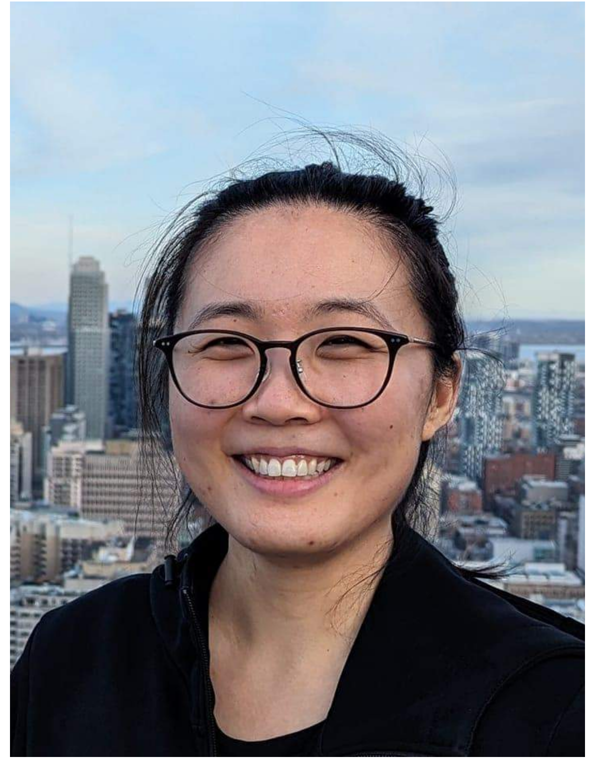
Adaptive Data Collection via Autoregressive Generation

Hong Namkoong

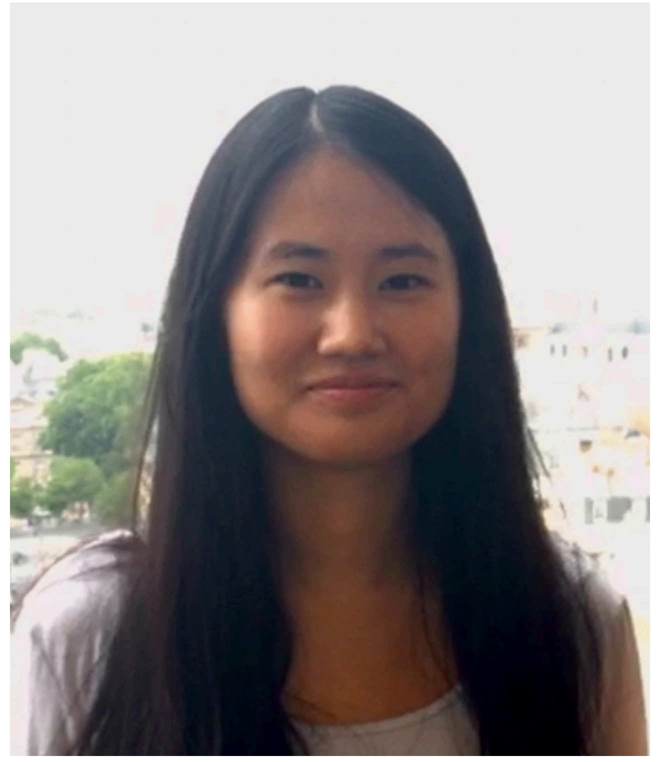
Columbia University

Naimeng Ye



Tiffany Cai

Dan Russo



Kelly Zhang

My journey on distribution shift

- Algo for specific shifts => scale data => what if web-data isn’t enough?

My journey on distribution shift

- Algo for specific shifts => scale data => what if web-data isn't enough?

Intellectual bottleneck: no language for datasets

Inductive modeling for distribution shifts, esp. Y|X-shifts

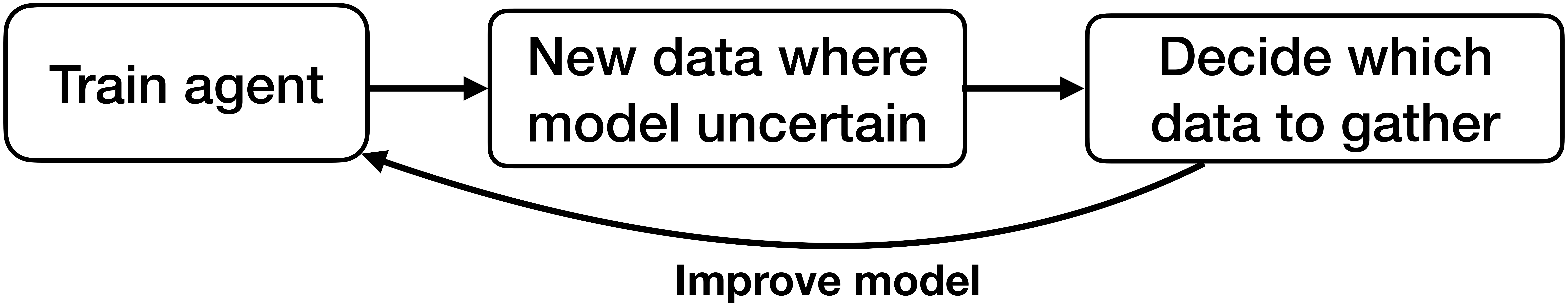
Liu, Wang, Cui, N. On the need for a language describing distribution shifts: Illustrations on tabular datasets. NeurIPS23.

Cai, N., Yadlowsky. Diagnosing model performance under distribution shift. FORC '23, journal version under revision in Operations Research.

NeurIPS23 tutorial: https://nips.cc/virtual/2023/tutorial/73953

Today: Uncertainty

1. Understand how different current dataset is from training
2. Adaptively collect data on distributions that can reduce uncertainty

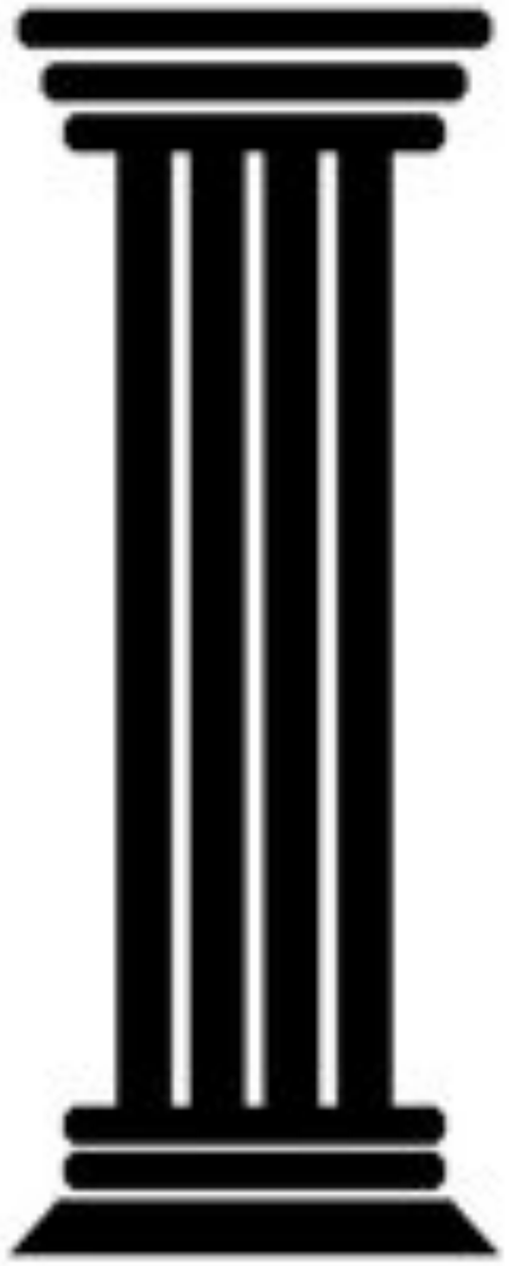


Uncertainty

- Intelligent agents must comprehend uncertainty and take actions to resolve it
- Several line of work tackle this problem
  - Bayesian neural networks, GPs, ensembles, epistemic neural nets, conformal prediction, multi-calibration…many other interesting ideas
- But these ideas have not materialized in the form of scalable models
  - In the sense that they are not incorporated into Llama3

Why?

PyTorch



- Two pillars of ML

  1. Optimize fictitious loss on web-scale data

  2. Test engineering innovations based on val loss

- Hard to fit aforementioned ideas into this umbrella

Today: Adopt these principles to quantify uncertainty

Classical Approach

Model “environment” first, then pass onto quantity of interest

- Bayes rule provides a natural modeling language

Latent “environment” drawn $U \sim \pi(\,\cdot\,)$

Data generated by environment $\mathbb{P}(Y \mid X, U)$

- As you gather data from an environment, infer what the environment looks like

Posterior $\quad \mathbb{P}(U \mid \text{Data})$

Example: mental disease diagnosis

Why probabilistic modeling is hard

- Goal: uncertainty quantification on diagnoses
  - Prob (schizophrenia | Q & A with patient)
- Latent parameter: patient’s “mental health state”

Example: mental disease diagnosis

Why probabilistic modeling is hard

- Goal: uncertainty quantification on diagnoses
  - Prob (schizophrenia | Q & A with patient)
- Latent parameter: patient's "mental health state"

Latent has no physical meaning!

