From Documents to Dialogues

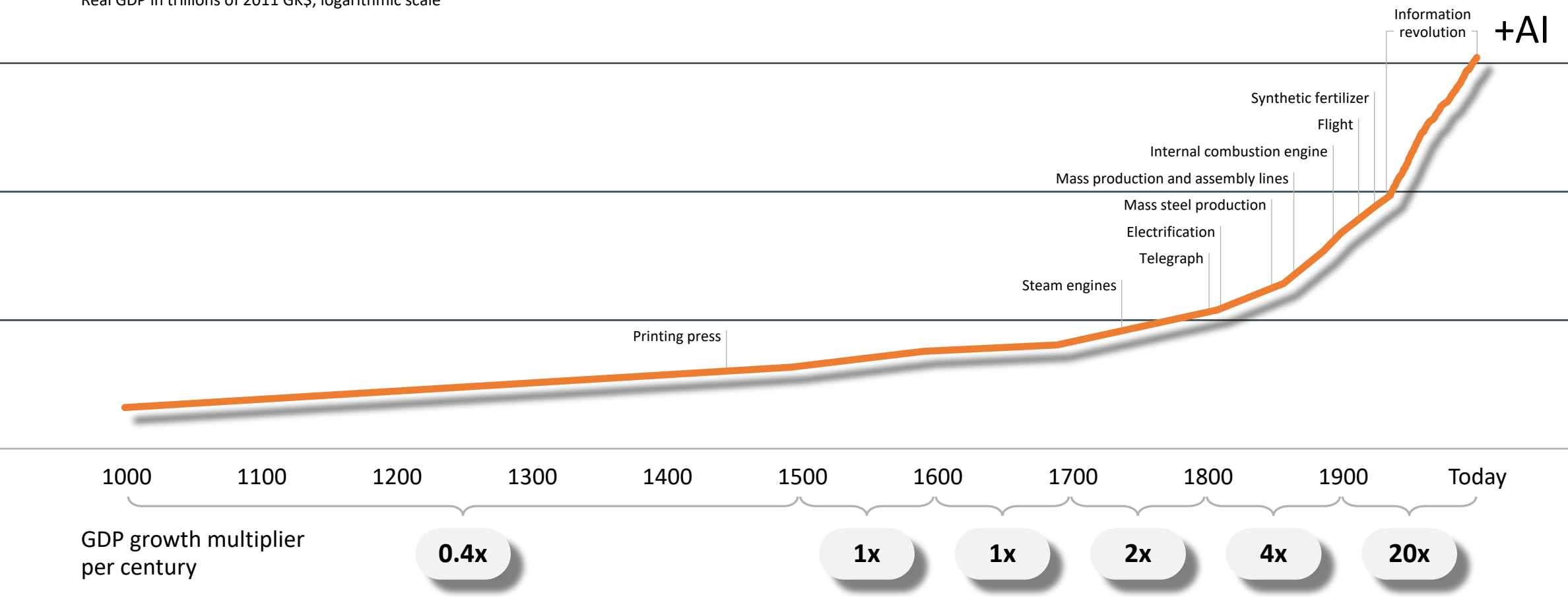
Jaime Teevan

Microsoft

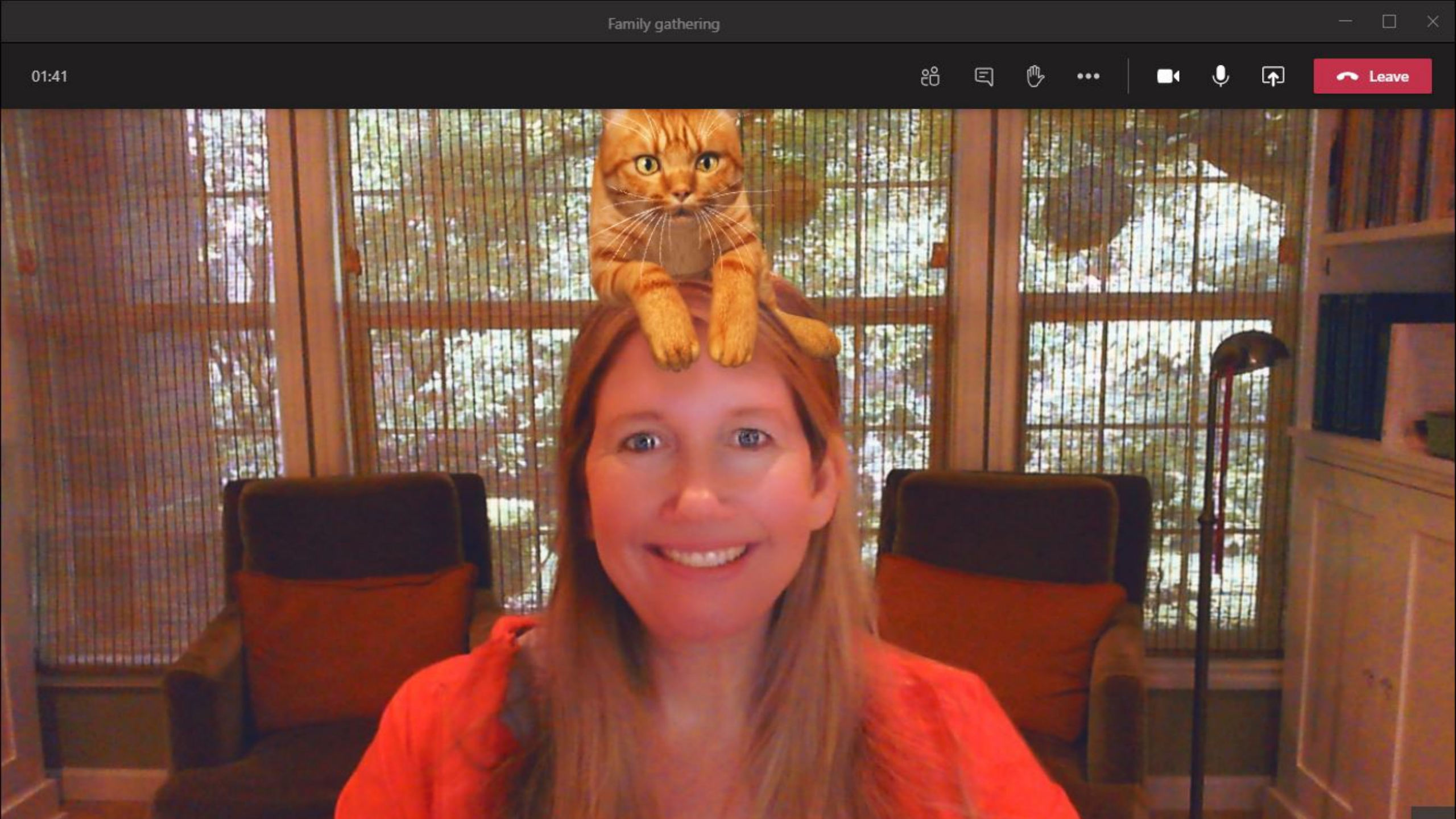
Pursuit of GDP Growth

Global GDP and technological revolutions

Real GDP in trillions of 2011 GK$, logarithmic scale



Source: Maddison Project, OurWorldInData.org


Family gathering
01:41
Leave

Enterprise Grade AI

