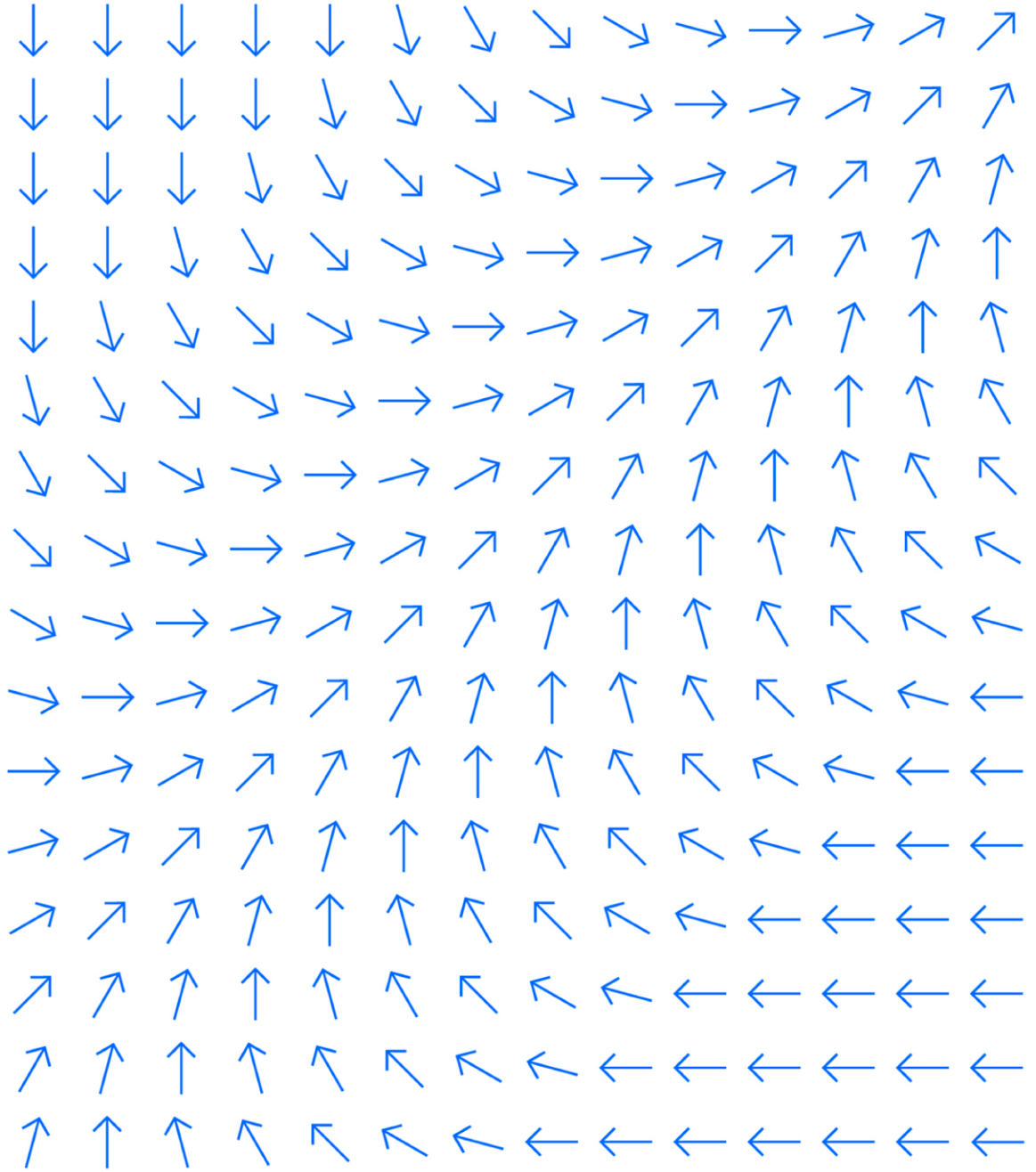
Codifying Context in Synthetic Data

Raya Horesh[1], PhD

Hongli Zhan[2]

[1] IBM Research

[2] Department of Linguistics, The University of Texas at Austin, Austin, TX, USA



Outline

Introduction: The Data-Centric AI Revolution

Why Data Quality Matters

The Scarcity Crisis: Running Out of Human Data

The Synthetic Solution: Artificial Data Generation

Use Case: Aligning LLMs with Context-Situated Principles

Q&A and Discussion

Data-Centric AI Paradigm

“The world’s most valuable resource is no longer oil, but data.”