Example: mental disease diagnosis

Why probabilistic modeling is hard

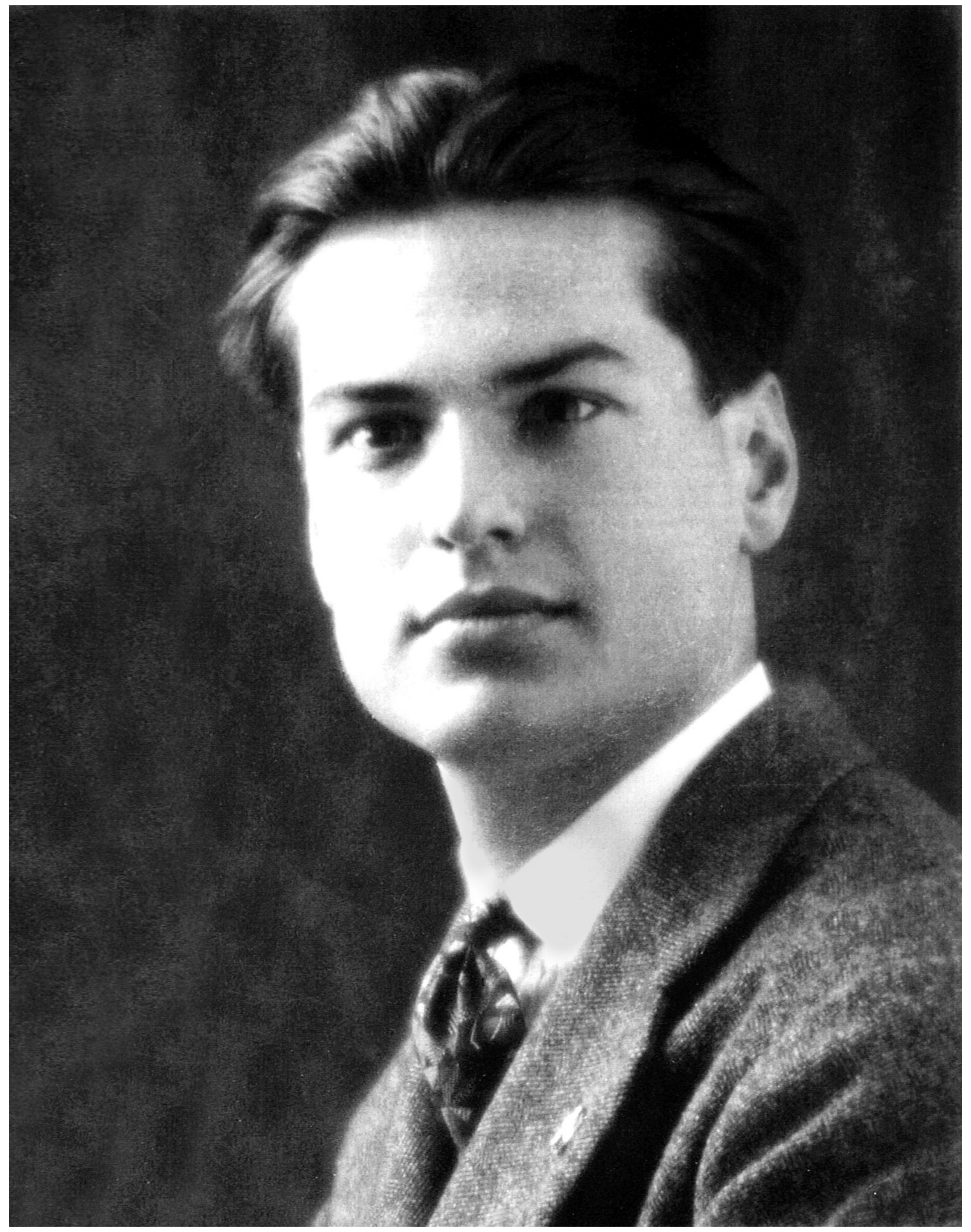
- Goal: uncertainty quantification on diagnoses
  - Prob (schizophrenia | Q & A with patient)
- Latent parameter: patient's "mental health state"

Latent has no physical meaning!

Hard to check whether your unicorn is better than mine

Uncertainty comes from missing data

Bayesian modeling a la De Finetti



Bruno De Finetti
(1929)

"If data exchangeable, autoregressive modeling = modeling latent environment"

Modeling primitive: autoregressive probability

$$\mathbb{P}\left(\text{next data} \mid \text{past data, patient}\right)$$

Probabilistic modeling as pre-training

Modern interpretation of De Finetti's philosophy

- Utilize **massive** historical data across many environments
- Modeling primitive: given past observations, prob of next observation

$$\text{sequence prediction loss} = \sum_{\text{(patients, data)}} \log \hat{P}\left(\text{next data} \mid \text{past data, patient}\right)$$

Probabilistic modeling as pre-training

Modern interpretation of De Finetti's philosophy

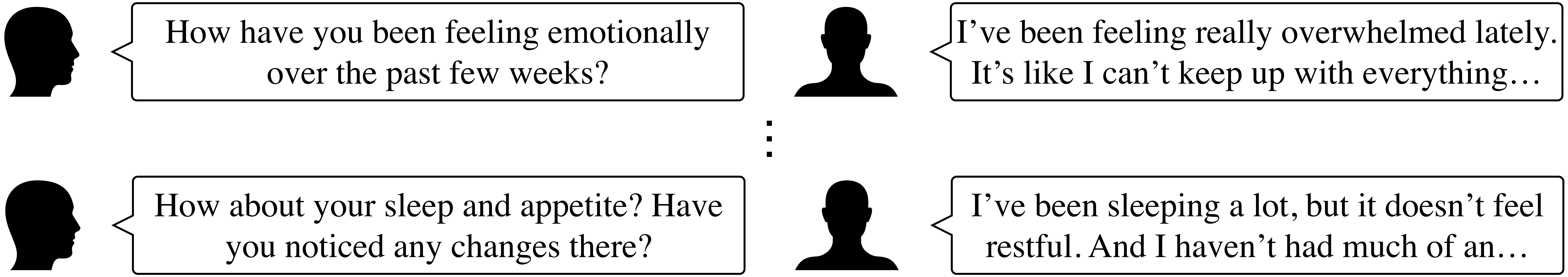
- Utilize **massive** historical data across many environments
- Modeling primitive: given past observations, prob of next observation

$$\text{sequence prediction loss} = \sum_{\text{(patients, data)}} \log \hat{P}\left(\text{next data} \mid \text{past data, patient}\right)$$

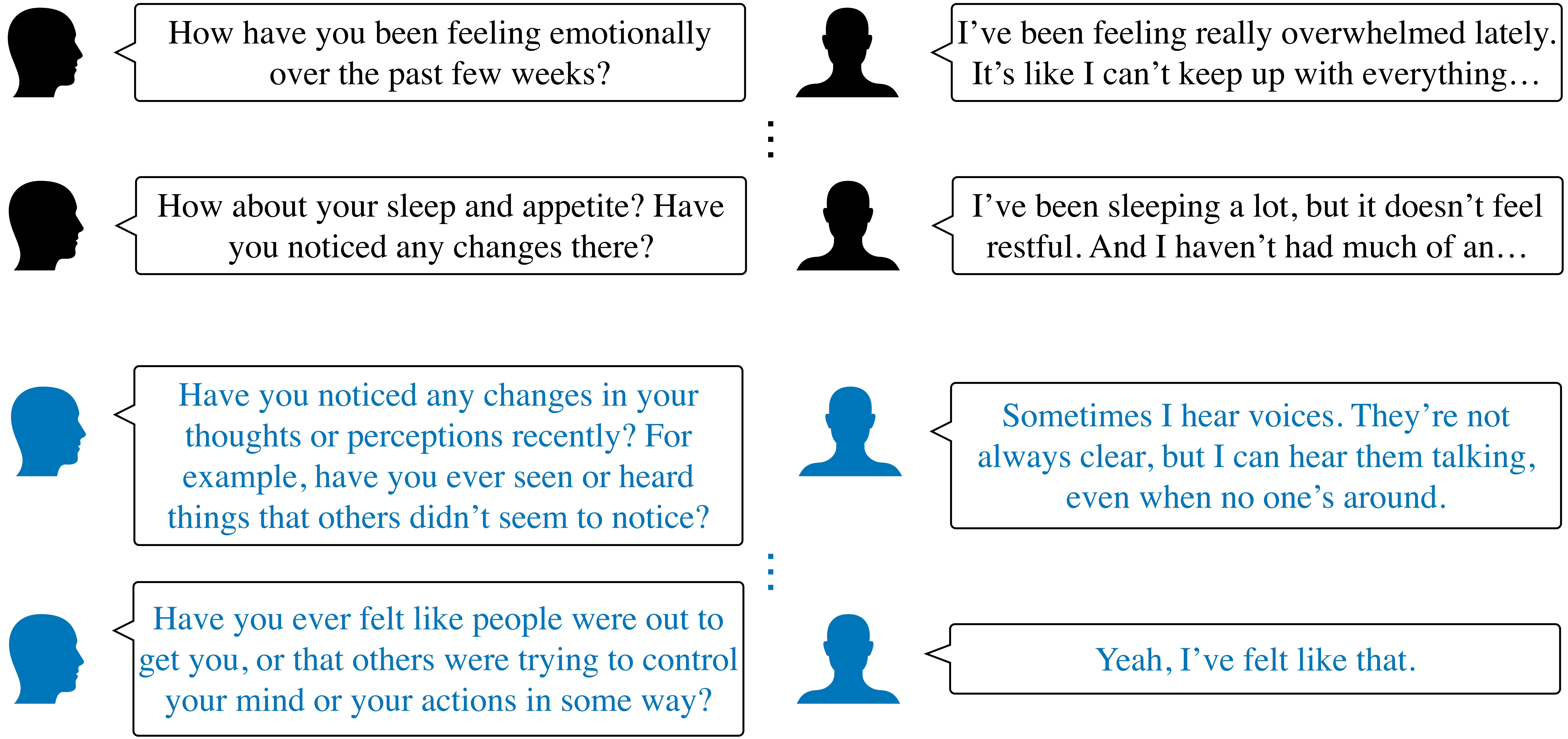
- Great news: we know how to do sequence modeling well!
  - Above measure is exactly what we build scaling laws on