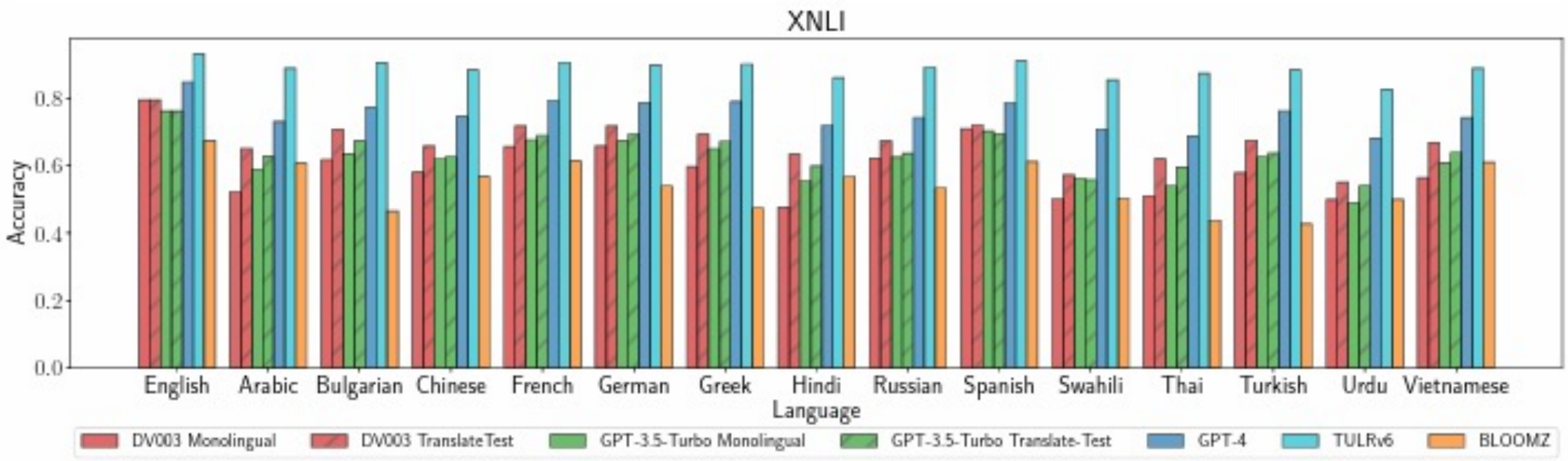
Global Scale

Grounded in Your Data

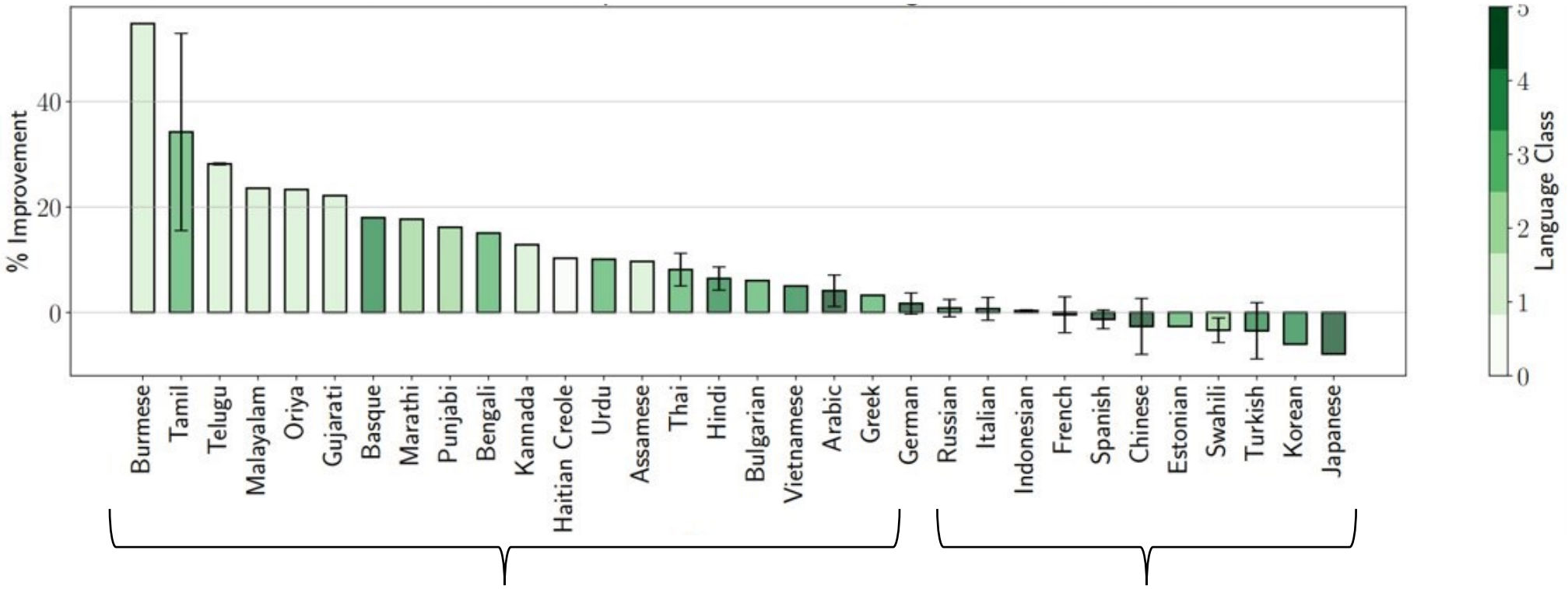
Trustworthy

Embedded in Existing Workflows

Global Scale: *Multilingual*



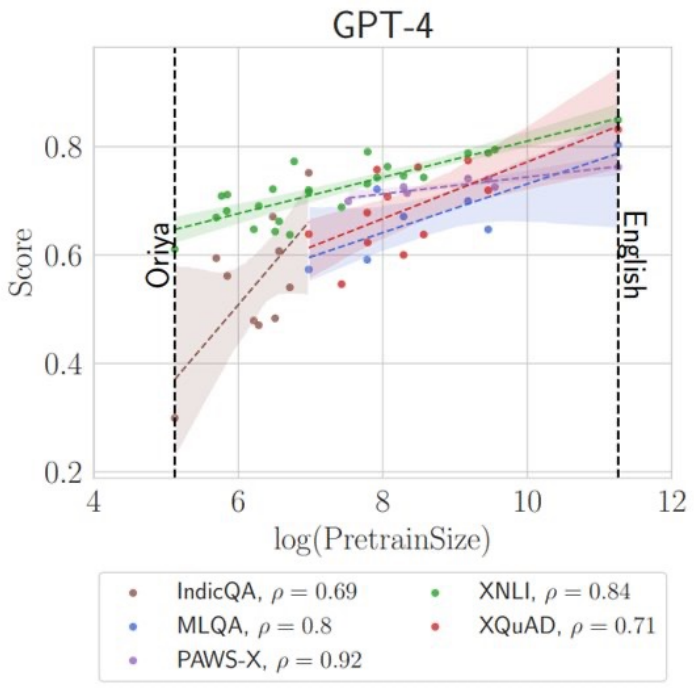
Global Scale: *Multilingual*



Translating through EN improves performance

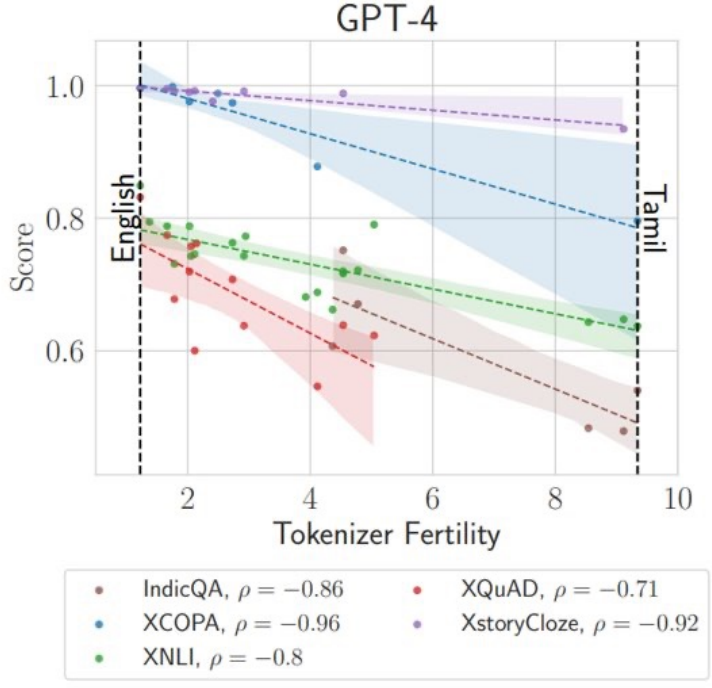
Translating through EN degrades performance