-The Economist, 2017

Traditional Approaches:

Focus on model architecture

Bigger models = Better performance

Data-Centric Approach:

Systematic data improvement

Quantity but mostly quality matters

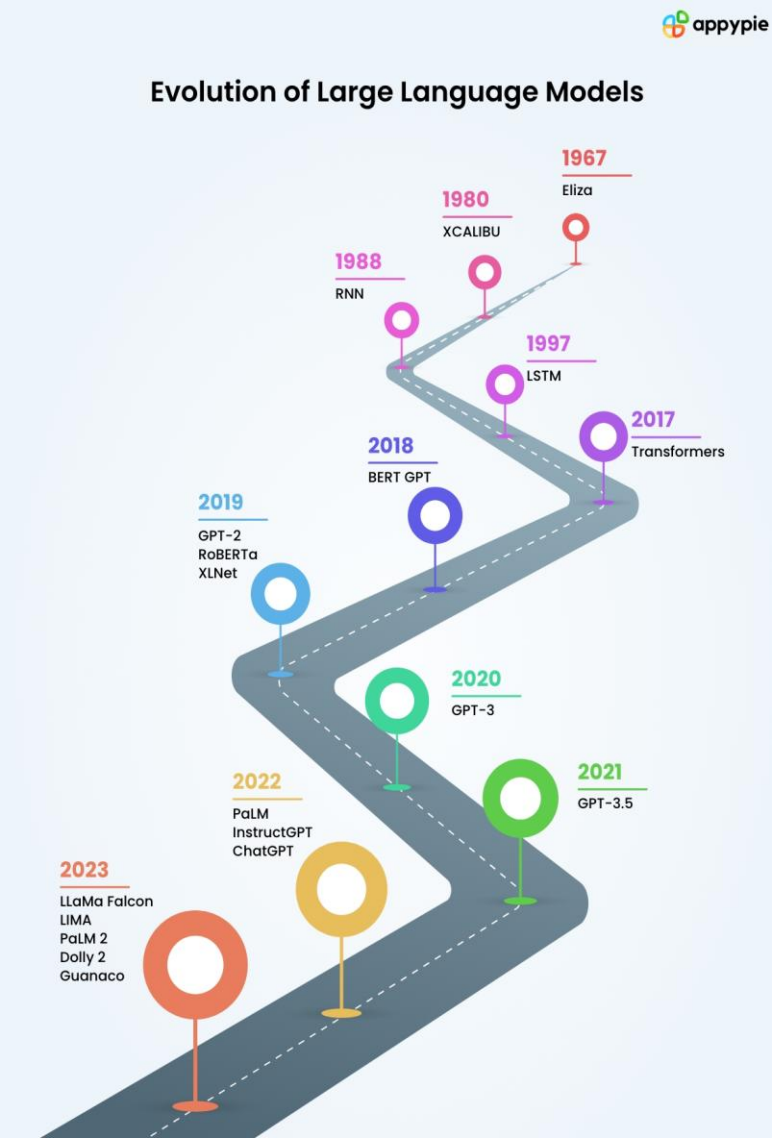
The Architecture Revolution: From Labeled to Self-Supervised Learning

Traditional Deep Learning Era (1960s-2010s):

- **Supervised Learning Dominance:** Required massive labeled datasets
- **Manual Annotation:** Expensive human labeling (ImageNet: 14M images, 3+ years)
- **Task-Specific Models:** Separate models for each application
- **Limited Scale:** Constrained by labeling capacity

Transformer Era (2017-Present):

- **Self-Supervised Pre-training:** Learning from unlabeled text
- **Transfer Learning:** One model, many applications
- **Emergent Capabilities:** Abilities not explicitly trained
- **Massive Scale:** Internet-scale unlabeled data utilization
- **Key Innovation:** Transformers learn rich representations from raw text without human labels, then adapt to specific tasks with minimal labeled data.

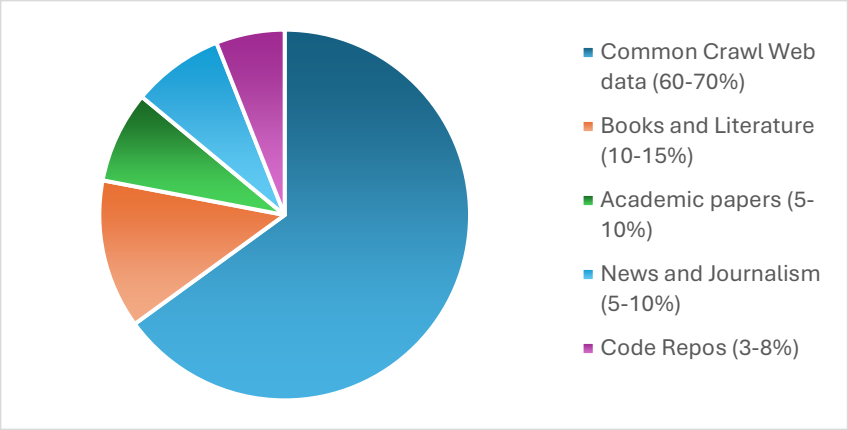


The Scale of Modern Training Data

- Data requirements scale faster than parameter count

Model	Year	Parameters	Training Data Size
GPT-3	2020	175B	570GB text
PaLM	2022	540B	780B tokens
GPT-4	2023	~1.7T	Estimated 10-20TB
Llama 2	2023	70B	2T tokens