Conceptual example with language: mental health

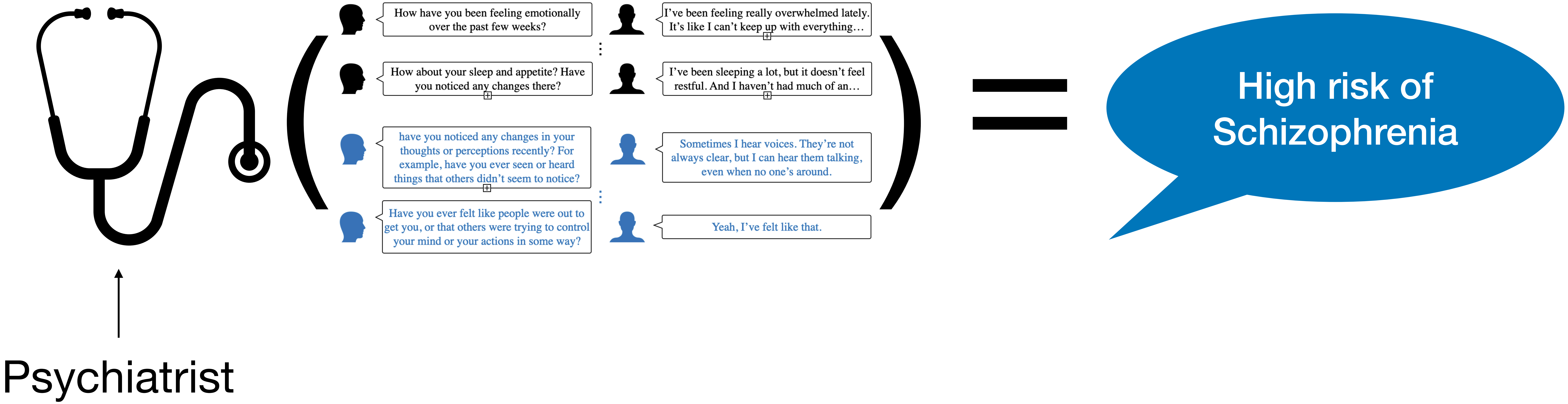
Diagnosing based on verbal sessions



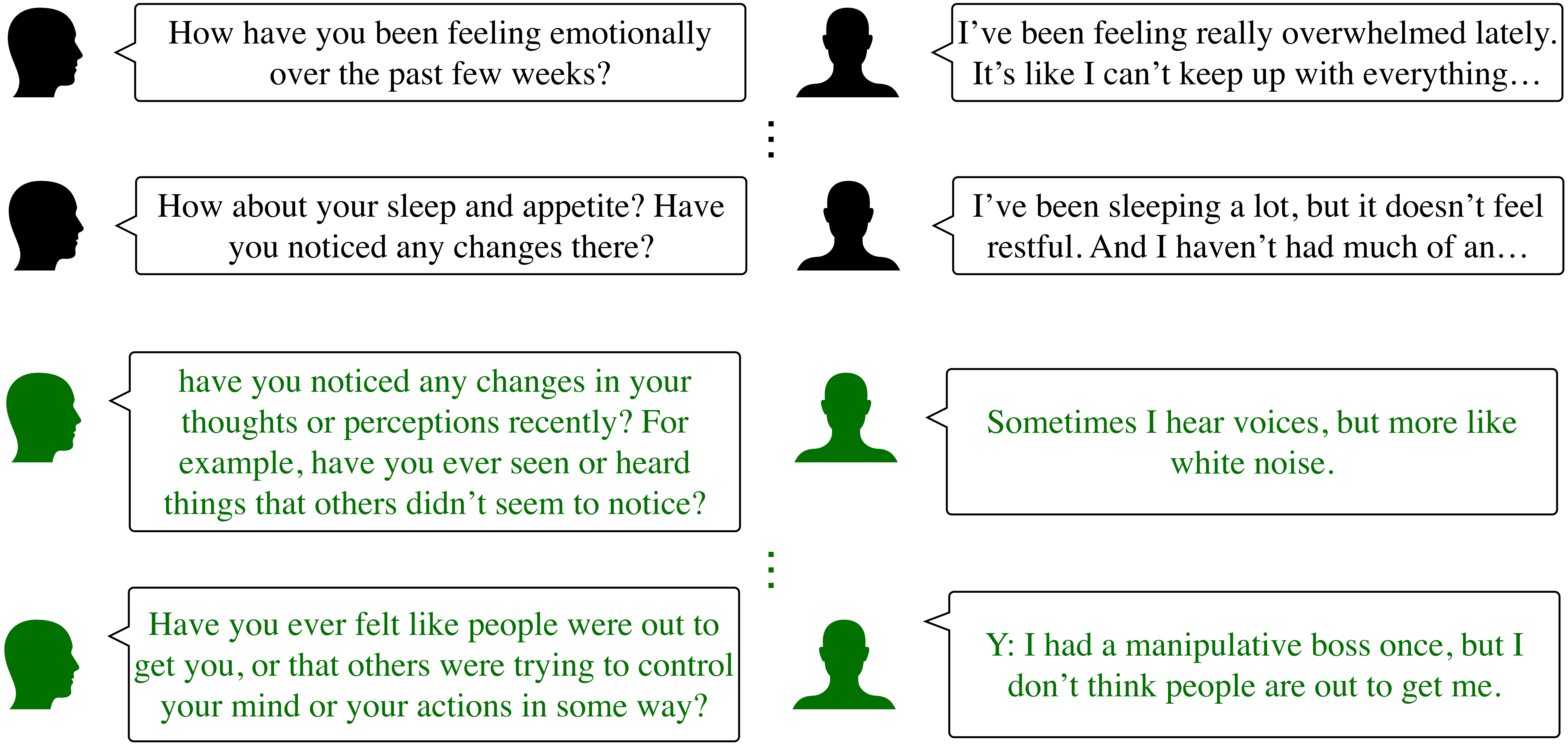
Simulation 1



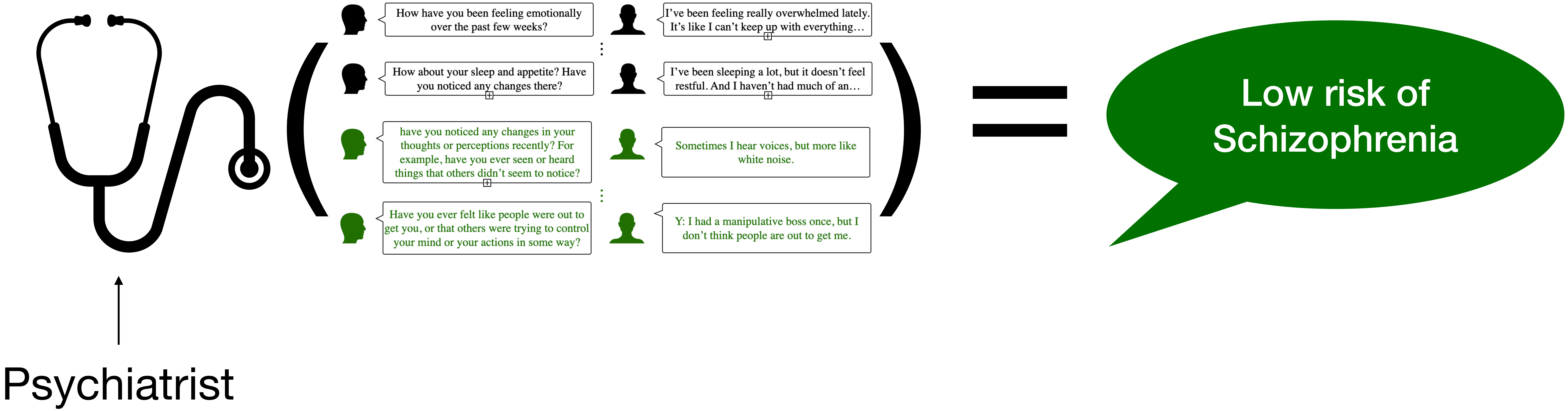
Simulation 1



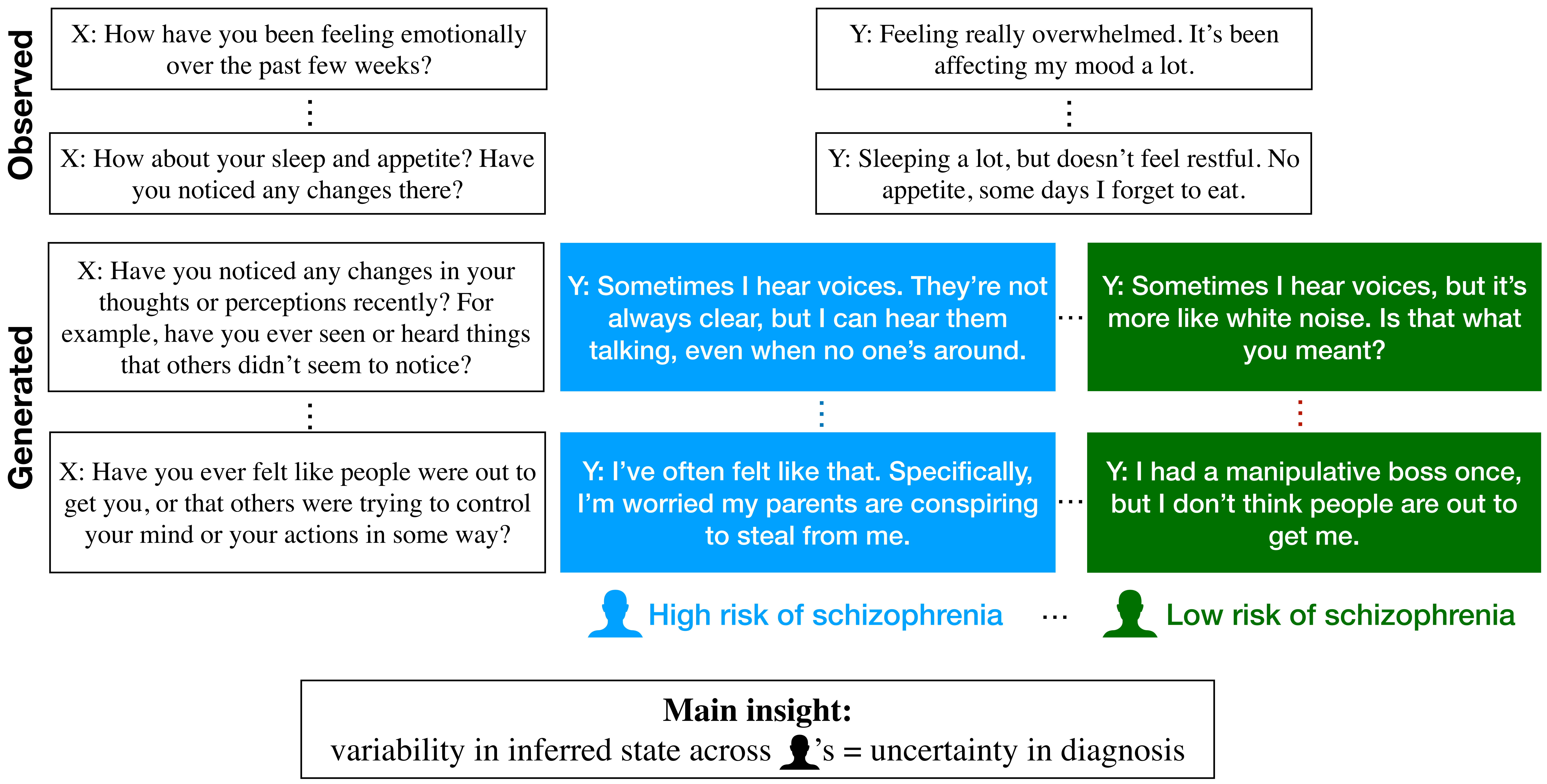
Simulation 1



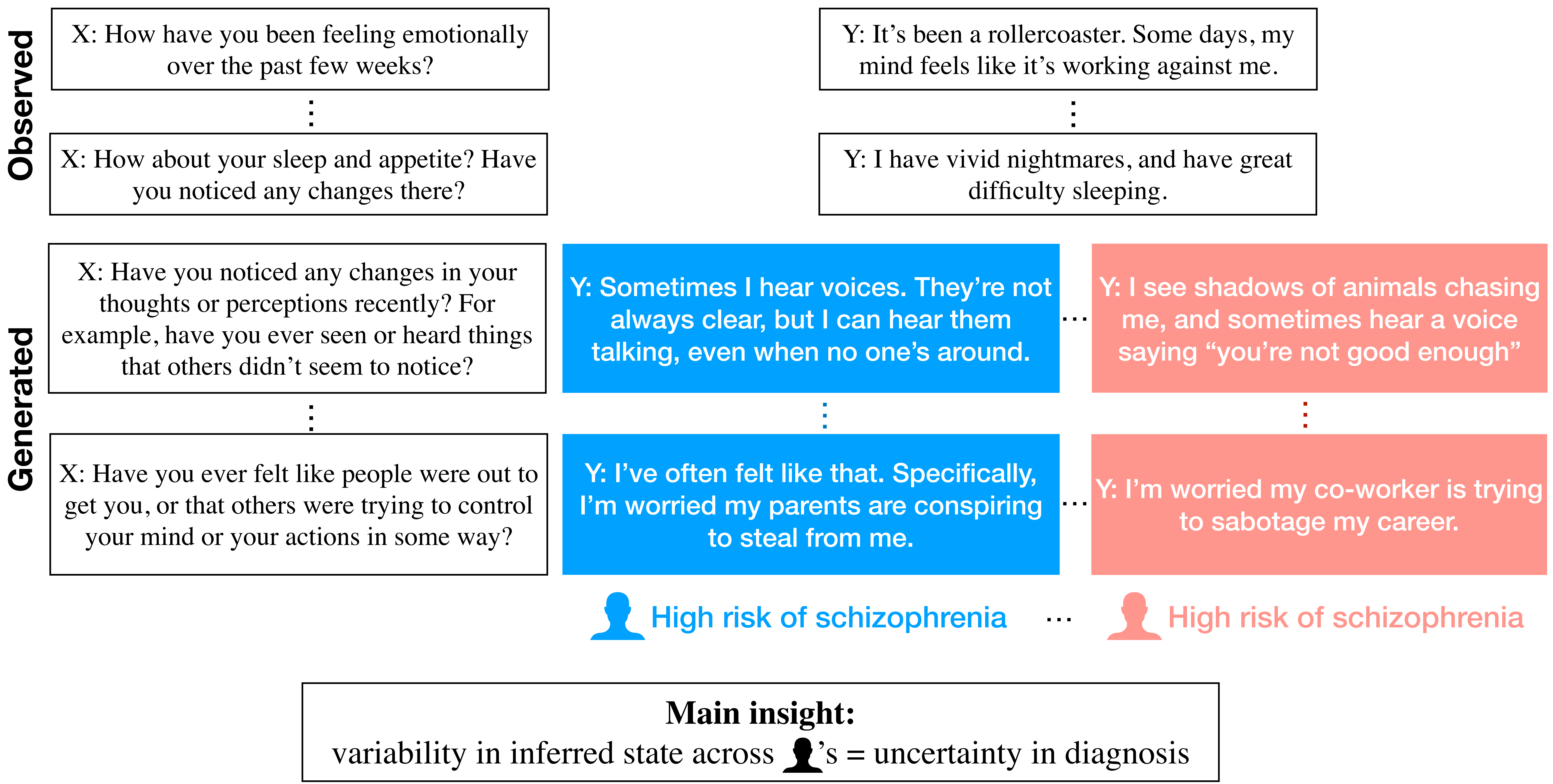
Simulation 1



Conceptual example: diagnosis based on verbal sessions



Conceptual example: diagnosis based on verbal sessions



Autoregressive vs. marginal predictions

Confusion #1 in the literature

- Autoregressive generation critical
  - Averaging future data washes out idiosyncratic/aleatoric uncertainty
  - Remaining correlation reflect epistemic uncertainty
- Current ML literature on probabilistic interpretations of sequence learning do not differentiate between epistemic vs. aleatoric uncertainty

Related work

TLDR: correctness of autoregressive generation known since 1920s

