Probe-Free Low-Rank Intervention

Chonghe Jiang
Chinese University of Hong Kong
POMS-HK, 2025

Probe-Free Low-Rank Intervention

LLM Task + OR Technique =

Chonghe Jiang
Chinese University of Hong Kong

joint work with
Bao Nguyen (CUHK)
Anthony Man-Cho So (CUHK)
Viet Anh Nguyen (CUHK)

LLM can give untruthful answers

What happens if we eat melon seeds?

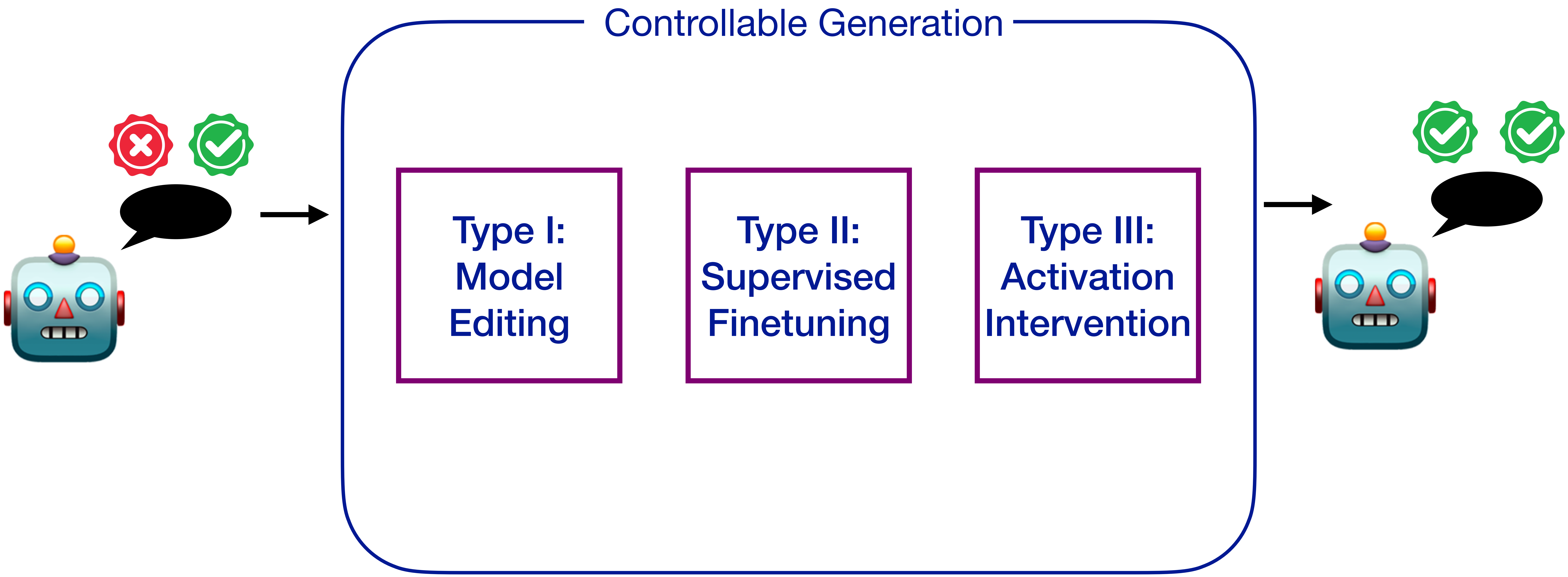
Death

Nothing

Key question: With *minimal invasion*, how can we promote truthful generation?