Global Scale: *Multilingual*



(b) Correlation between pre-training size and performance for GPT-4

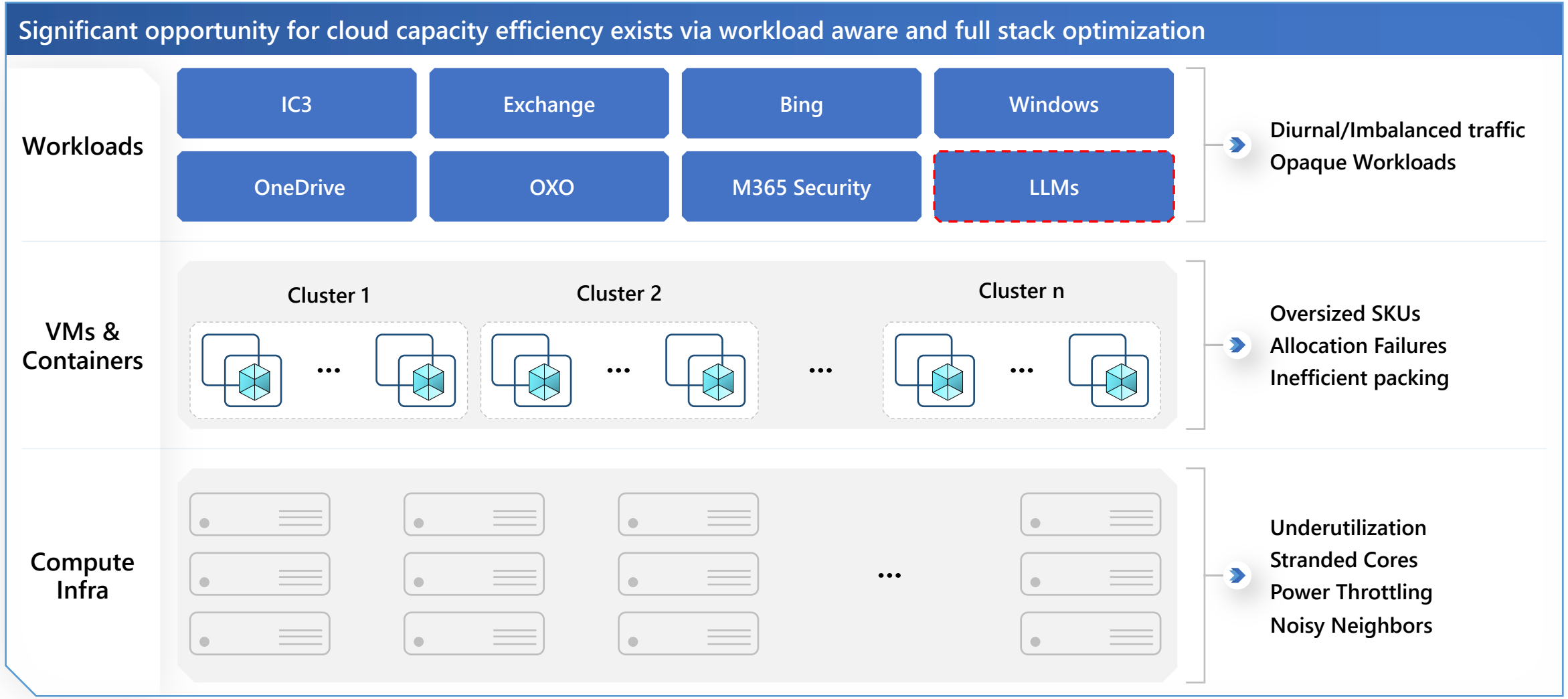
Training data size is dominant driver in performance



(b) Correlation between tokenizer fertility and performance for GPT-4

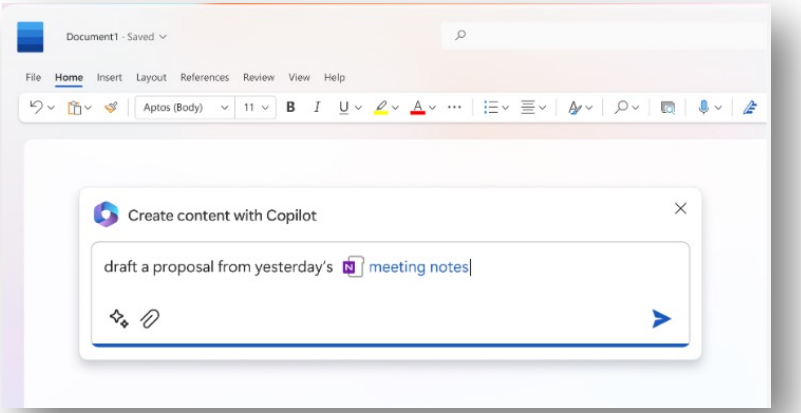
Non-Latin script languages still lag in performance