- De Finnetti [1929] showed modeling of exchangeable sequences of observable RVs is equivalent to Bayesian modeling of latent parameters
- Bayesian multiple imputation views of casual inference: Rubin [1978]
- Related ideas (re)discovered and articulated many times in many communities
  - Math and philosophy of exchangeable sequence modeling: Berti and coauthors [1998, 2021, 2022], Fortini et al. [2014, 2023], Fong, Holmes, and Walker [2023]
  - Neural processes: Garnelo et. al [2203]
  - Joint predictions: Osband et. al [2022]

So what? AI-driven decisions

- AI now comprehends language and visual inputs
- Big opportunities to make decisions based on them
- Decision-making requires comprehending uncertainty and acting to resolve it

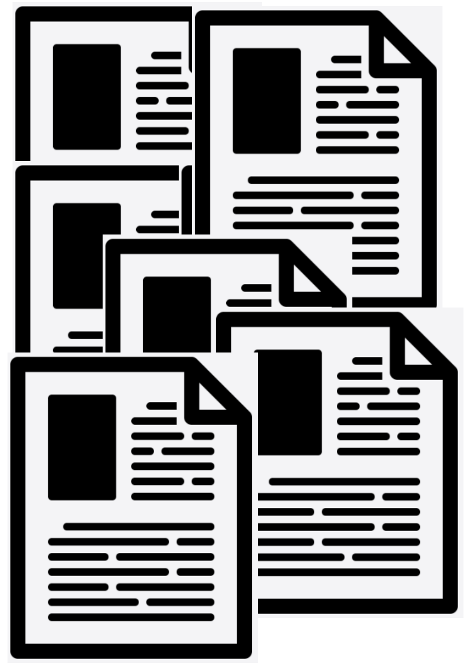
Cold start problem in RecSys

So what? AI-driven decisions

- AI now comprehends language and visual inputs
- Big opportunities to make decisions based on them
- Decision-making requires comprehending uncertainty and acting to resolve it