non-toxic/fair/ …

Solution - Controllable Generation



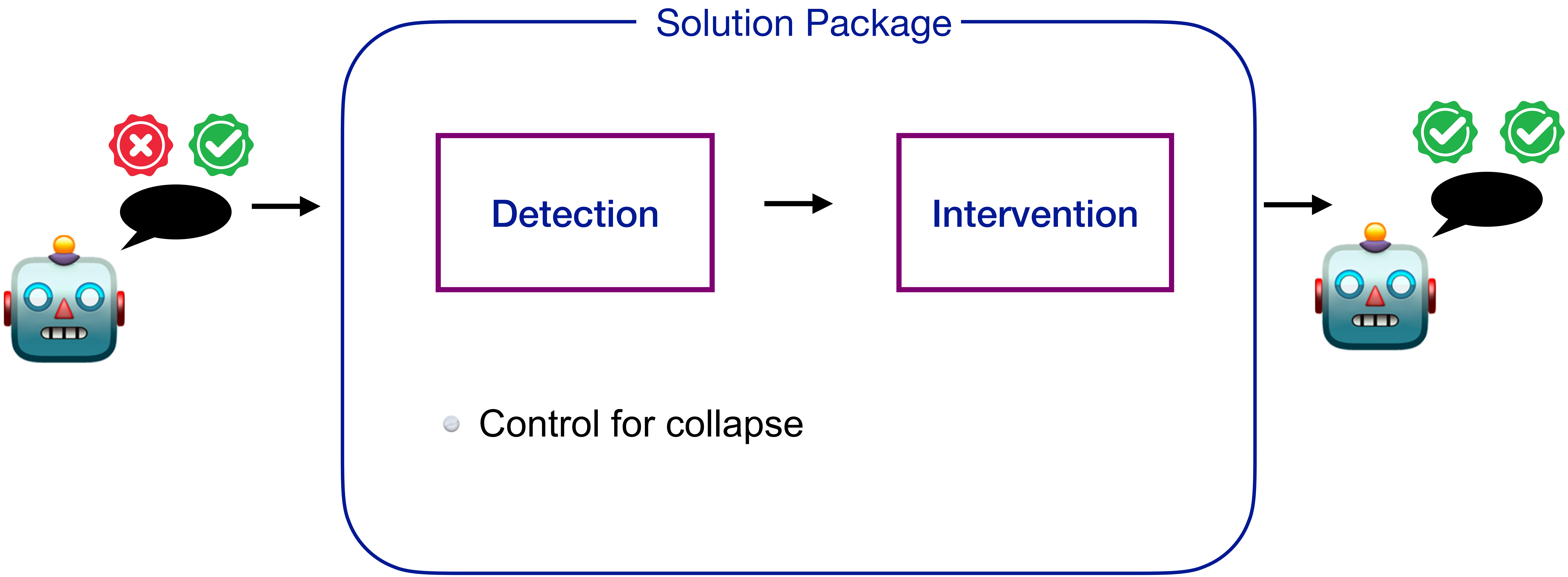
Activation Intervention - Advantages & Challenges

- **Advantage:**
  - It does not require text samples to change the model
  - Instead, it edits the activation vector in the inference time

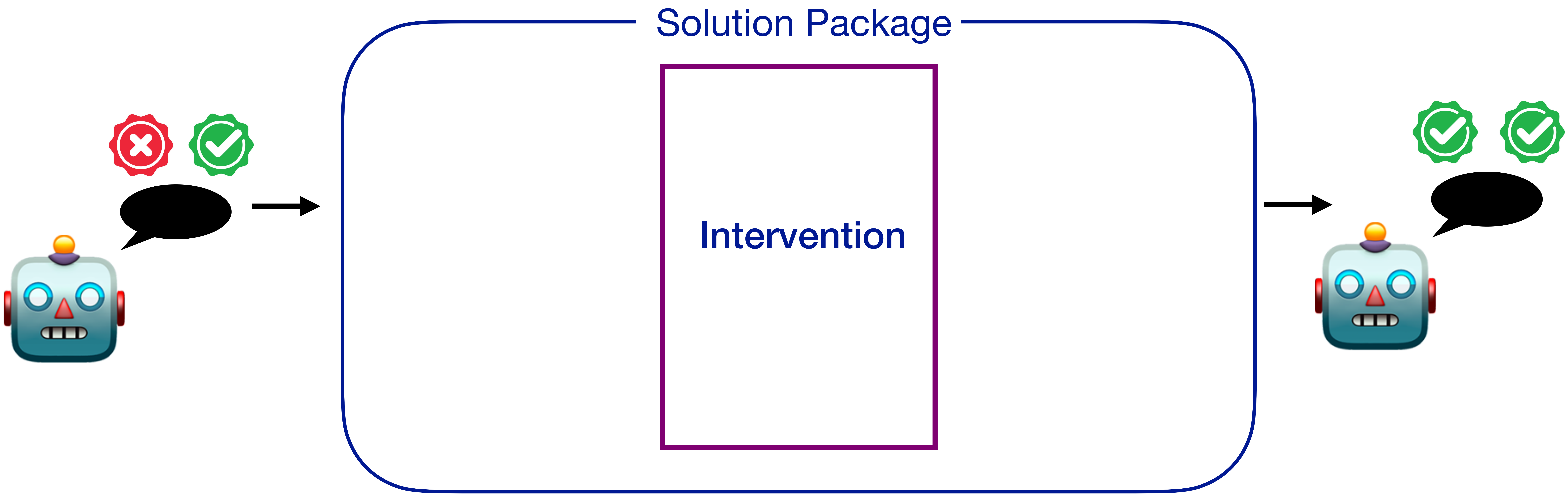
- **Challenge I:** Detect true vs untrue during generation (inference)
  - Human uses words, computer uses number
  - LLMs are complex

- **Challenge II:** Operational constraints
  - Resources for training (memory)
  - Resources for deployment (memory, time)
  - Explainability

An Overview of **Prior Arts**



Can we eliminate the detection step?



- Ideas: drop detection, intervene all-the-time
- Goal: high quality + efficient
  - Truthful activations should not be modified too much
  - Untruthful activations should be corrected
  - The intervention method should be computationally efficient

LLM Generation Mechanism

- We need to understand how LLMs represent "knowledge" and generate texts

Patching Question and Answer

What happens if we eat melon seeds? Nothing.