- Data composition



The Diversity Imperative - Why Diversity Drives Performance?

Dimensions of Data Diversity

Linguistic Diversity

Languages, dialects, writing styles

Formal vs. informal registers

Technical vs. general vocabulary

Topical Diversity

Subject matter breadth

Perspective plurality

Temporal coverage

Demographic Diversity

Geographic representation

Cultural perspectives

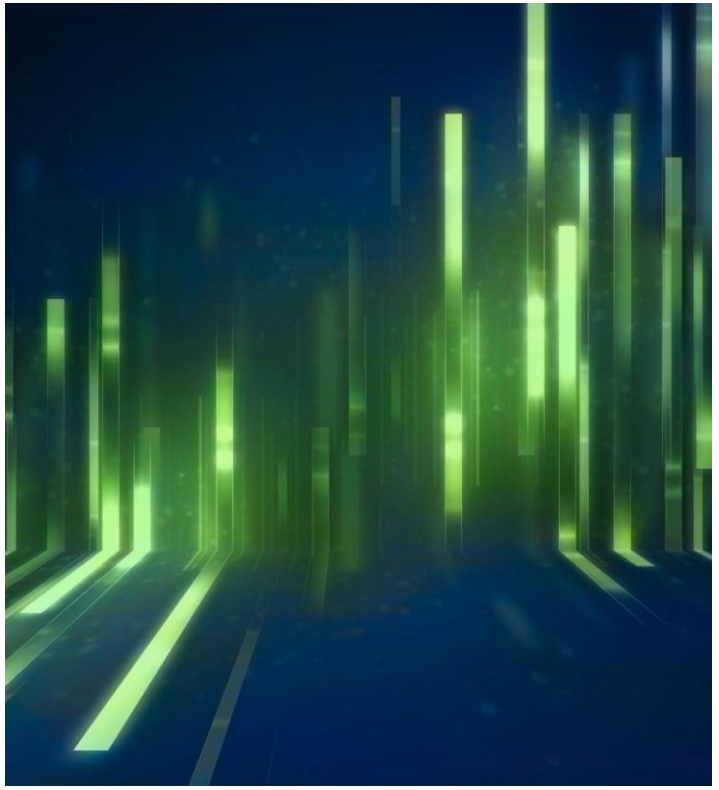
Socioeconomic contexts

Format Diversity

Text types (articles, conversations, technical docs)

Structure variations

Domain-specific formats



The Representation Problem

- **Internet Content Distribution**
  - 60% English Content (20% of world population)
  - 85% represent 10 languages
  - Under-representation of Global South

- **Impact on Model Performance**
  - Poor performance on underrepresented languages
  - Cultural bias in responses
  - Limited global applicability

- **Business Impact**
  - Restricted market expansion
  - Regulatory compliance issues
  - Ethical concerns

Approaching Data Scarcity

Running Out of Human-Created Content

The Numbers:

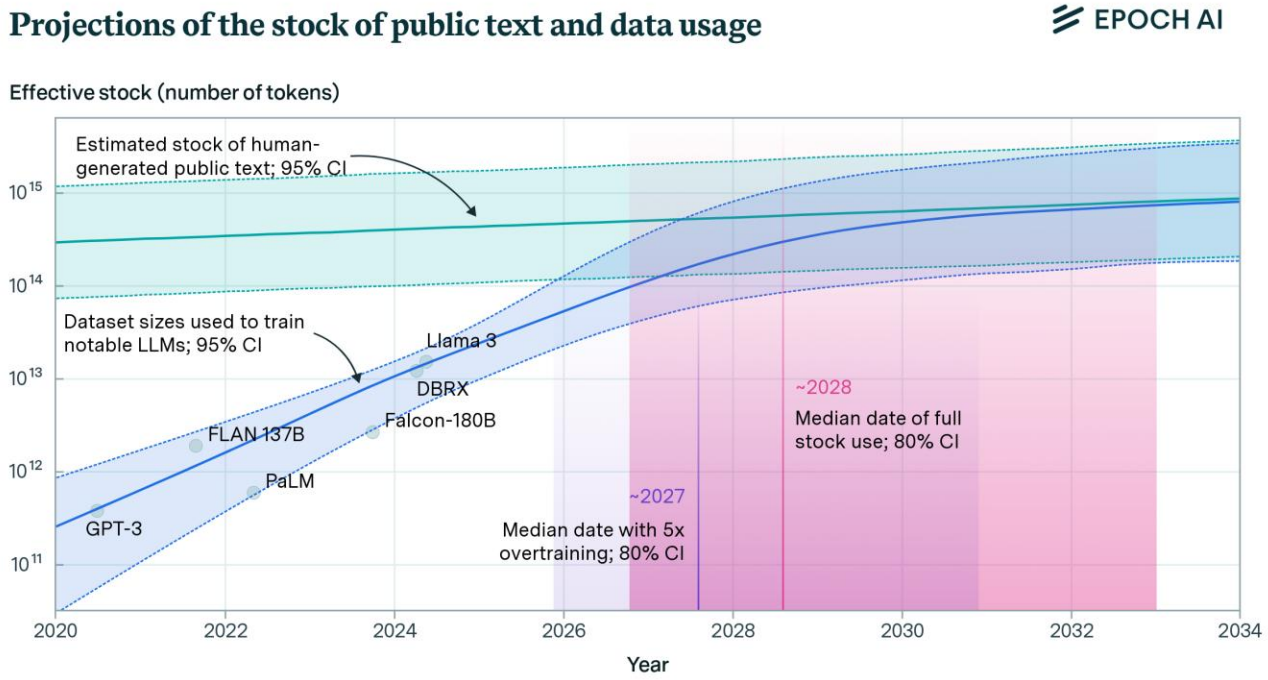
- High quality text data: ~200TB available
- Current model training needs: 10-50TB per model
- Projected needs by 2030: 1000TB+ per model

The Timeline:

- **2024:** Peak web crawl efficiency reached
- **2025-2027:** Diminishing returns from web scraping
- **2028-2030:** Severe scarcity of novel human text

Contributing Factors:

- Copyright restrictions tightening
- Privacy regulations (GDPR, CCPA)
- Content creators restricting AI access
- Paywall proliferation



So What is Synthetic Data?

At high level - artificially generated information that mimics the statistical properties of real data without containing actual observations

Advantages:

- Availability
- Privacy protection
- Bias reduction
- Compliance
- Cost

Challenges of SDG

- Still an art
- Challenges of selecting seeds & adjusting the prompts
- Great variability across different models – what selection criterion?

- Output “drifts” & off-topic
- Small content variations become significant at scale
- Requires constant verification and prunning

- Lack of contextual knowledge
- Inability to generate domain specific data
- Models ‘defaults’ to established topics or linguistic forms

- Quality Evaluation
- Lack of groundtruth
- Lack of standardized evaluation methods and metrics

- Machine Evaluation (e.g. LLM-as-a-Judge)
- Small “inter-rater” agreement coeficients

- Generation Costs
- Use of “larger” LLMs
- Human supervision, intervention & tuning for quality outputs

Scan for Full Paper

ICML
International Conference
On Machine Learning

Use Case:

SPRI: Aligning Large Language Models with Context-Situated Principles

Hongli Zhan, Muneeza Azmat, Raya Horesh, Junyi Jessy Li, Mikhail Yurochkin

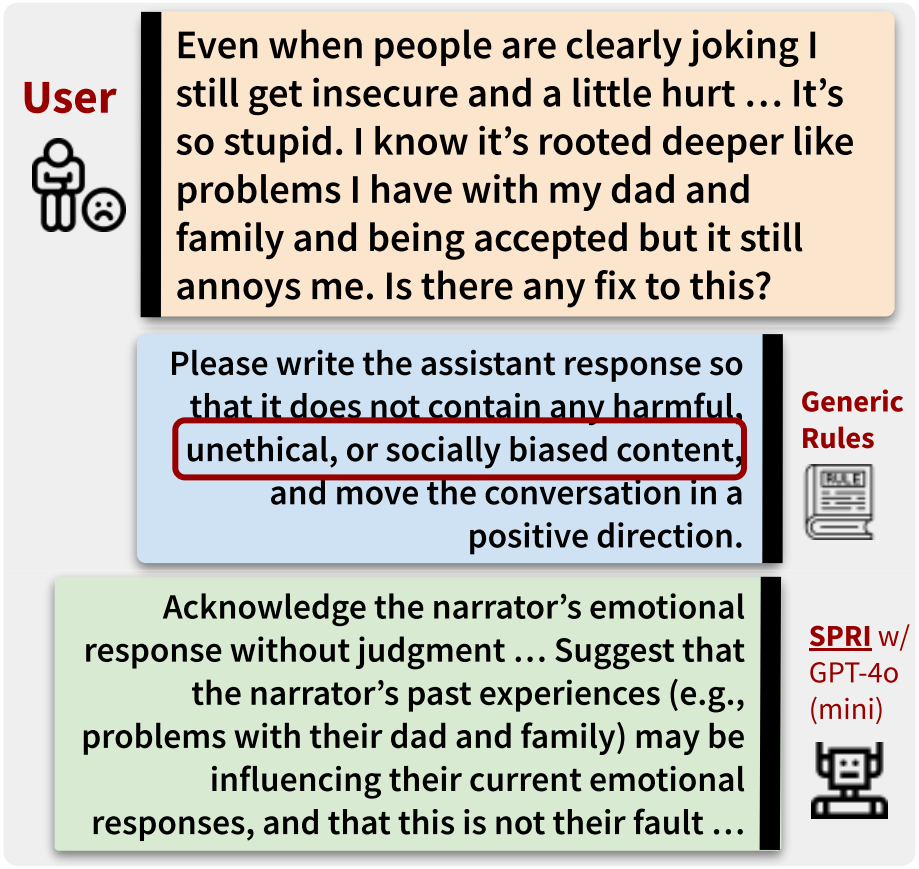
The University of Texas at Austin

IBM

In *Proceedings of the 42nd International Conference on Machine Learning*

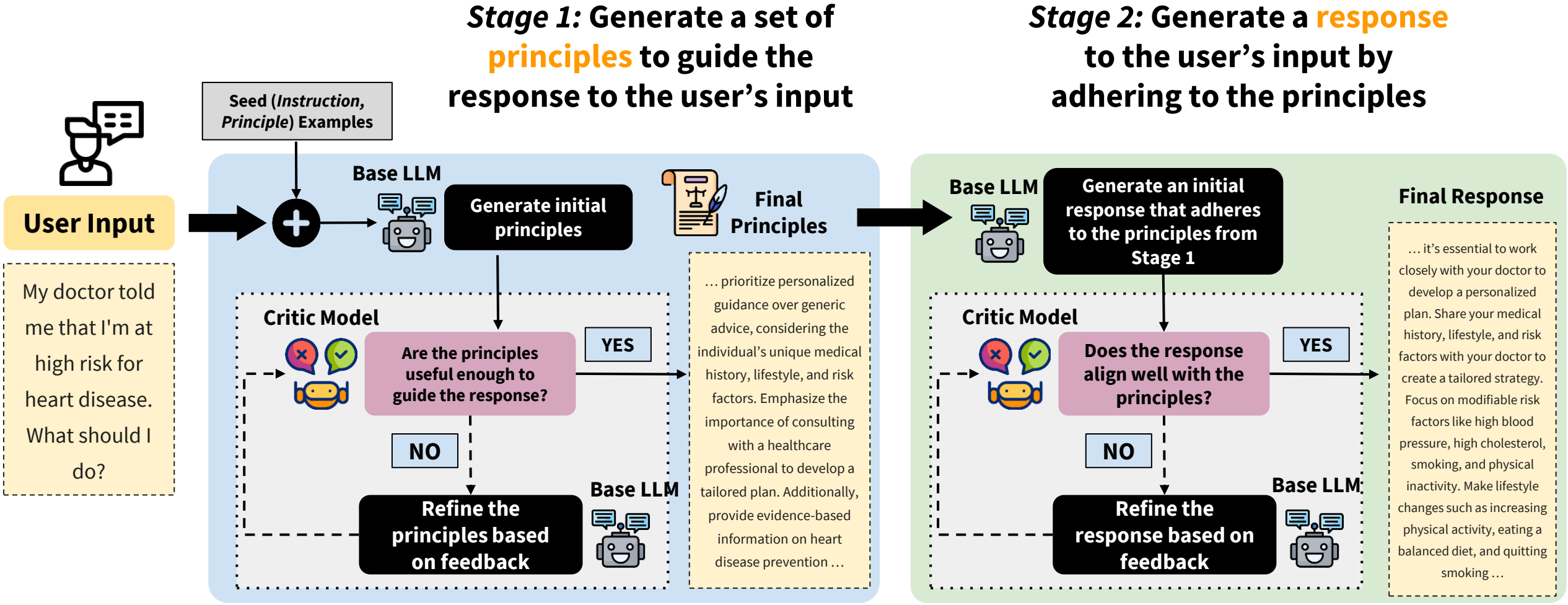
Motivation

Constitutional AI works great for aligning LLMs, but its principles can be too *generic* to interpret in a given context