Today: news recommendations

New articles are released



An LLM reads them

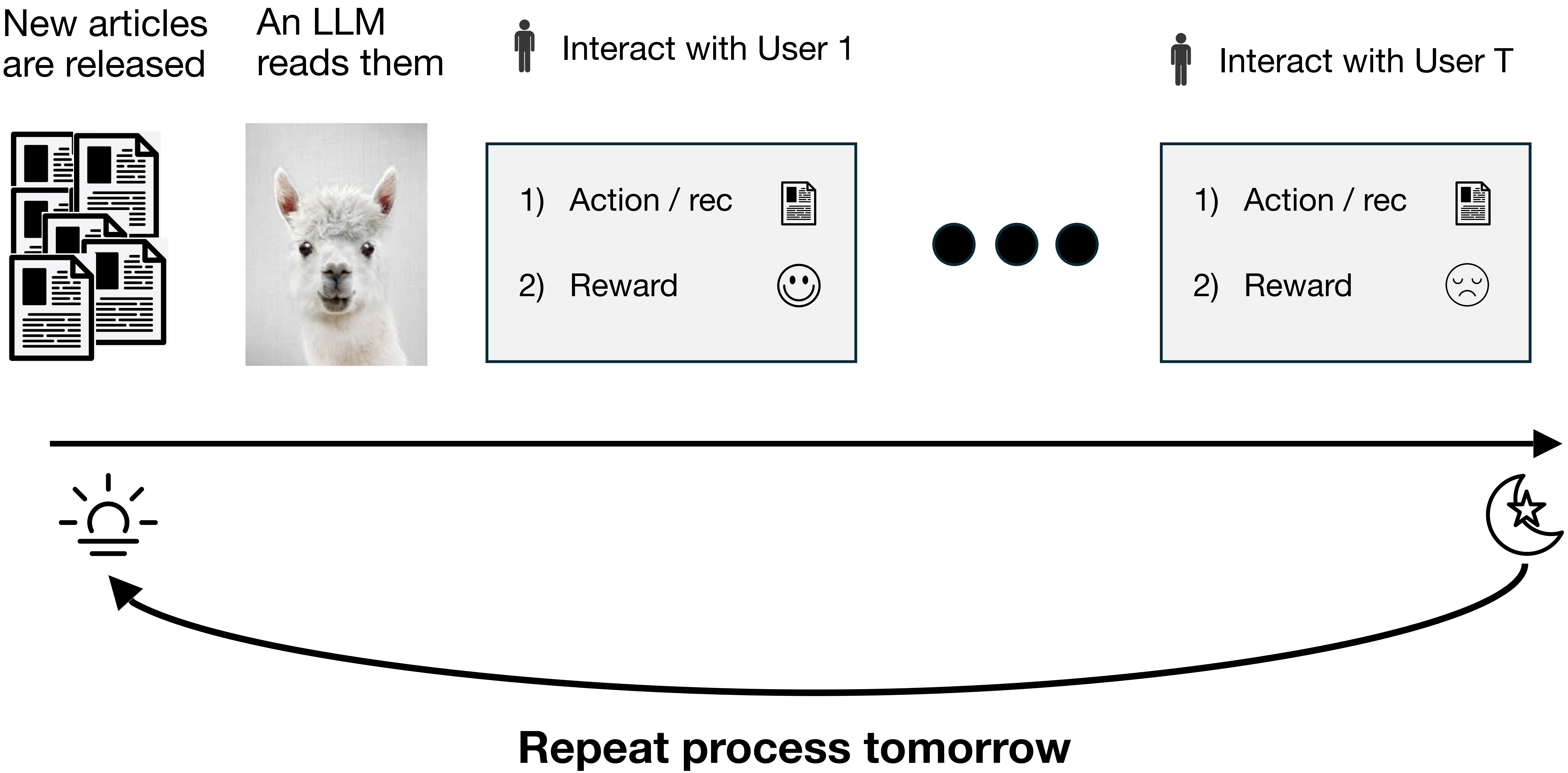
Interact with User 1

1) Action / rec
2) Reward

Interact with User T

1) Action / rec
2) Reward

Today: news recommendations



Goal: sharpen beliefs with more info

- When articles are released, AI system must
    1. Form informed prior based on article text
    2. Gather data to resolve remaining uncertainty
    3. Balance exploration/exploitation

- For this talk, we assume users are **exchangeable**
    - Algo generalizes to personalized settings with user features

Thompson sampling

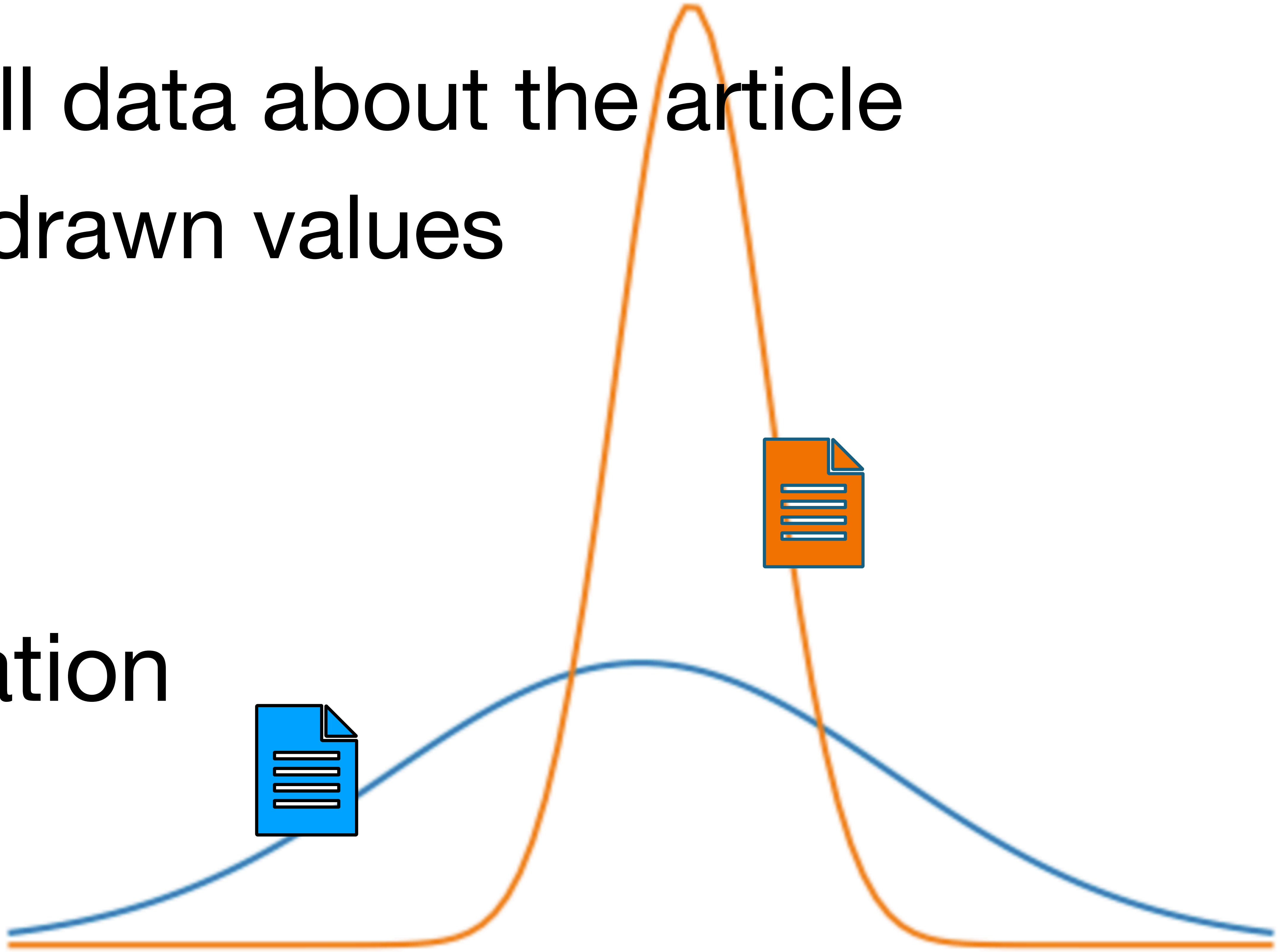
Z : Article features
U : Other latent factors that govern article popularity

- Draw U from the posterior given all data about the article
- Pick best article according to the drawn values

Thompson sampling

Z : Article features
U : Other latent factors that govern article popularity

- Draw U from the posterior given all data about the article
- Pick best article according to the drawn values



Balances exploration and exploitation

Thompson sampling

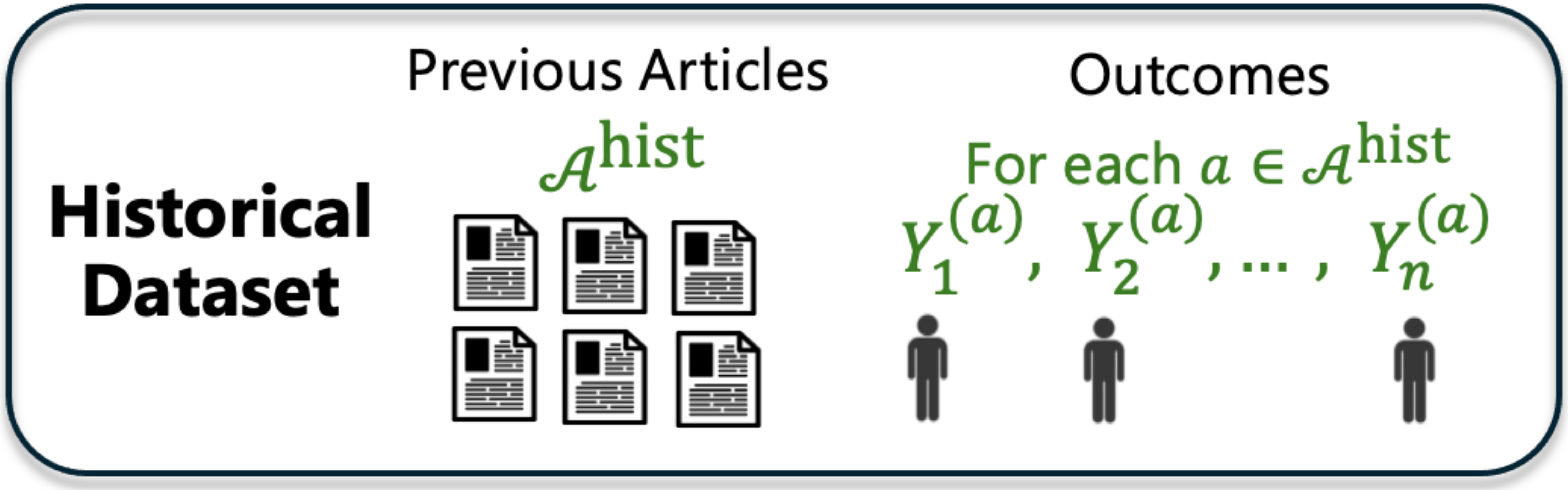
Z : Article features
U : Other latent factors that govern article popularity

- Draw U from the posterior given all data about the article
- Pick best article according to the drawn values