LLM Generation Mechanism

- We need to understand how LLMs represent “knowledge” and generate texts

Simplified tokenization: word level

What	happens	if	we	eat	melon	seeds	?	Nothing	.
1	2	3	4	5	6	7	8	9	10

LLM Generation Mechanism

- We need to understand how LLMs represent "knowledge" and generate texts

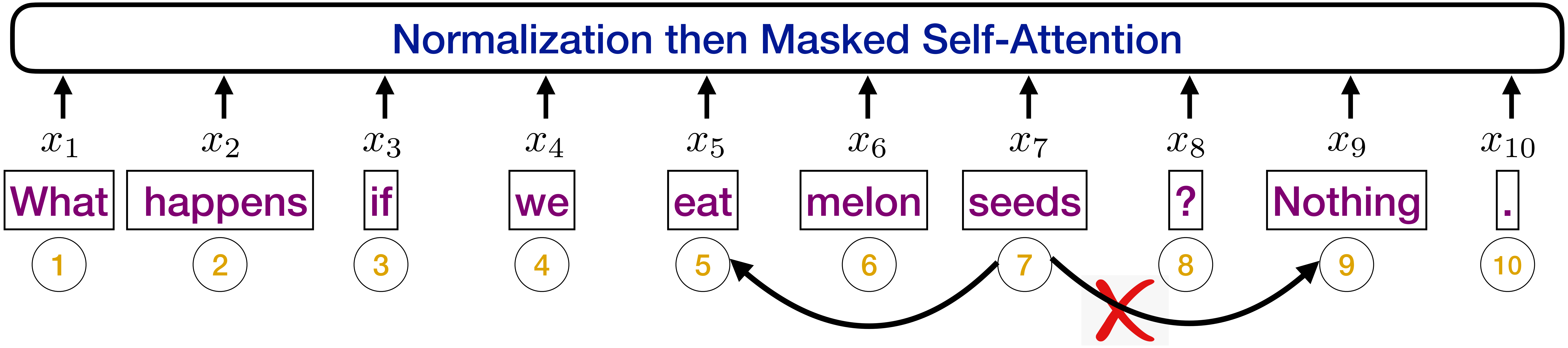
Token + positional embeddings

$x_i \in \mathbb{R}^{4096}$

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
What	happens	if	we	eat	melon	seeds	?	Nothing	.
1	2	3	4	5	6	7	8	9	10

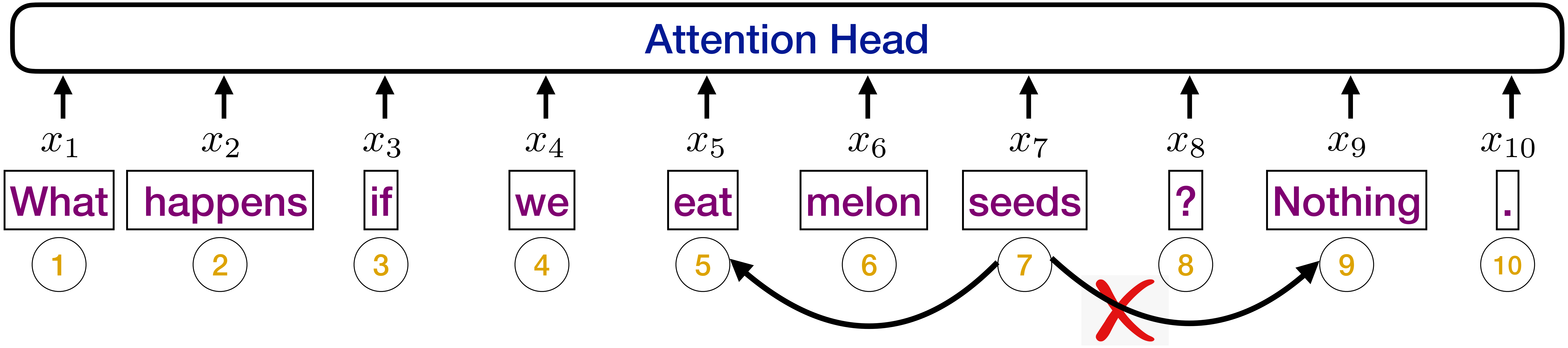
LLM Generation Mechanism

$$normalize(\text{input}) = \frac{\text{input} - \text{mean}}{\text{standard deviation}}$$