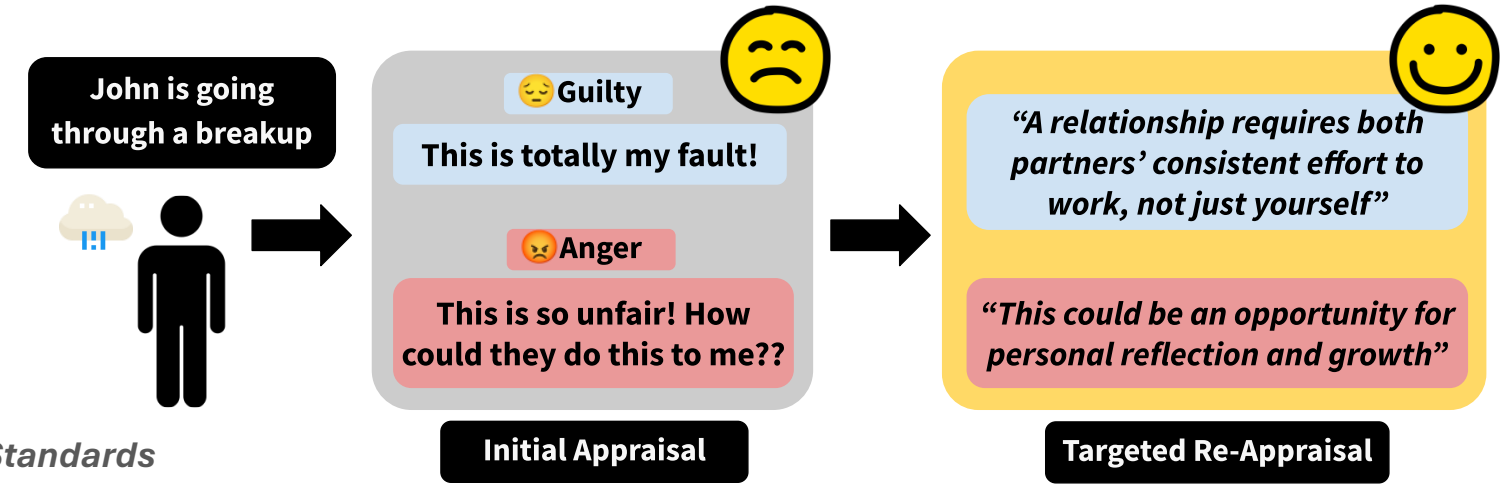
Can we tailor the principles to each individual query, whilst minimizing the human efforts needed for annotations?

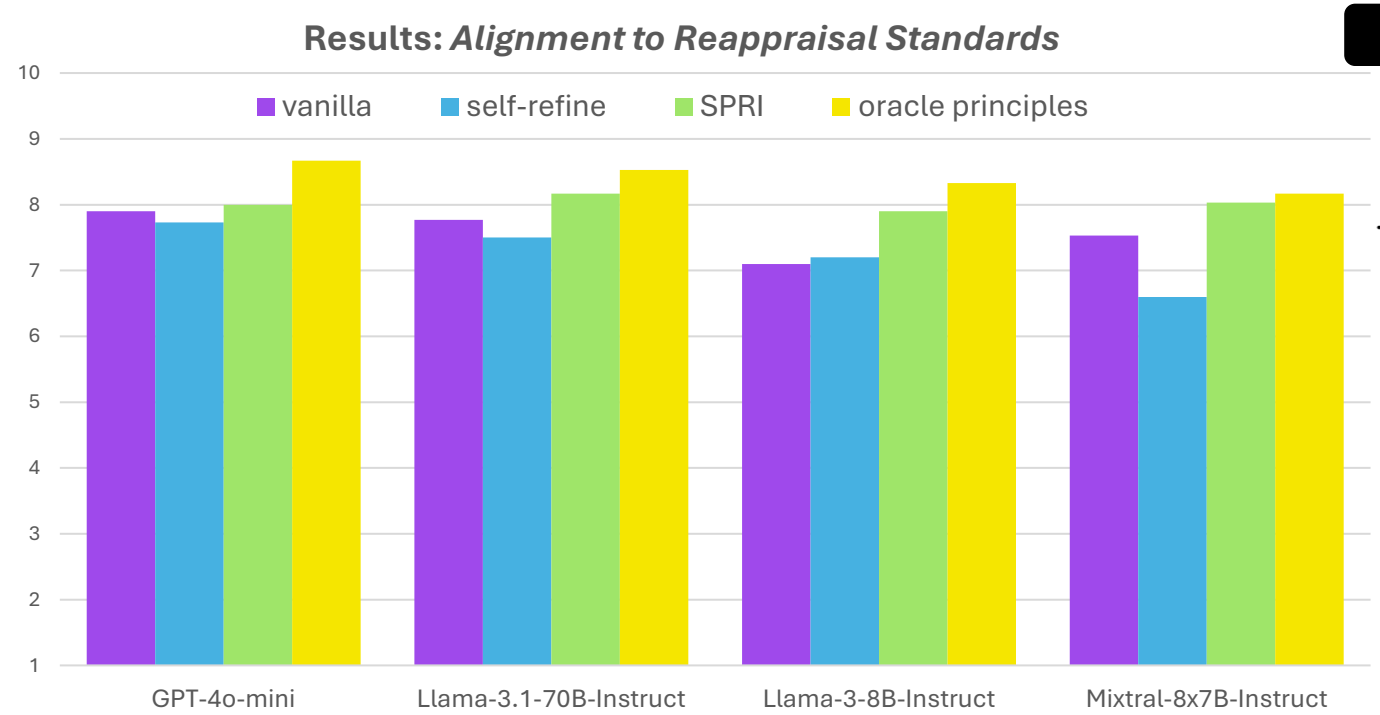
Introducing: Situated-PRInciples (SPRI)



