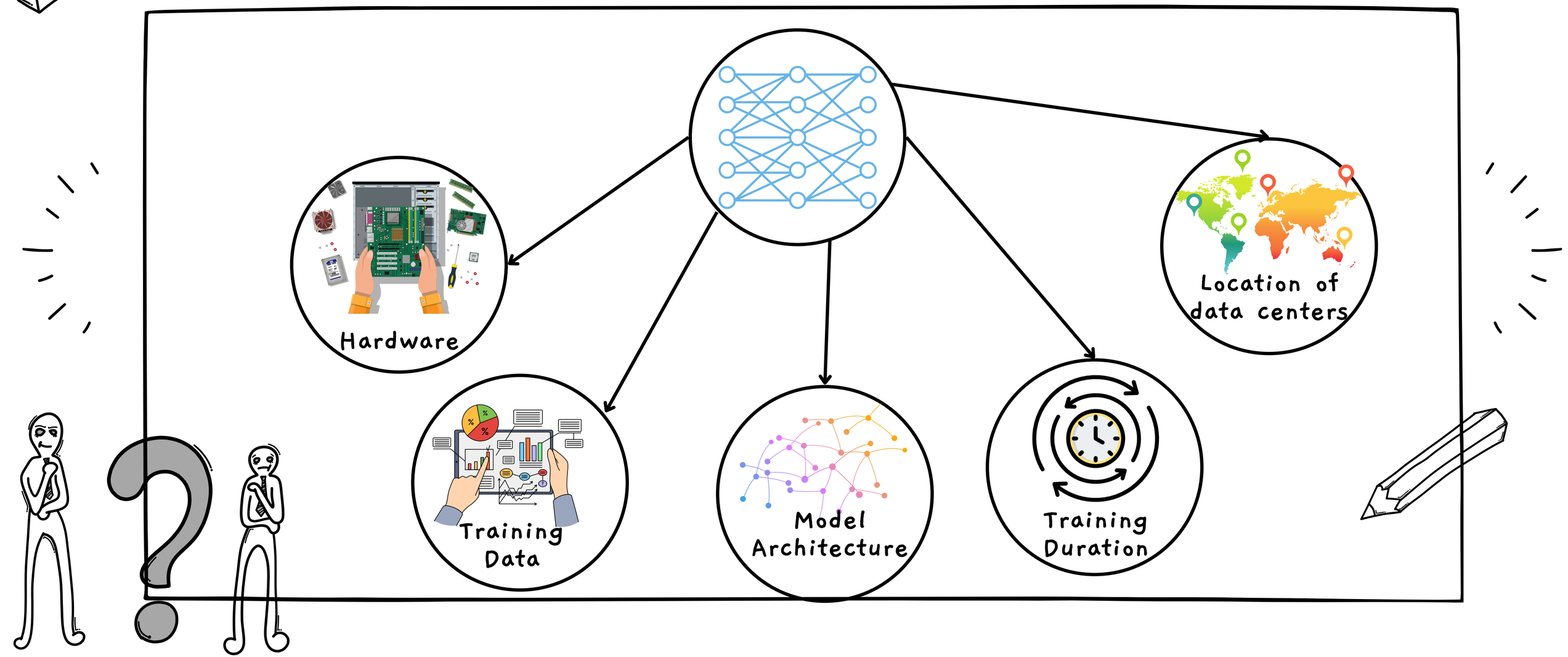
1. Contribute to the Code: Add new optimization patterns, templates, or features "Even small improvements can save millions of tokens globally!"
2. Use the Browser Extension: Start optimizing your own prompts today "Be amazed at how much more efficient your AI interactions become!".
3. Share Your Feedback: Help us make the tools better for everyone "Your insights drive our community's innovation!"
4. Spread the Word: Convince 5 friends to try ScaleDown "The average user saves 40-60% of their AI tokens—imagine that at scale!
5. Contribute to the Template Library: Share your most effective prompts "Your expertise could help thousands of other users!"