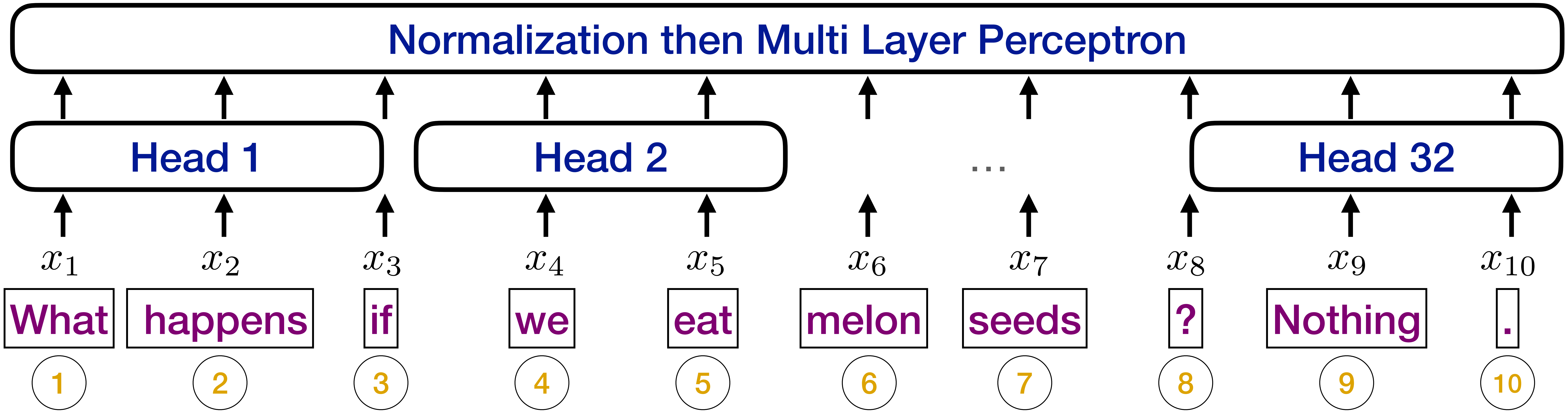


LLM Generation Mechanism

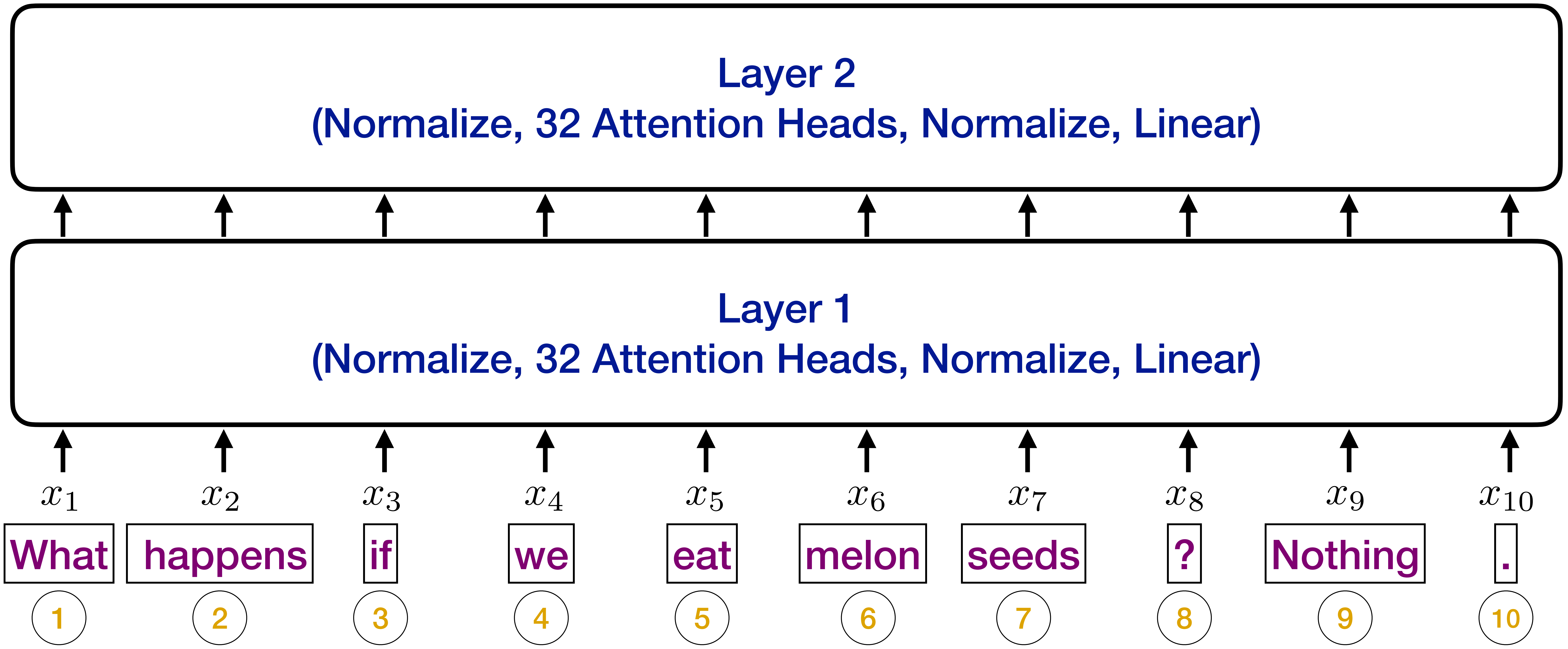
$$normalize(\text{input}) = \frac{\text{input} - \text{mean}}{\text{standard deviation}}$$