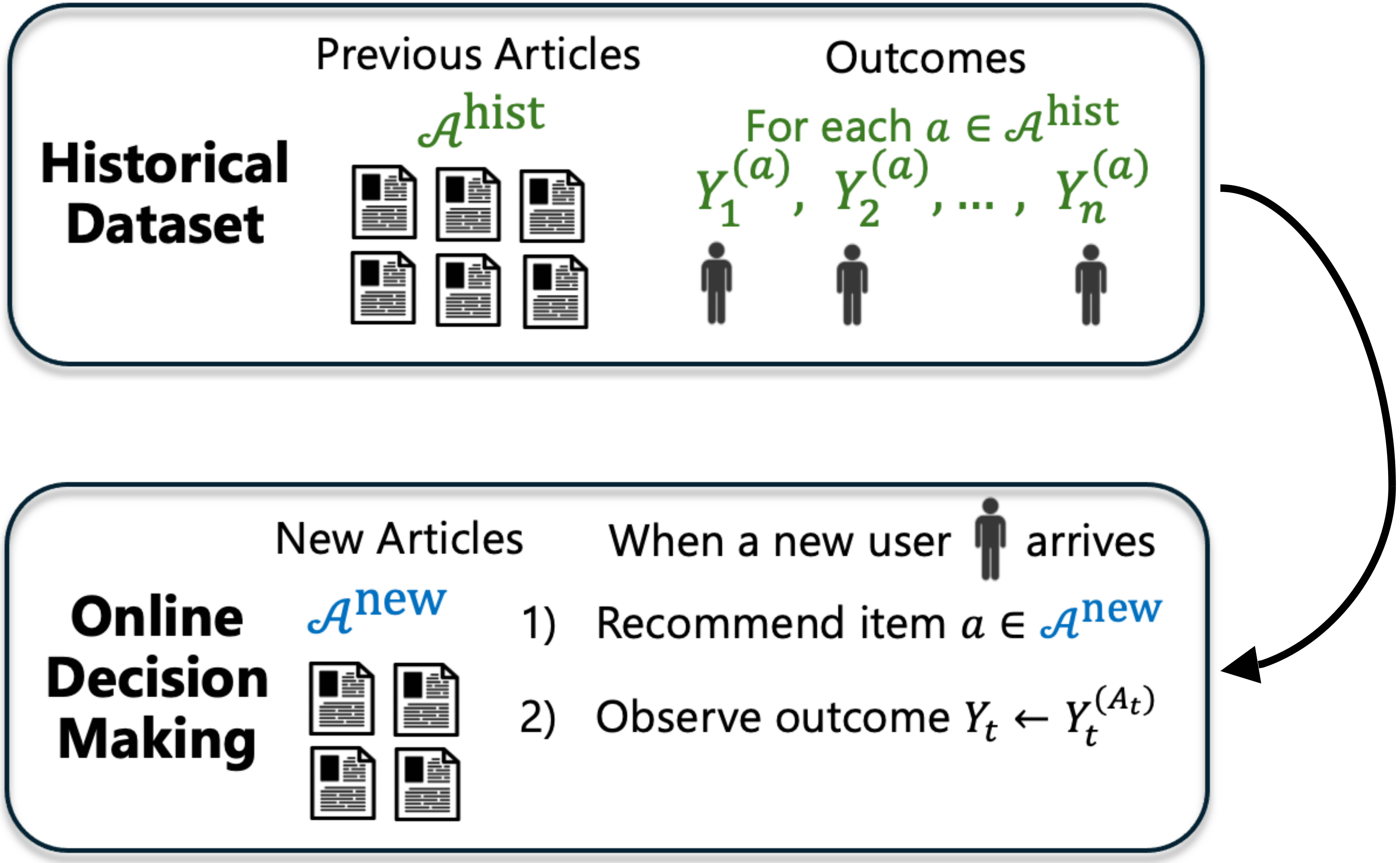
Main challenge

Informed exploration requires probabilistic model over text

Our solution: use massive historical data



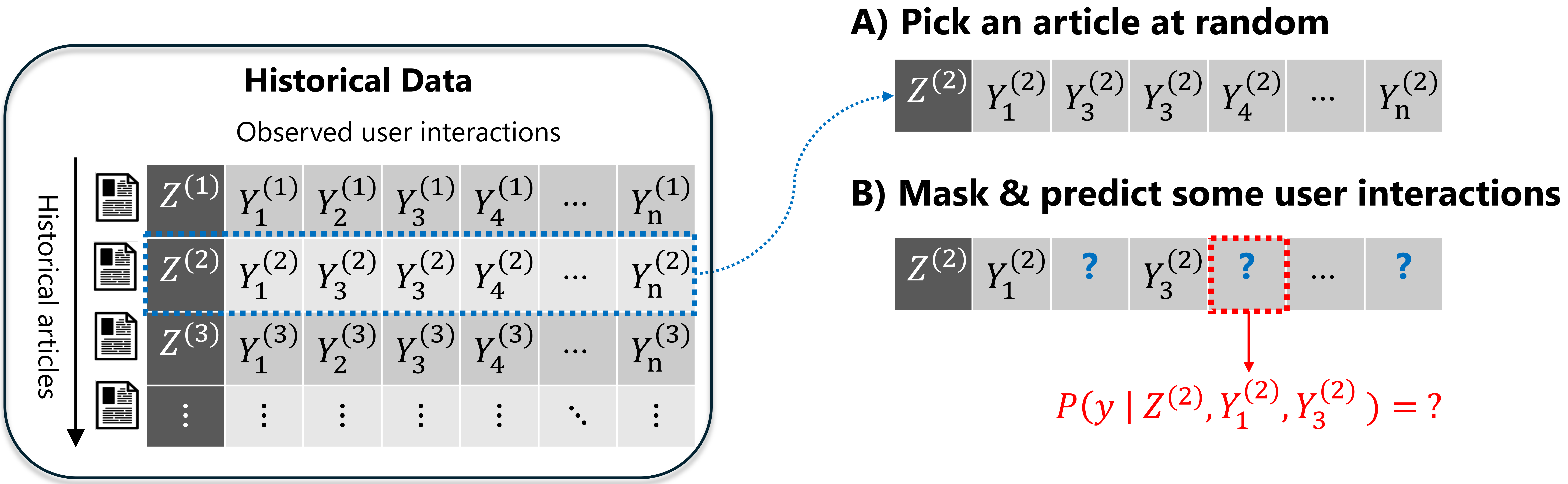
Step 1: Pretrain a sequence model



Use vast historical data to **warm-start** online decision-making

Step 1: Pretrain a sequence model

Low loss requires ability to sharpen beliefs



Step 1: Pretrain a sequence model

Low loss requires ability to sharpen beliefs

Predict well having observed few Y's

$Z^{(2)}$	$Y_1^{(2)}$	?	$Y_3^{(2)}$	?	…	?

$$P(y \mid Z^{(2)}, Y_1^{(2)}, Y_3^{(2)}) = ?$$

Predict well having observed more Y's

$Z^{(2)}$	$Y_1^{(2)}$	$Y_2^{(3)}$	$Y_3^{(2)}$	?	…	$Y_n^{(3)}$

$$P(y \mid Z^{(2)}, Y_1^{(2)}, Y_2^{(2)}, Y_3^{(2)}, Y_n^{(2)}) = ?$$

Sequence predictions enable probabilistic reasoning

1. When to heavily weigh the "prior" based on text Z?
2. When user interactions overwhelm the "prior"?

Step 2: Act to resolve uncertainty

Autoregressive generation reveals actions that *might* have great performance

1) Fill in missing outcomes by autoregressive generation

Observed and 'missing' outcomes

Potential user interactions

Today's articles

$Z^1 \;\; Y_1^1, Y_2^1, \hat{Y}_3^1, \cdots, \hat{Y}_T^1$

$\vdots$

$Z^k \;\; Y_1^a, \hat{Y}_2^a, \hat{Y}_3^a, \cdots, \hat{Y}_T^a$

Step 2: Act to resolve uncertainty

Autoregressive generation reveals actions that *might* have great performance

1) Fill in missing outcomes by autoregressive generation

Observed and 'missing' outcomes

Potential user interactions

Today's articles

$Z^1 \quad Y_1^1, Y_2^1, \hat{Y}_3^1, \cdots, \hat{Y}_T^1$

$\vdots$

$Z^k \quad Y_1^a, \hat{Y}_2^a, \hat{Y}_3^a, \cdots, \hat{Y}_T^a$

2) Compute reward rates under hypothetical table

Average

$$\hat{\mu}^1 = \frac{Y_1^1 + Y_2^1 + \hat{Y}_3^1 + \cdots + \hat{Y}_T^1}{T}$$

Average

$$\hat{\mu}^a = \frac{Y_1^a + \hat{Y}_2^a + \hat{Y}_3^a + \cdots + \hat{Y}_T^a}{T}$$