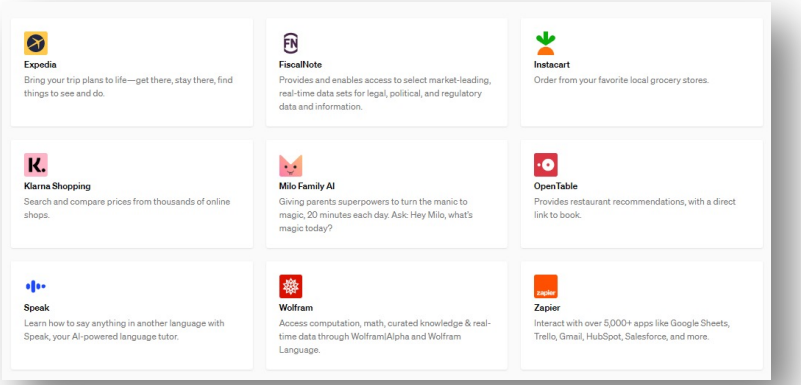
Global Scale: *Efficient*



Global Scale: *Efficient*

- High computing demand
  - 175 Billion parameters for GPT-3 – more for GPT-4
  - Require expensive GPU resources
- Ever-expanding scenarios
  - Large-scaled commercial LLM-powered services
  - Billions of users across world
  - Emerging market for plug-ins
- Opportunity to leverage years of research on cloud efficiency for LLM Infrastructure
  - Profile and characterize LLM workloads for resource optimization with temporal, spatial and power shaping
  - Guarantee high reliability and availability of LLM-powered services





Global Scale: *Efficient*







Data Platform

Collect and process relevant data and accumulate critical knowledge for supporting down streaming tasks

- Runtime Telemetry
- Static Metadata
- Domain Knowledge
- ...

Workload Intelligence

Profile characteristics of workloads for supporting workload-agnostic algorithms

- Resource Requirement
- SLO Metrics
- Temporal Pattern
- Spatial Distribution
- Operation Characteristics
- Service Properties
- ...

Optimization Methods

Develop new optimization framework to handle multi-objective, multi-constraints efficient problems under uncertainty

- Chance-Constraint Optimization
- X-Layer Parameter Optimization
- Proactive Design
- Combinatorial Searching
- Reinforcement Learning
- ...

Efficiency Scenarios

Adapt workload intelligence and optimization methods to different efficiency scenarios for improvements

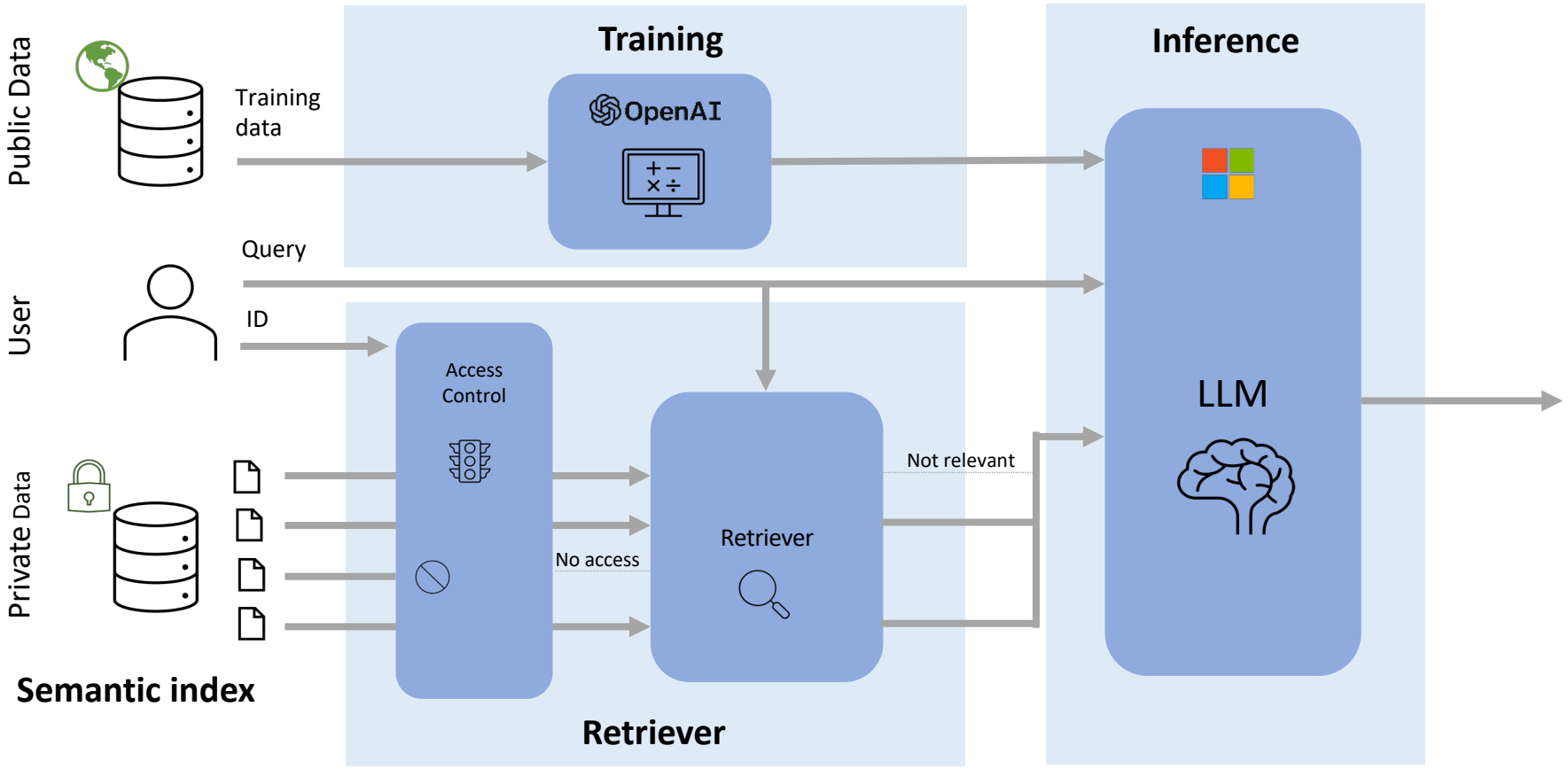
- SKU Recommendation
- Oversubscription
- SPOT/Harvest Migration
- Power Capping
- Region Agnostic Placement
- Multi-Availability Data-Center
- Allocation Failure Prediction
- ...