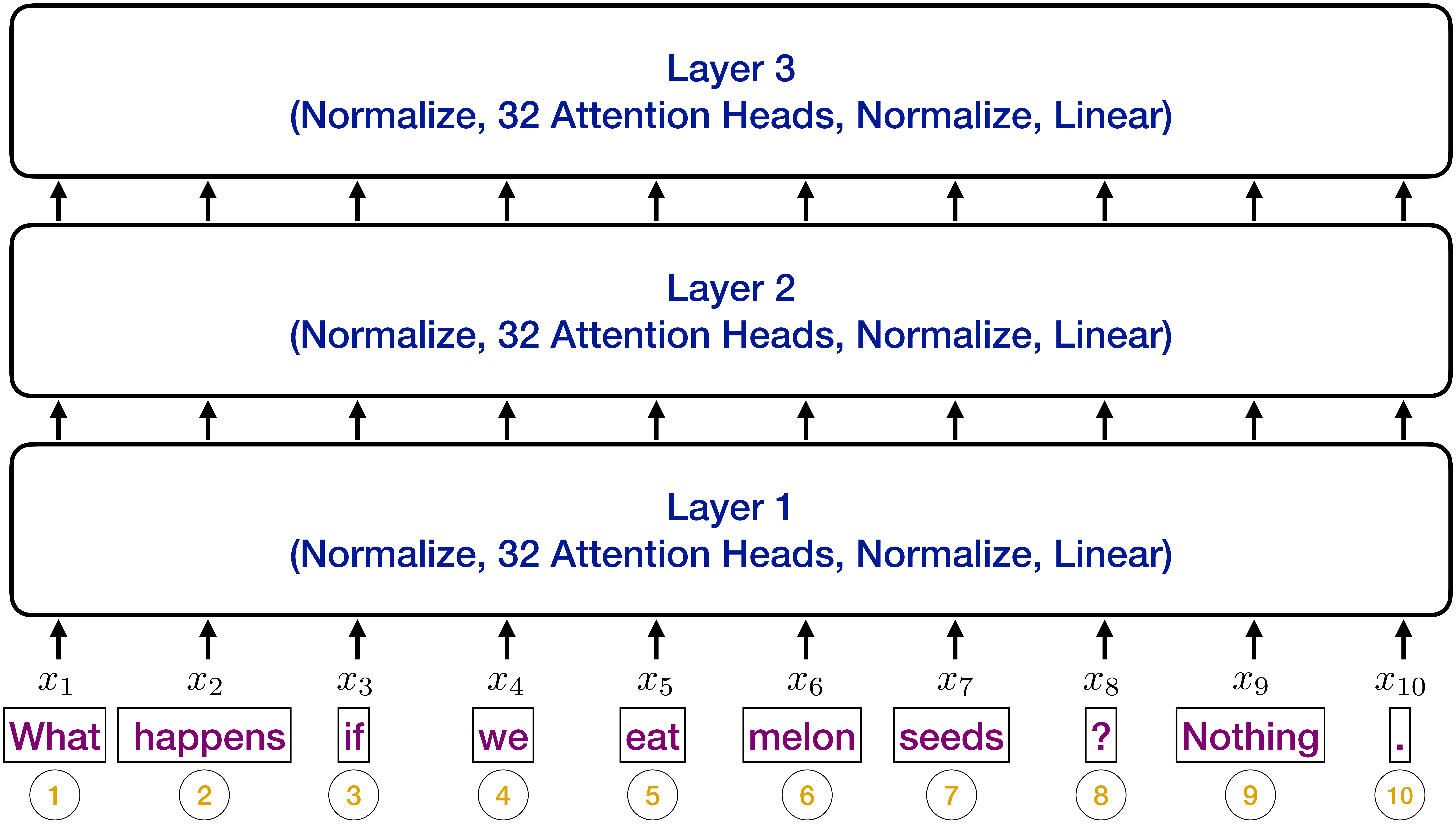


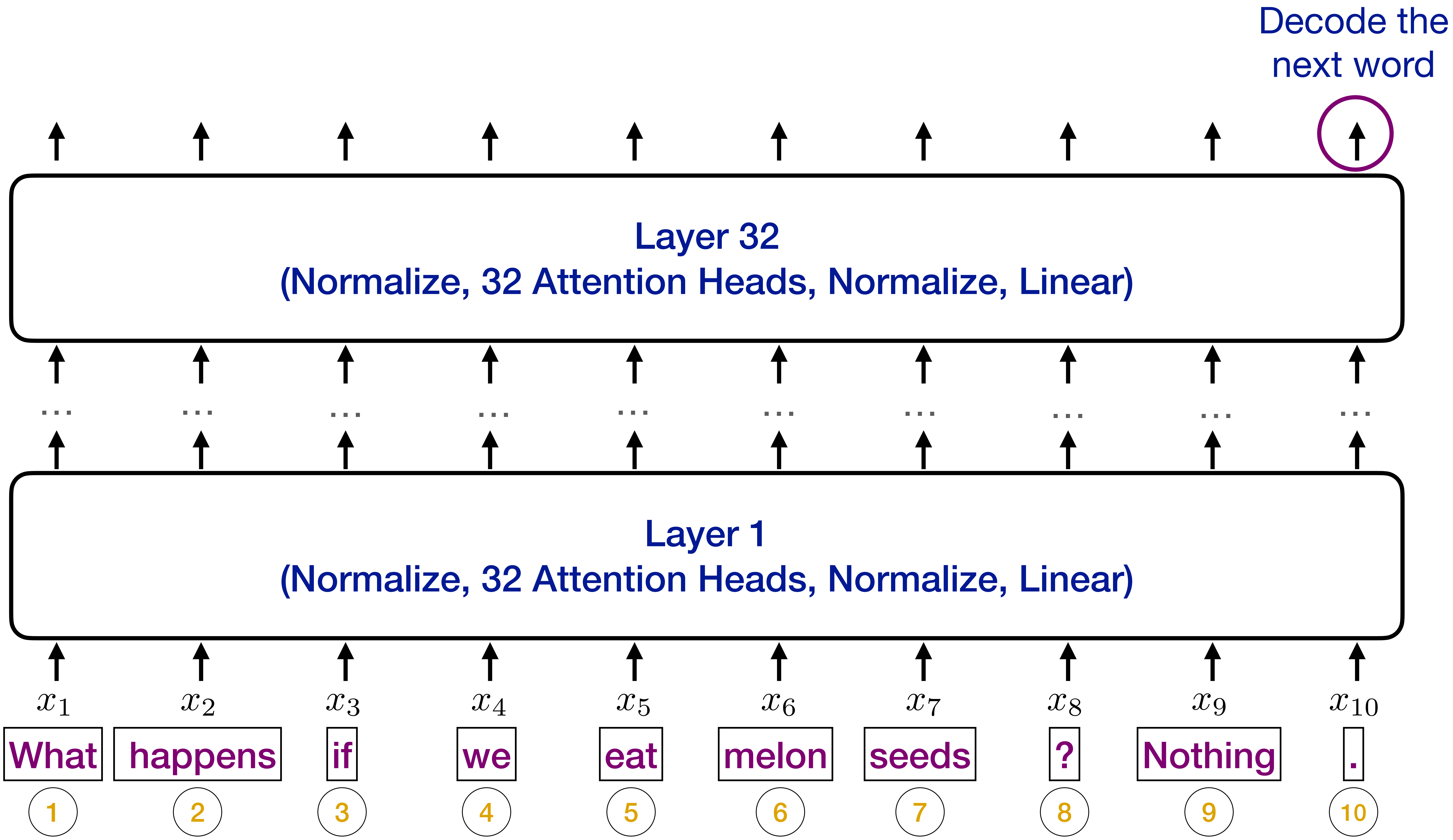
LLM Generation Mechanism



LLM Generation Mechanism

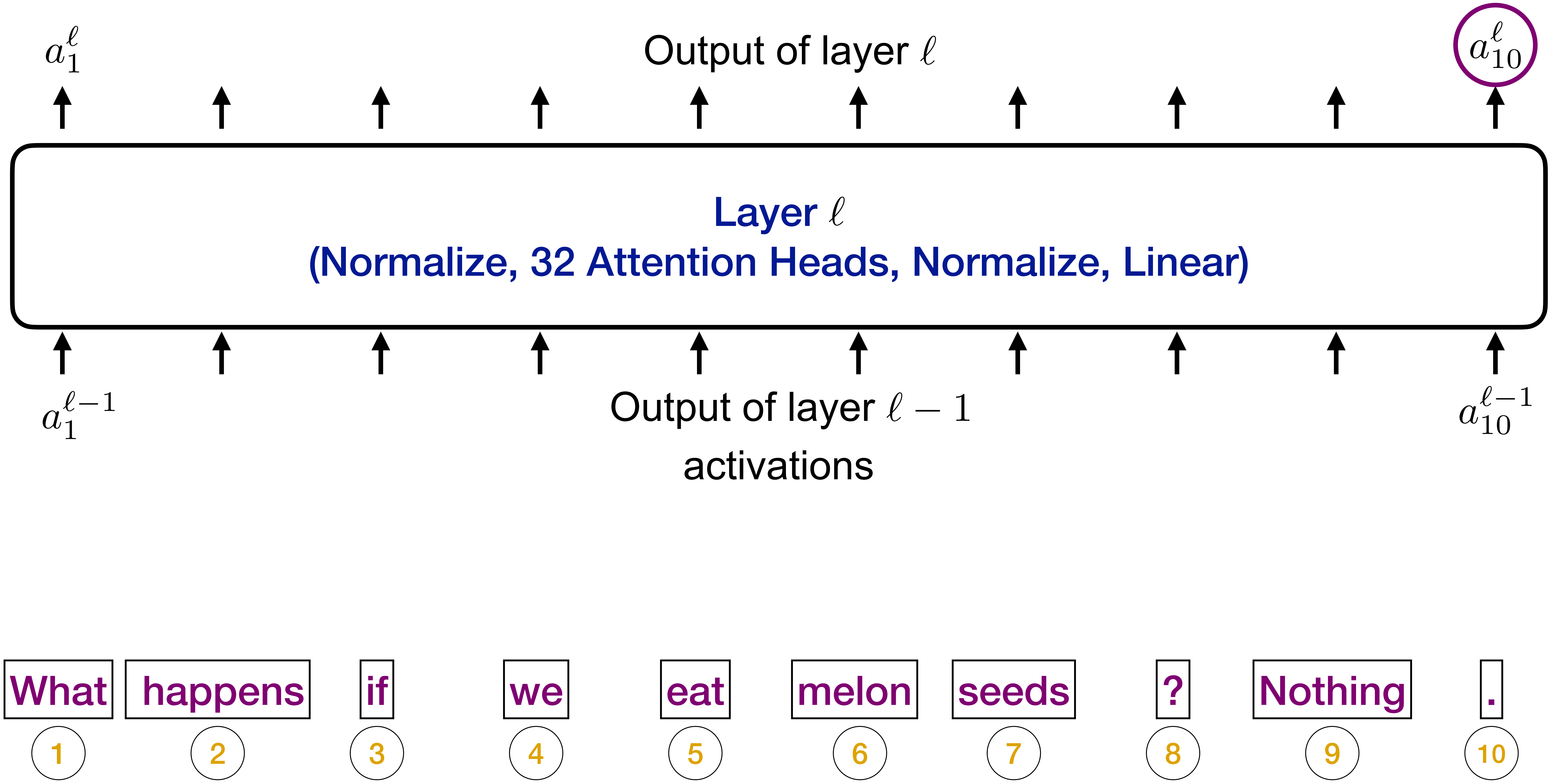



Layer 3
(Normalize, 32 Attention Heads, Normalize, Linear)
Layer 2
(Normalize, 32 Attention Heads, Normalize, Linear)
Layer 1
(Normalize, 32 Attention Heads, Normalize, Linear)
x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8 x_9 x_{10}
What happens if we eat melon seeds ? Nothing .
1 2 3 4 5 6 7 8 9 10


Decode the next word
Layer 32
(Normalize, 32 Attention Heads, Normalize, Linear)
Layer 1
(Normalize, 32 Attention Heads, Normalize, Linear)
x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8 x_9 x_{10}
What happens if we eat melon seeds ? Nothing .
1 2 3 4 5 6 7 8 9 10