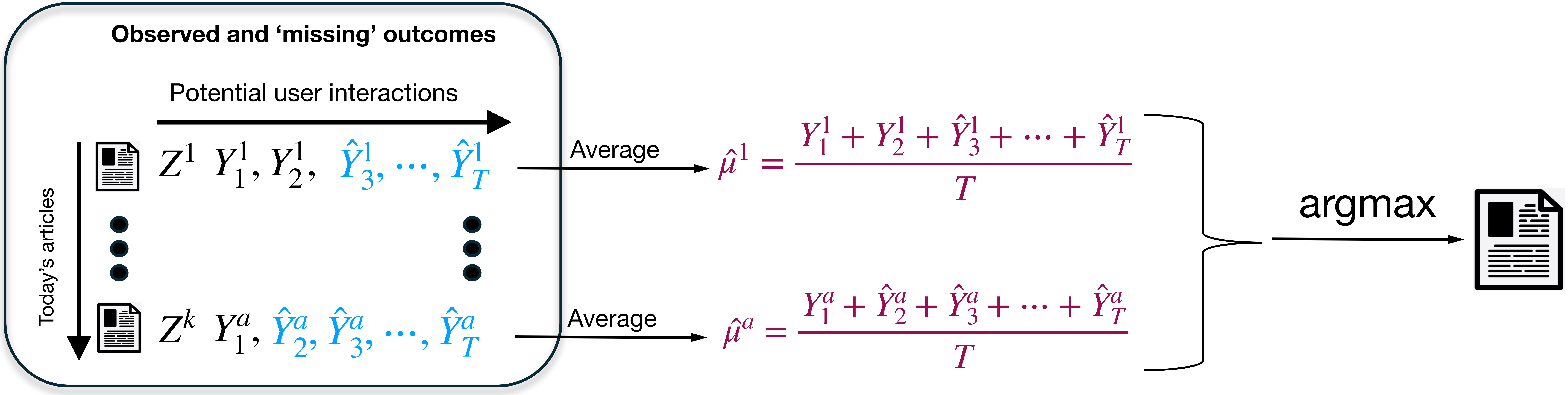
Step 2: Act to resolve uncertainty

Autoregressive generation reveals actions that *might* have great performance

1) Fill in missing outcomes by autoregressive generation

2) Compute reward rates under hypothetical table

3) Pick item with highest hypothetical reward



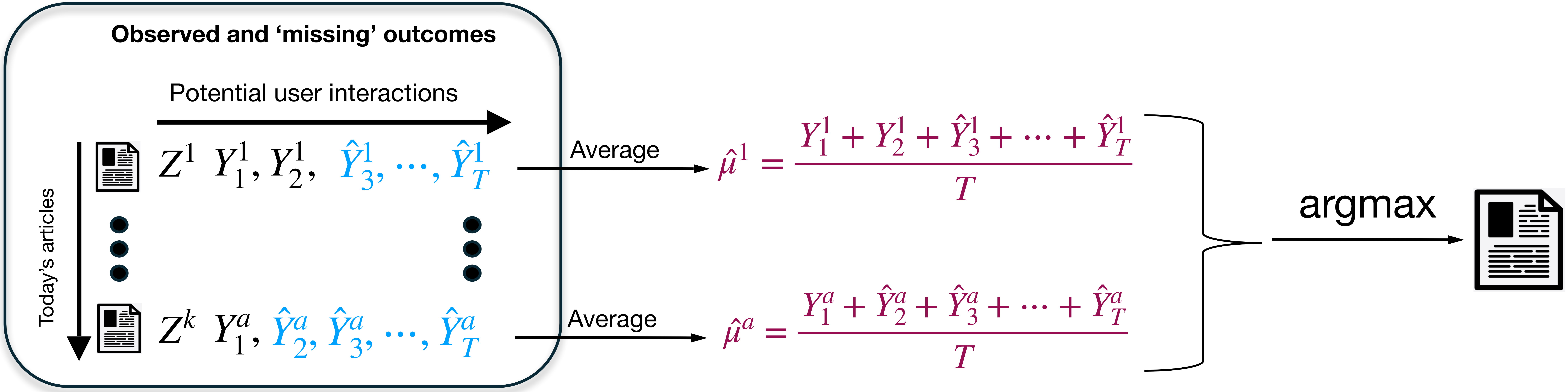
Step 2: Act to resolve uncertainty

Autoregressive generation reveals actions that *might* have great performance

1) Fill in missing outcomes by autoregressive generation

2) Compute reward rates under hypothetical table

3) Pick item with highest hypothetical reward



Action with high potential has a fair chance

Theorem: regret controlled by perplexity

Data assumptions

1. Independently drawn articles.

- Text/ potential outcome sets $\left(Z^{(a)}, Y_1^{(a)}, \ldots, Y_T^{(a)}\right)$ are independent across articles.

2. Historical data is representative of future days

- Distribution $\left(Z^{(a)}, Y_1^{(a)}, \ldots, Y_T^{(a)}\right)$ is the same for articles in historical data as what governs tomorrow's draw..

3. Exchangeability across users

- $P^*\left(Y_1^{(a)}, \ldots, Y_T^{(a)} \mid Z^{(a)}\right) = P^*\left(Y_{\sigma(1)}^{(a)}, \ldots, Y_{\sigma(T)}^{(a)} \mid Z^{(a)}\right)$

Theorem: regret controlled by perplexity

Optimal decision under imagined data = best under posterior draw

Two assumptions: 1) bounded rewards, 2) training length sequence exceeds T

Regret when using autoregressive model p_θ $\leq$ Regret of Thompson sampling with true prior

$$+ \sqrt{2 \cdot \text{no. articles each day} \cdot (\ell_T(p^\star) - \ell_T(p_\theta))}$$

Optimal autoregressive prediction

Validation loss

Theorem: regret controlled by perplexity

Optimal decision under imagined data = best under posterior draw

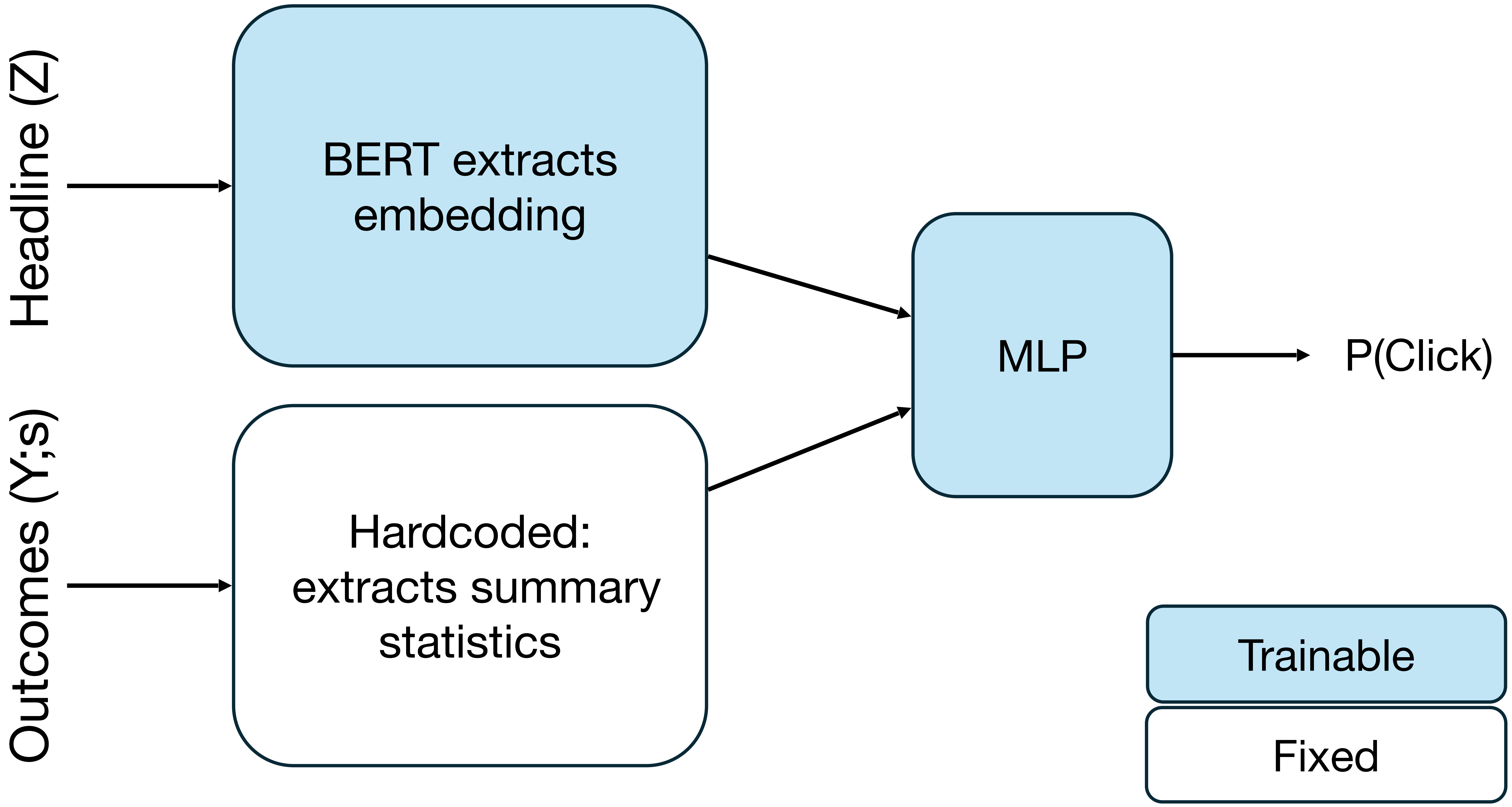
Two assumptions: 1) bounded rewards, 2) training length sequence exceeds T

Regret when using autoregressive model p_θ $\leq$ Regret of Thompson sampling with true prior

$$+ \sqrt{2 \cdot \text{no. articles each day} \cdot (\ell_T(p^\star) - \ell_T(p_\theta))}$$

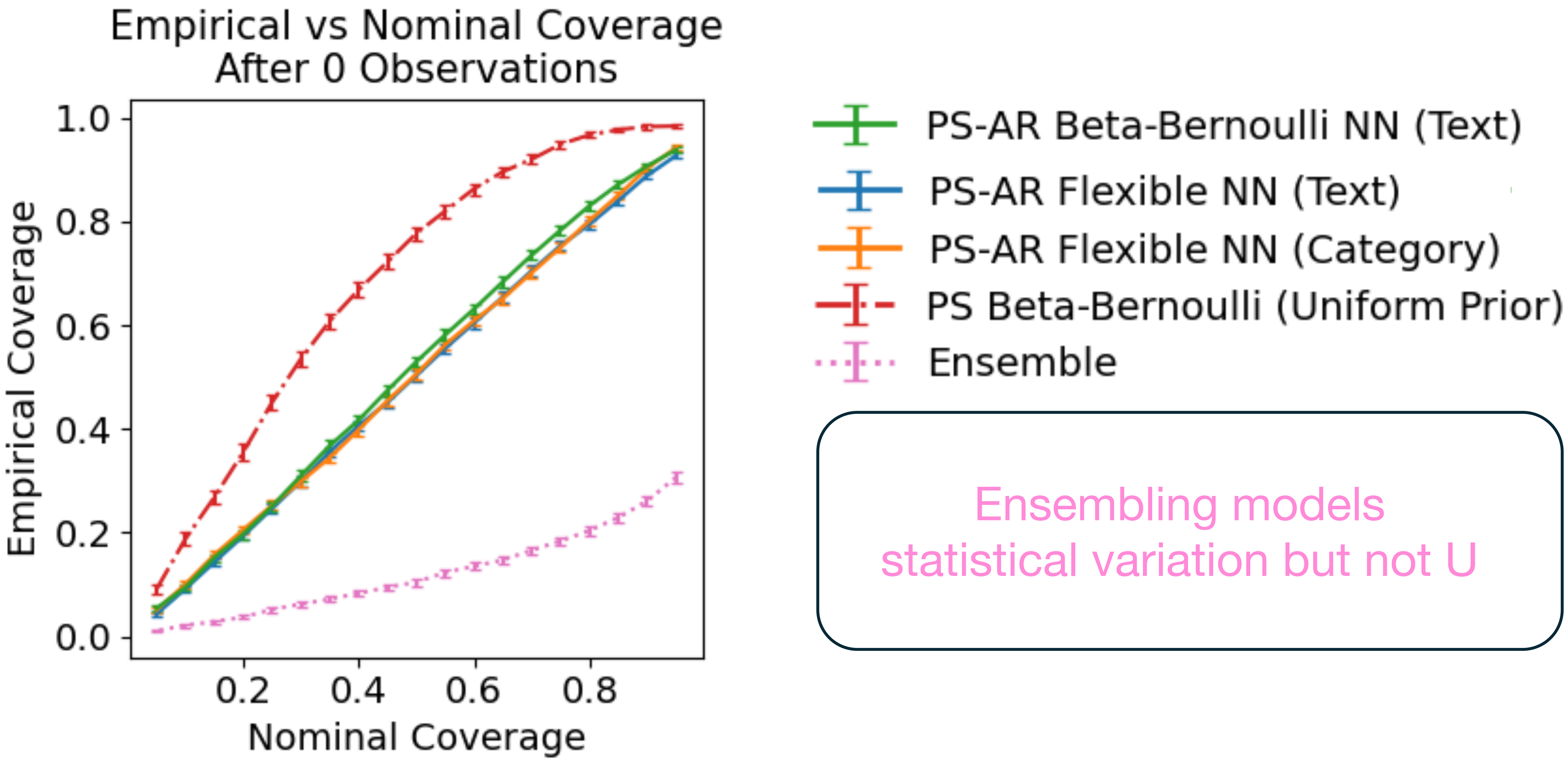
Scaling laws over user interactions control recommendation quality!