Check out the app at extension.scaledown.ai

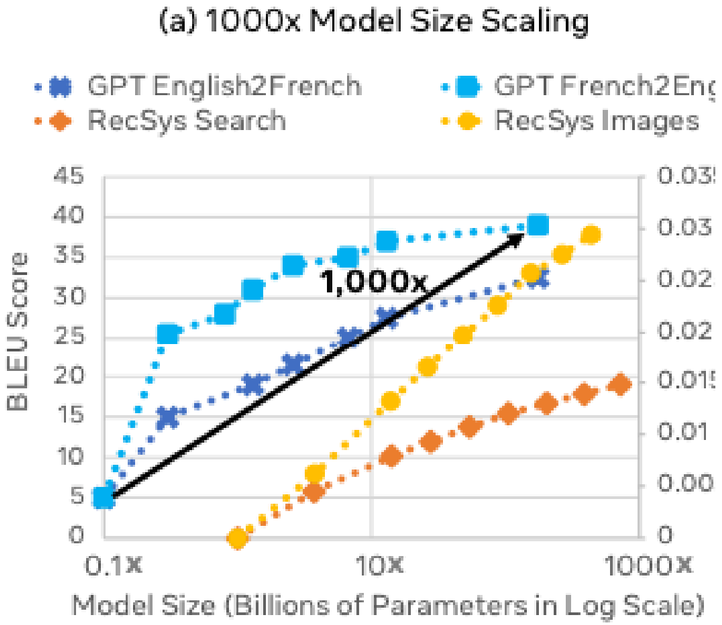


Appendix

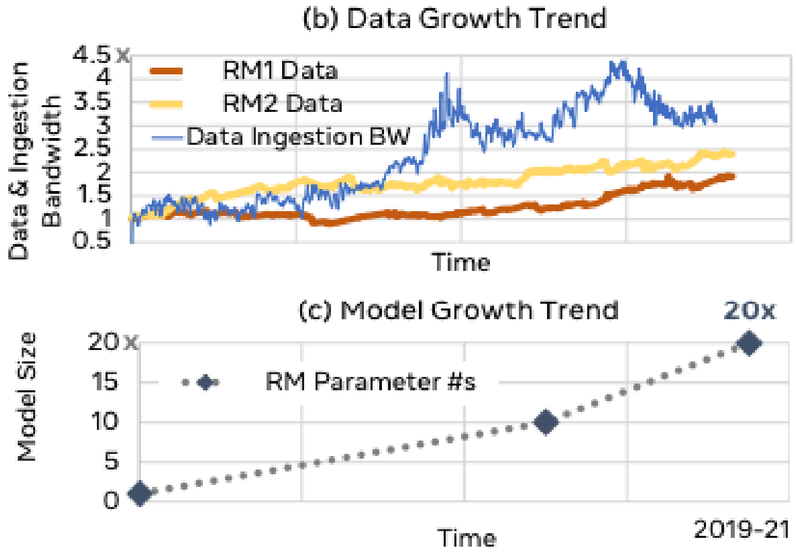
The Factors Affecting Carbon Footprint of ML Models



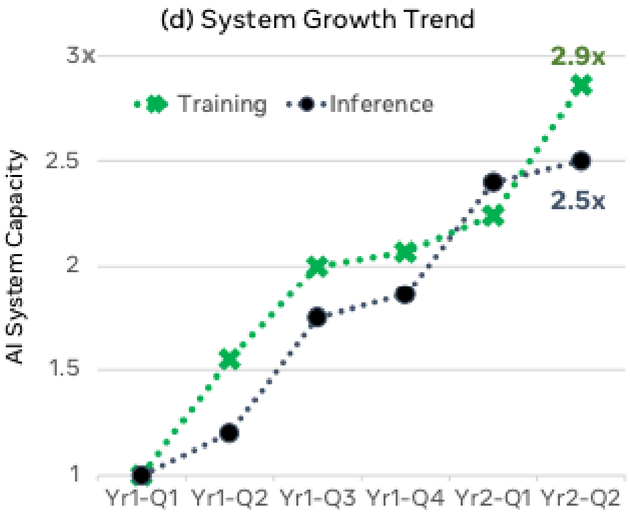
" Deep learning has witnessed an exponential growth in data, model parameters, and system resources over the recent years" .



The 1000× model size growth has led to higher model accuracy for various ML tasks



At Facebook, the amount of data for recommendation use cases has roughly doubled between 2019 and 2021, leading to 3.2 times increase in the data ingestion bandwidth demand.) Facebook's recommendation and ranking model sizes have increased by 20 times during the same time period



The explosive growth in AI has driven 2.9× and 2.5× capacity increases for AI training and inference

Results from the Paper

Model	Hardware	Power (W)	Hours	kWh·PUE	CO_2e
$\text{Transformer}_{base}$	P100x8	1415.78	12	27	26
Transformer_{big}	P100x8	1515.43	84	201	192
ELMo	P100x3	517.66	336	275	262
BERT_{base}	V100x64	12,041.51	79	1507	1438
BERT_{base}	TPUv2x16	—	96	—	—
NAS	P100x8	1515.43	274,120	656,347	626,155
NAS	TPUv2x1	—	32,623	—	—
GPT-2	TPUv3x32	—	168	—	—

Transformer:

- Base and big models, with 65M and 213M parameters, respectively.
- Training specifics: 8 NVIDIA P100 GPUs for the base model (12 hours) and big model (84 hours).
- Neural Architecture Search (NAS) for these models required 32,623 TPU hours or 274,120 P100 GPU hours.

ELMo:

- Stacked LSTM architecture, trained for rich word representations.
- Training specifics: 3 NVIDIA GTX 1080 GPUs for 2 weeks (336 hours).

BERT:

- Transformer-based architecture for contextual representations.
- Training specifics: 16 TPU chips for 4 days (96 hours) or 64 Tesla V100 GPUs for 3.3 days (79.2 hours).

GPT-2:

- General-purpose token encoder with self-attention.
- Training specifics: 1542M parameters trained for 1 week (168 hours) on 32 TPUv3 chips.

Model	BERT finetune	BERT pretrain	6B Transf.	Dense 121	Dense 169	Dense 201	ViT Tiny	ViT Small	ViT Base	ViT Large	ViT Huge
GPU	4·V100	8·V100	256·A100	1·P40	1·P40	1·P40	1· V100	1·V100	1·V100	4·V100	4·V100
Hours	6	36	192	0.3	0.3	0.4	19	19	21	90	216
kWh	3.1	37.3	13,812.4	0.02	0.03	0.04	1.7	2.2	4.7	93.3	237.6

Power Consumption to train some common models

Note that the power required to train a 6B transformer is more than the generation capacity of Singapore

pc =the average powerdraw (in watts) from all CPU sockets during training,

pr =average power draw from all DRAM (main memory) sockets

pg =average power draw of a GPU during training,

g =be the number of GPUs used to train

Power Usage Effectiveness (PUE)=accounts for the additional energy required to support the compute infrastructure (mainly cooling)

PUE coefficient =1.58;average annual power usage effectiveness (PUE) ratio collected by Uptime

$$p_t = \frac{1.58t(p_c + p_r + gp_g)}{1000}$$