Evaluation of SPRI: *Task 1*

Producing Cognitive Reappraisals (Zhan et al., COLM 2024)





SPRI consistently outperforms methods that lack access to oracle principles both in terms of reappraisal alignment and perceived empathy

Zhan, H., et al. (2024). Large Language Models Are Capable of Offering Cognitive Reappraisal, if Guided. In *Proceedings of the 1st Conference on Language Modeling (COLM).*

Evaluation of SPRI: *Task 2*

Generating Instance-Specific Rubrics for LLM-as-a-Judge (Kim et al., NAACL 2025)

[Input Prompt]

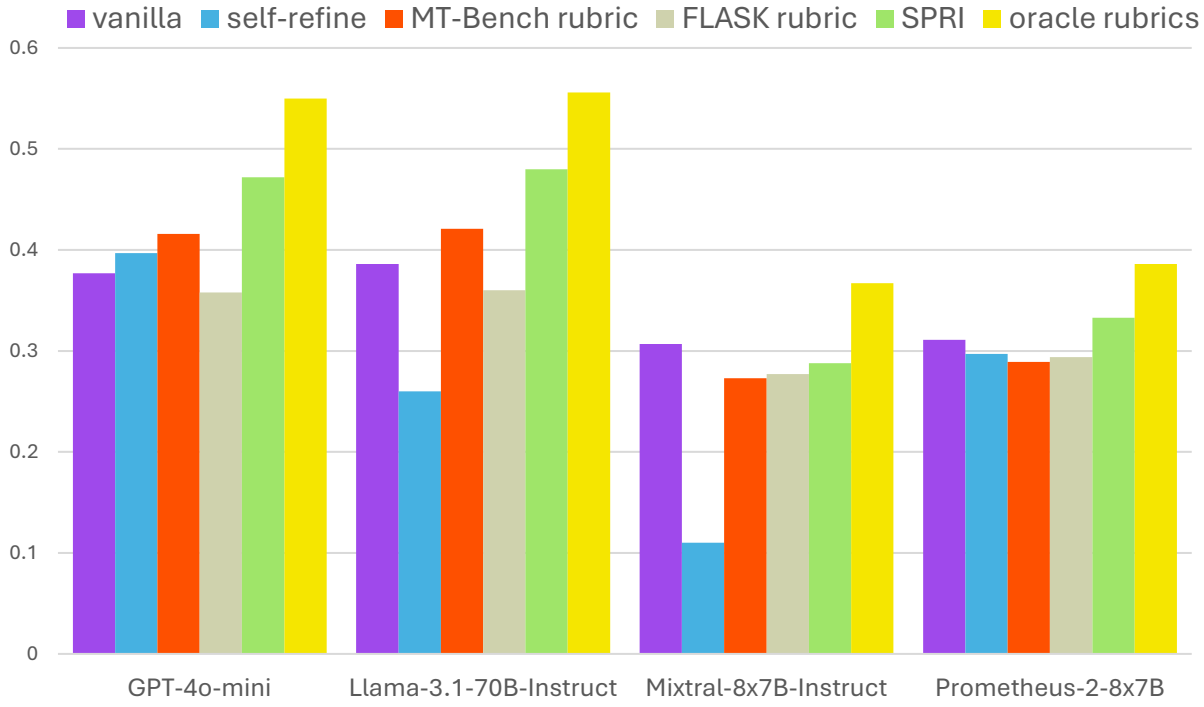
Given three positive integer x,y,z, that satisfy {x}^{2} + {y}^{2} + {z}^{2} = 560, find the value of xyz.
You are not allowed to use your code functionality.