Global Scale: *Sustainable*

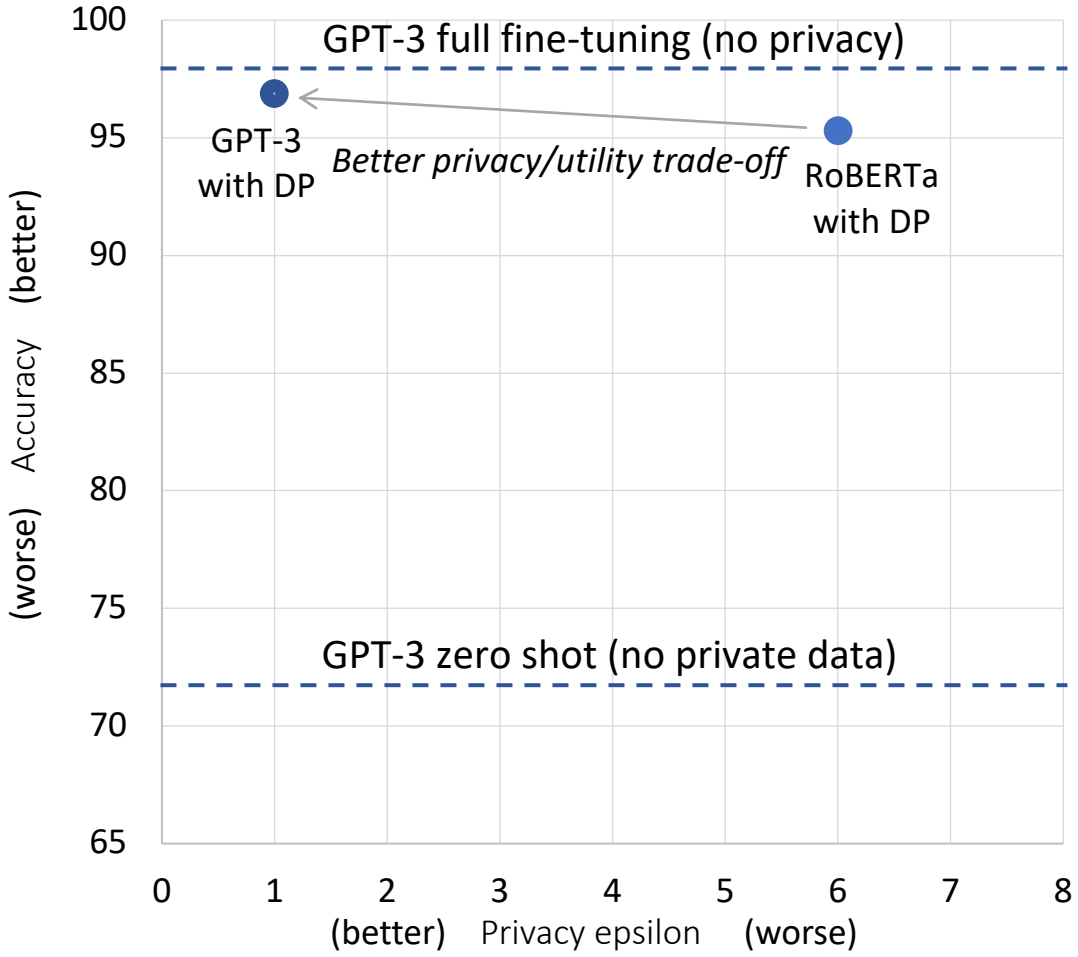
Grounded in Your Data: *RAG*

- Trend in using LLMs: In-context learning
  - Don’t fine-tune – keep the model weights static
  - Provide task instructions and data in the prompt
- Advantages of retrieval augmentation
  - Less information needs to be embedded in the model itself
  - Relevant, context-specific results via increased real-time personalization
  - Factually grounded, less hallucination
  - Traceable information flow with explicit disclosure policies (e.g., ACL filtering)
- Interesting research questions abound
  - New types of data: Documents, application context, chat history
  - Finding the right content: Source selection, query generation
  - Context compressing: Creating a working menu