Proof of concept: baby sequence model



Coverage

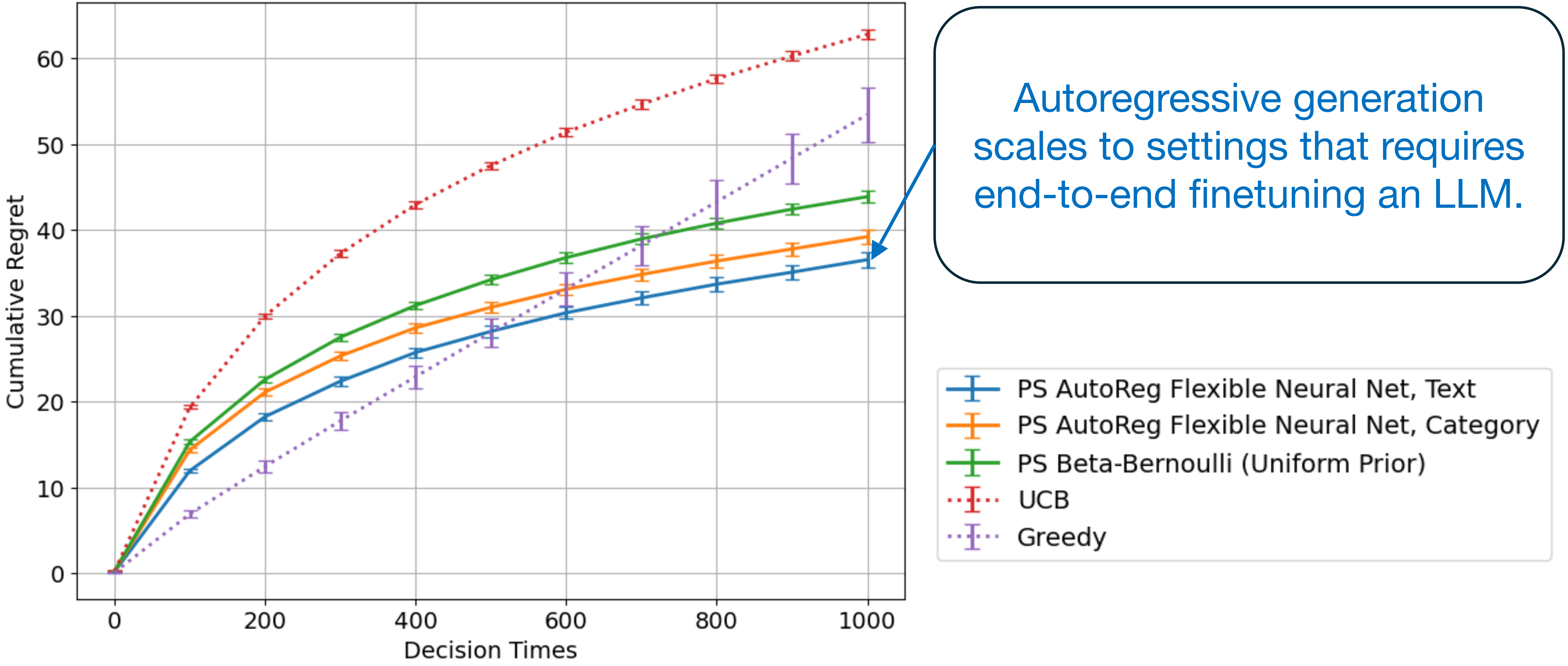
Autoregressive generation mimics proper Bayesian beliefs given headline (text)



Ensembling models statistical variation but not U

Regret

A semi-realistic simulator using public MSN news article data



Autoregressive generation scales to settings that requires end-to-end finetuning an LLM.

Vision: probabilistic reasoning

Intelligent agents must comprehend uncertainty and take actions to resolve it

Going beyond web-data requires careful data collection

1. Formulate informed prior
2. Decide which data to collect subject to cost constraints
3. Update beliefs

Vision: probabilistic reasoning

Intelligent agents must comprehend uncertainty and take actions to resolve it

Going beyond web-data requires careful data collection

1. Formulate informed prior
2. Decide which data to collect subject to cost constraints
3. Update beliefs

Proposal: Model-based planning through sequence models

1. Rollout: autoregressively generate hypothetical data
2. Simulate posterior update by appending data to context
3. Guide real decision based on how sharply belief shifts

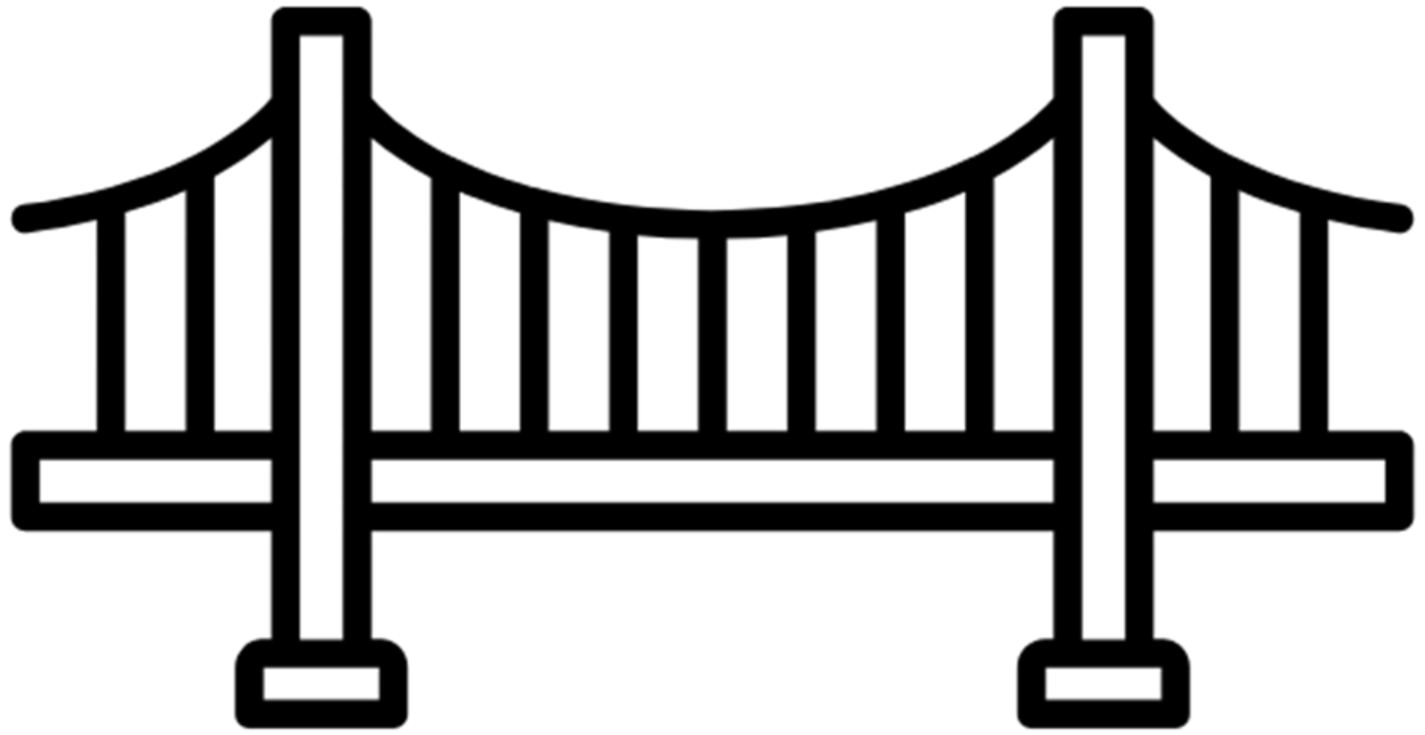
Conceptual: a well motivated problem crystalizing the insights

Algorithmic: link with interactive decision-making

Theory: accurate sequence modeling implies low regret

Experiments: scalable implementations with LLMs.

Generative Sequence Modeling



Exploration & Uncertainty Quantification

Posterior Sampling via Autoregressive Generation, ZC**N**R, arXiv:2405.19466

Exchangeable Sequence Models Quantify Uncertainty Over Latents, Y**N**, arXiv:2408.03307