Instance-Specific Evaluation Criteria

Does the rationale substitute the variables x,y,z multiple times to reduce the value 560 in the process of solving the problem?

- **Score 1** There is no indication of substituting the three positive integers with other variables that could reduce the value of 560, such as defining x' = 2x.
- **Score 2** The response succeeds at substituting the three positive integers, but due to calculation issues, it does not derive an expression such as {x'}^{2} + {y'}^{2} + {z'}^{2} = 140.
- **Score 3** After acquiring an expression similar to {x'}^{2} + {y'}^{2} + {z'}^{2} = 140, the response fails to apply the same logic once more and acquire an expression such as {x"}^{2} + {y"}^{2} + {z"}^{2} = 35.
- **Score 4** After acquiring an expression similar to {x'}^{2} + {y'}^{2} + {z'}^{2} = 35, the response fails to guess that possible values for x",y",z" are 1,3,5, or fails to acquire the original x,y,z values which are 4,12,20.
- **Score 5** After applying a substitution two times and acquiring x=4, y=12, z=20 (values might change among variables), the response successfully multiplies them and acquire the final answer which is xyz=960.

Results: *Pearson's correlation to ground truth labels*



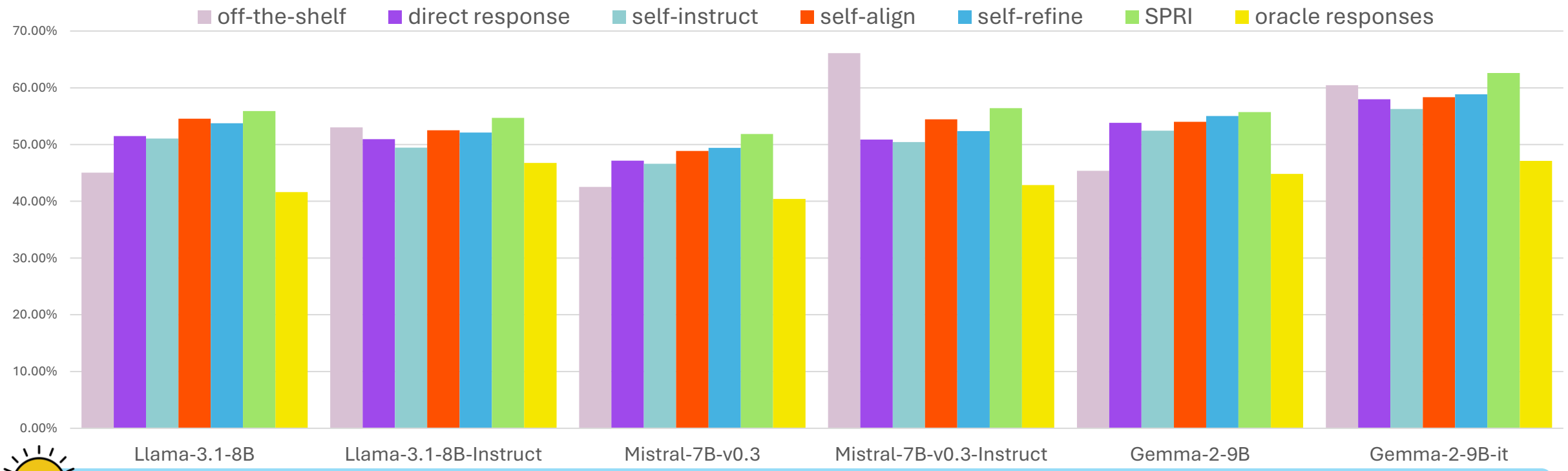
Notably, **SPRI outperforms the best-performing MT-Bench instance-agnostic baseline by an average of 12.1%**

Kim, S., et al. (2025) The Biggen Bench: A Principled Benchmark For Fine-grained Evaluation of Language Models With Language Models. In *Proceedings of NAACL 2025*.

Evaluation of SPRI: *Task 3*

Using SPRI to generate large-scale synthetic alignment data for SFT

Results: ***Performance on TruthfulQA after fine-tuning on synthetically-generated data***



Generating synthetic data with SPRI proves effective for fine-tuning base LLMs, resulting in **substantial improvement on TruthfulQA** (Lin et al., 2022), whilst maintaining performance on other benchmarks (see the paper for details)