Grounded in Your Data: *Private by Design*



Trustworthy: *Differential Privacy*

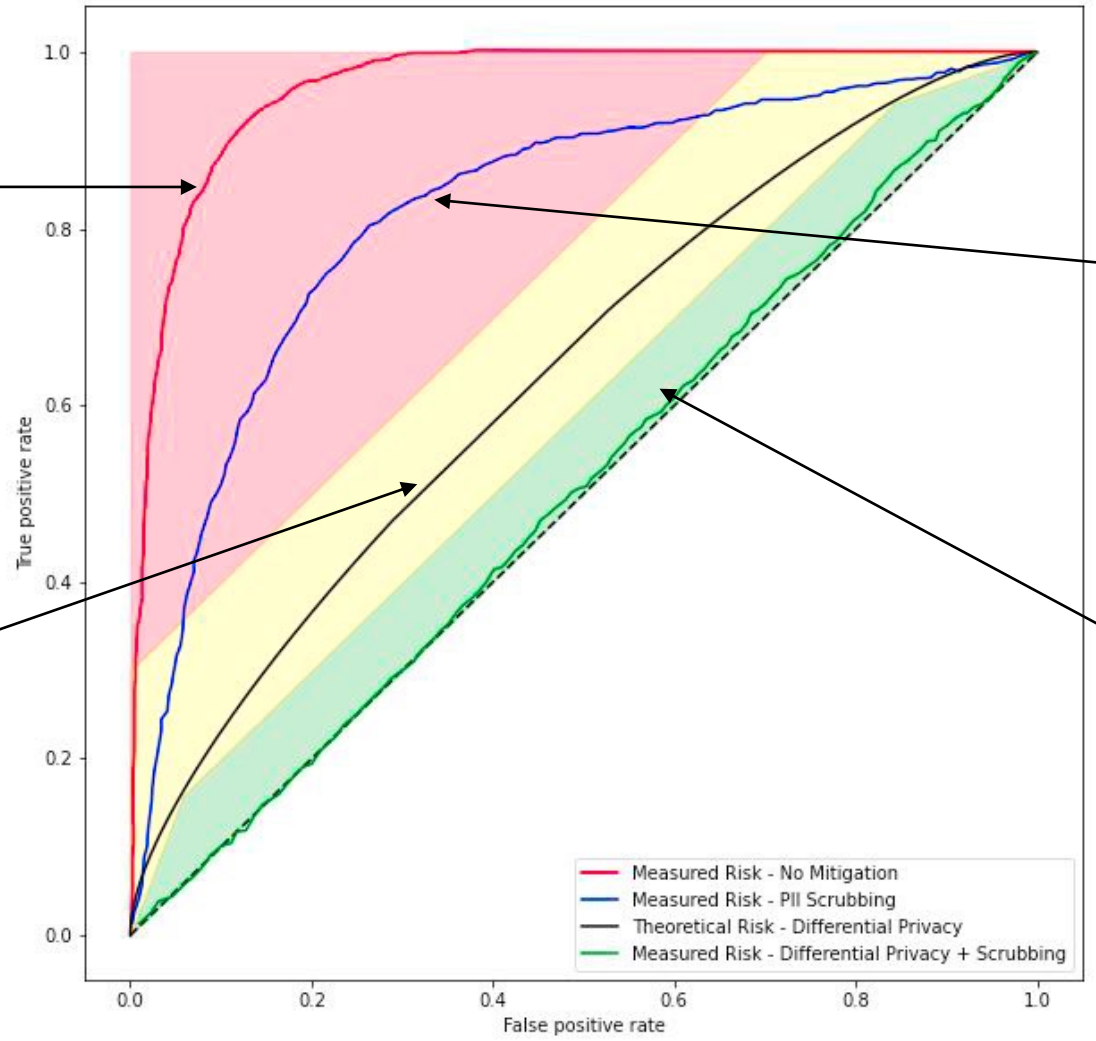


- Private data is valuable
  - DP fine-tuning on private data gives a significant boost in performance
- Solid privacy/utility trade-off
  - Similar accuracy for private v. non private fine-tuning
- Larger models fine-tune better
  - Can privately fine-tune with small epsilon values for very strong privacy
  - Can use a larger epsilon but have small # of parties contribute data

Trustworthy: *Measuring Privacy*

High risk without mitigations – easy attack possible

Corresponding theoretical bound from DP



Removing PII not sufficient to mitigate re-identification risk in this scenario (implementation dependent)

Combining DP and PII scrubbing effectively eliminates risk of re-identification (with loose DP bounds)

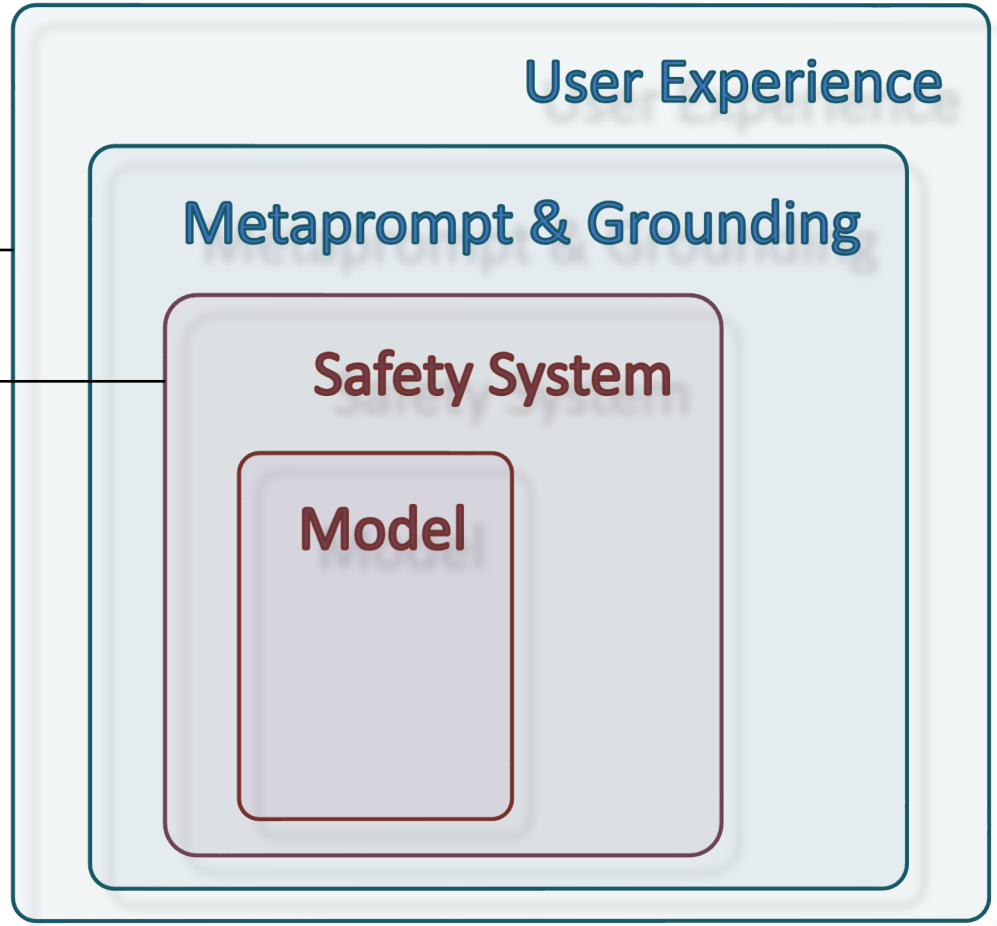
Trustworthy: *Responsible*

- Requires range of mitigation layers

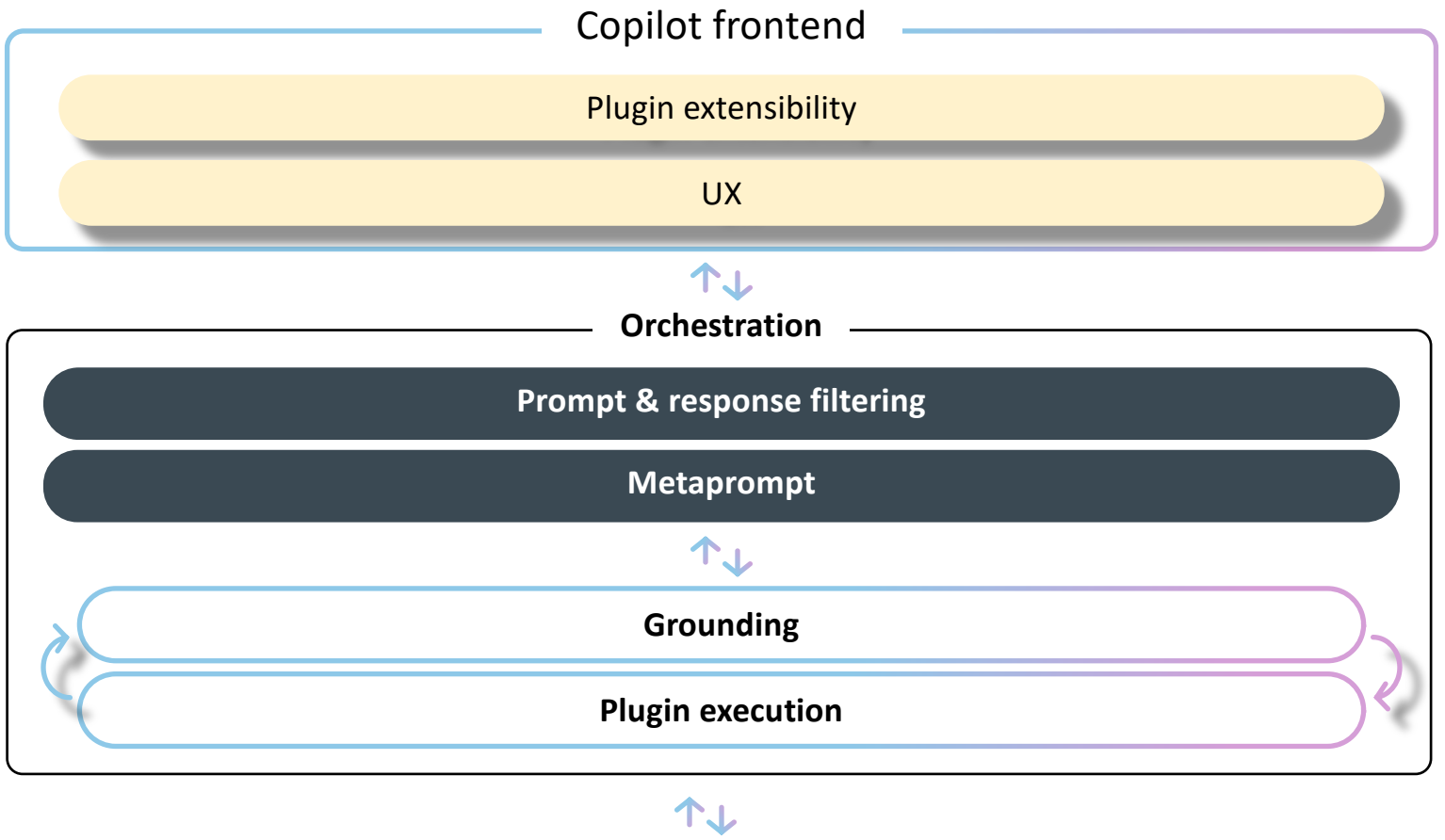
Application

Platform

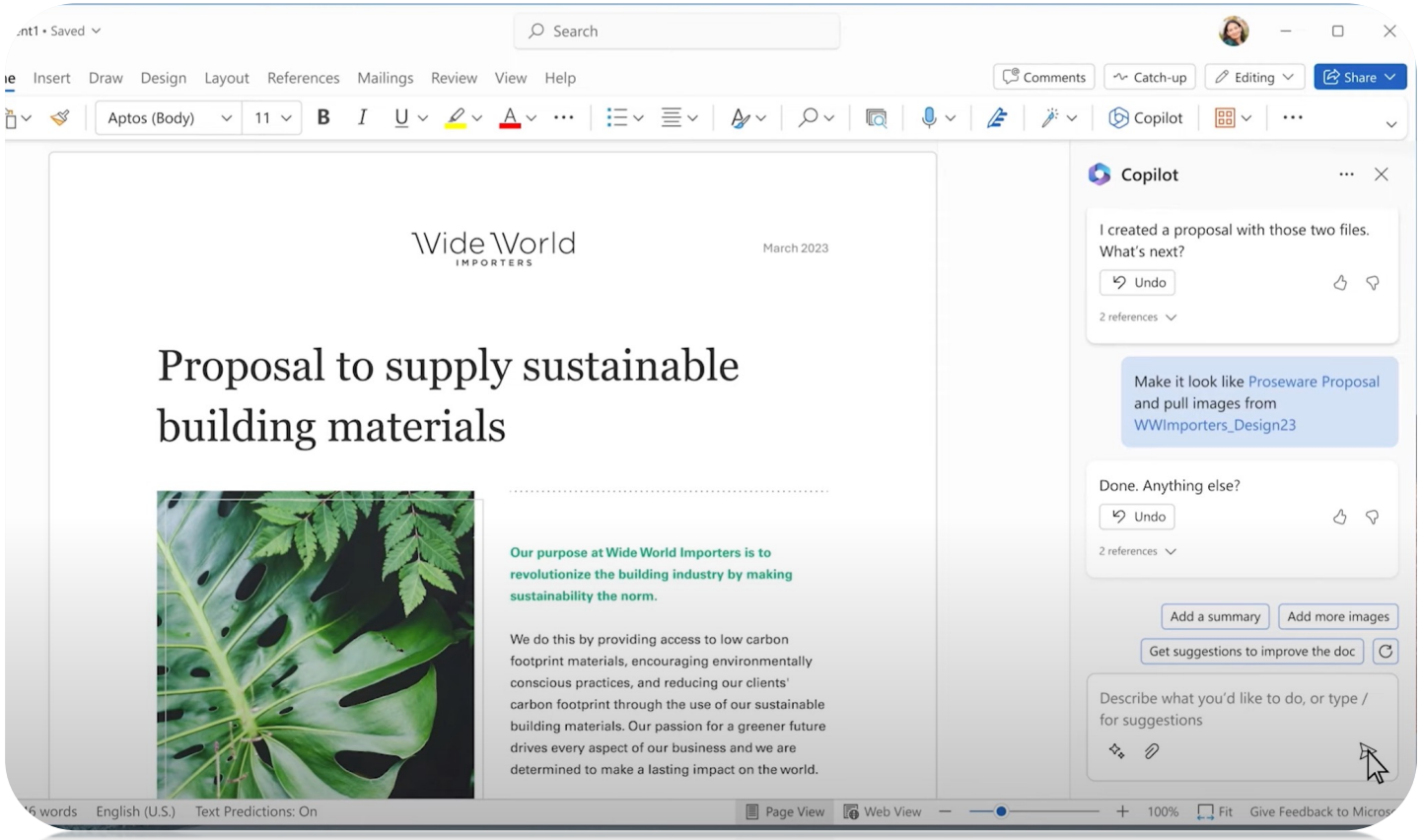
- LLMs foreground new challenges
  - Hallucination and errors
  - Jailbreaks and prompt injection
  - Harmful content and code
  - Manipulation, human-like behavior
- Enterprise v. consumer context



Embedded in Existing Workflows: *Copilot*



Embedded in Existing Workflows: *ODSL*



Knowledge in Conversations

From documents to dialogues

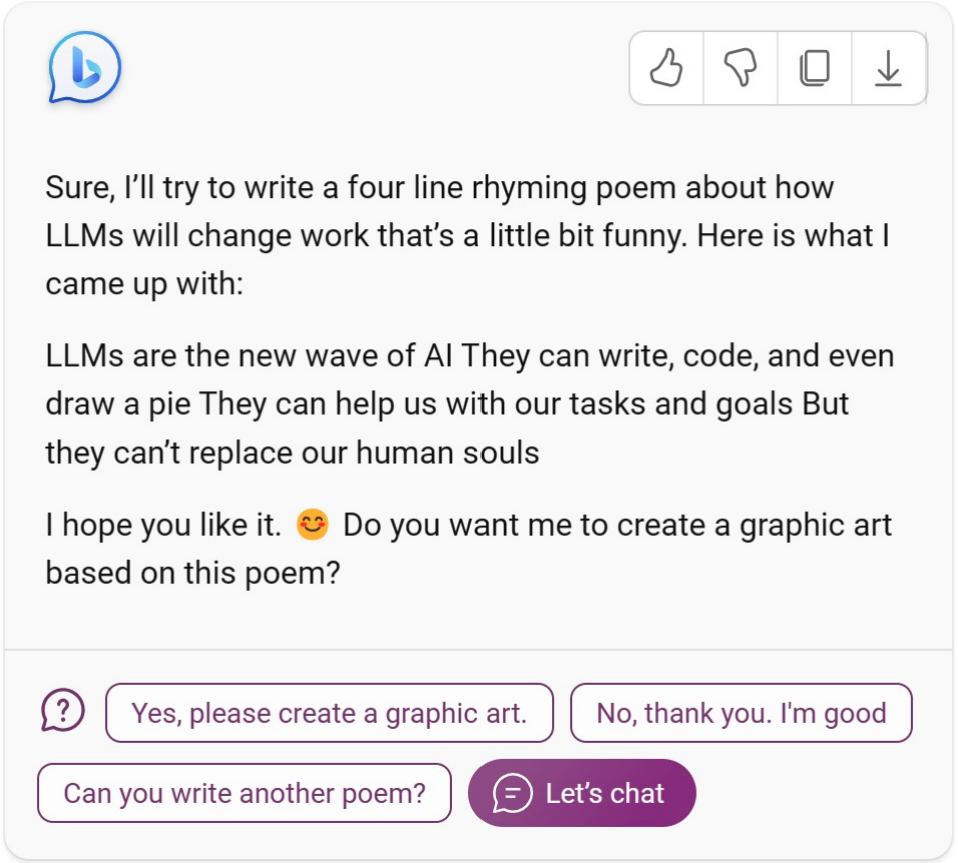
With other people

With your computer

rounded

Chat Log Analysis

- We expect unexpected uses
  - People can: Learn, find, compare, write, solve, plan, produce, …
- Chat logs teach us about these uses
  - People create complex prompts, iterate, learn as they go
  - Chat logs contain rich data, including feedback, knowledge, style, process
- LLMs can help with the analysis
  - Make it possible to extract meaning from billons of conversations
  - Important to do this analysis in a privacy preserving manner



Chat Log Analysis: *Prompting Do's*

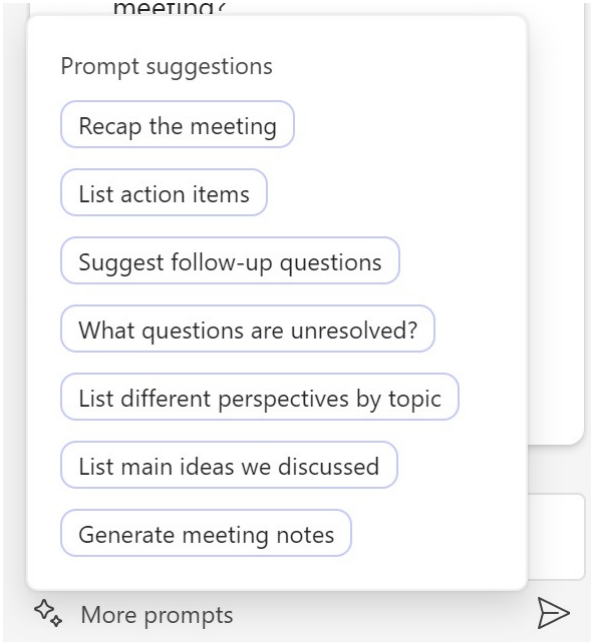
- Do provide **clear and specific instructions** to the AI writing assistant, such as the **topic**, the **purpose**, the **tone**, and the **length of the writing**. For example, "*Write a short summary of this article in a formal tone*." This will help the AI understand your writing goal and generate relevant and coherent outputs
- Do **give feedback to the AI writing assistant** when it produces good or bad outputs, so that it can learn from your preferences and improve its performance. For example, "*This sentence is too vague, can you be more specific?*" or "*Thank you, this is much better. Can you please add a sentence that summarizes the main point of the paragraph?*"
- Do **use polite and respectful language** when communicating with the AI writing assistant, such as saying "*please*", "*thank you*", and "*I appreciate your help*". This will help create a positive and collaborative atmosphere and improve the AI's responsiveness and performance.
- Do **use the AI writing assistant as a source of inspiration and guidance**, **not as a replacement** for your own writing. For example, you can use the AI to generate some ideas, sentences, or paragraphs, but you should also use your own creativity, logic, and judgment to edit, revise, and polish the writing.
- Do use questions to **elicit more information or feedback from the AI**. For example, "*Who is the main character of the story?*" or "*How can I improve this sentence?*" This will help the AI understand your needs and preferences better and provide more relevant and helpful responses.

Chat Log Analysis: *Prompting Don’ts*

- Don't use **vague or ambiguous instructions or feedback** to the AI writing assistant, such as "write something" or "make it better". This will confuse the AI and result in poor or irrelevant outputs. Instead, try to be as specific and detailed as possible.
- Don't expect the AI writing assistant to **write perfect or original outputs every time**. The AI writing assistant is a tool that can help you with your writing tasks, but it is not a substitute for your own creativity, critical thinking, and editing skills. You should always review and revise the AI's outputs before using them for your purposes.
- Don't use the AI writing assistant for **inappropriate or unethical purposes**, such as writing fake reviews, misleading information, or harmful content. The AI writing assistant is not responsible for the content or the consequences of your writing, and you should respect the laws, the rules, and the rights of others.
- Don't use **slang, jargon, or informal language when writing to the AI writing assistant**, unless you specify the tone and the audience of your writing. This may make the AI produce outputs that are inappropriate or unprofessional for your intended purpose. For example, "Write a dope intro bout how meditating is lit af." instead of "Write a short introduction paragraph for a blog post about the benefits of meditation.”
- Don't **interrupt or change the topic of the conversation abruptly**, as this might disrupt the flow or coherence of the writing process. For example, don't ask the AI to write a story, then a usage manual, then a story again, without finishing or closing the previous task.

Prompt Support: *Creating the LLM "Ribbon"*

- Support learning
  - Identify and surface tips in situ
- Support sophisticated prompting
  - Identifying good templates
  - Template recommendation
- Evaluating model output is key
  - Especially across different contexts (e.g., model versions)
- Opening up the metaprompt
  - Capture personalization and style
  - Enable goal-directed AI



Lead Like a Scientist

Develop and test hypotheses
Build on the state-of-the-art
Validate and debate
Consider the externalities

Thank you:

Maxamed Axmed, Chetan Bansal, Caitlin Cummings, Brent Hecht, Daniel Jones, Morris Kabuage, Victor Rühle, Robert Sim, Rujia Wang, Longqi Yang, Lukas Wutschitz, and the Calliope Team