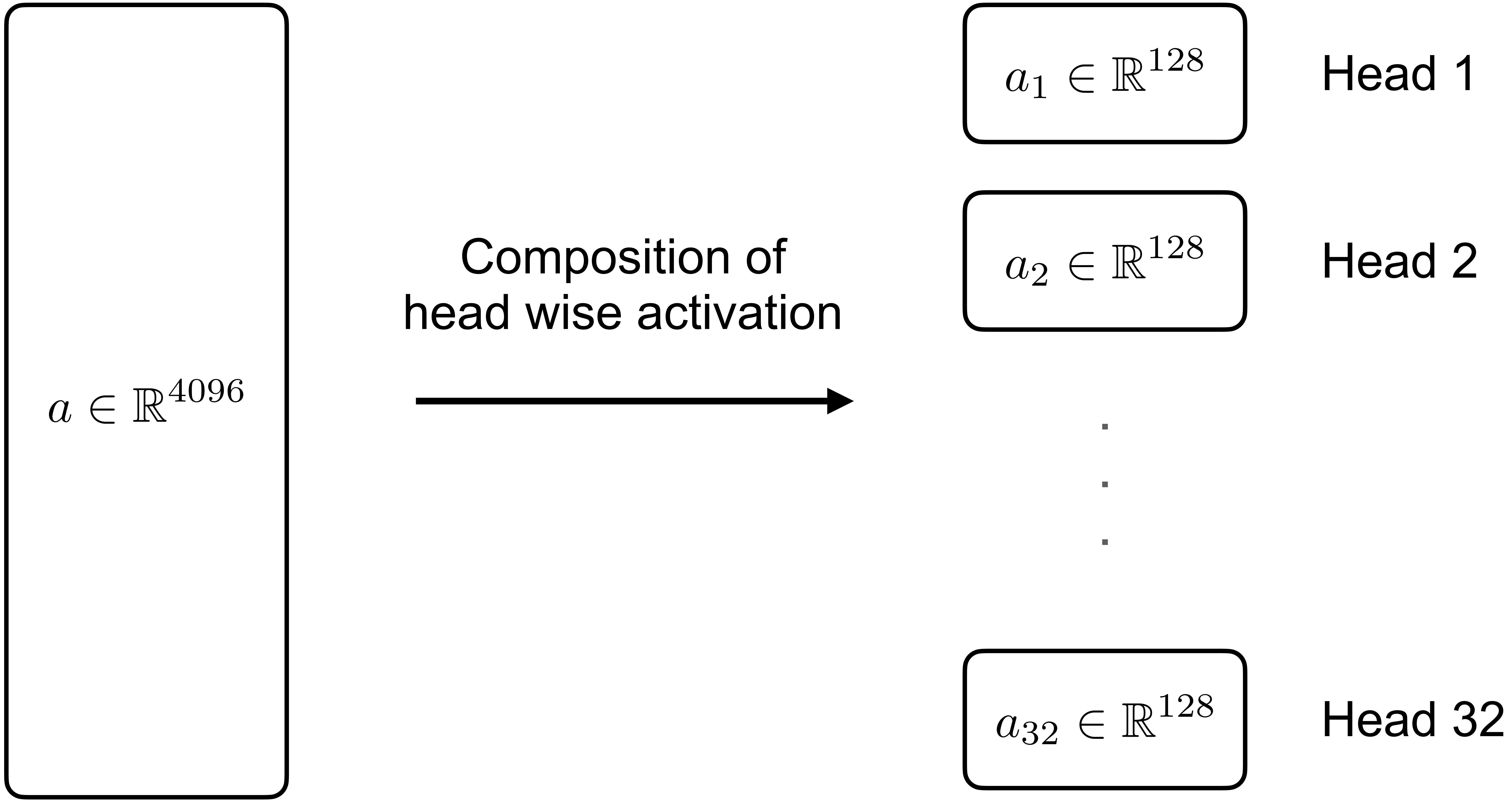
Focus on one layer

Contains the largest
amount of information

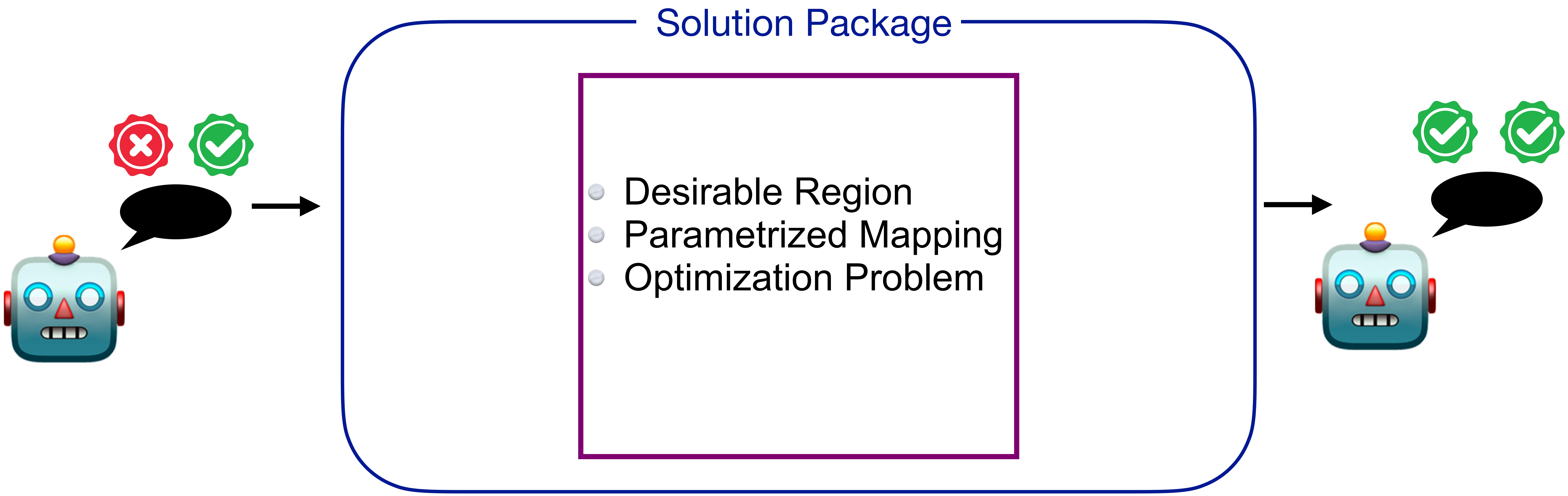


Activations of the last token

- Patch text, pass it through the LM to get activations of the last token at layer ℓ
- We always take the last token, so we drop the index $a^{\ell}_{10} \to a$

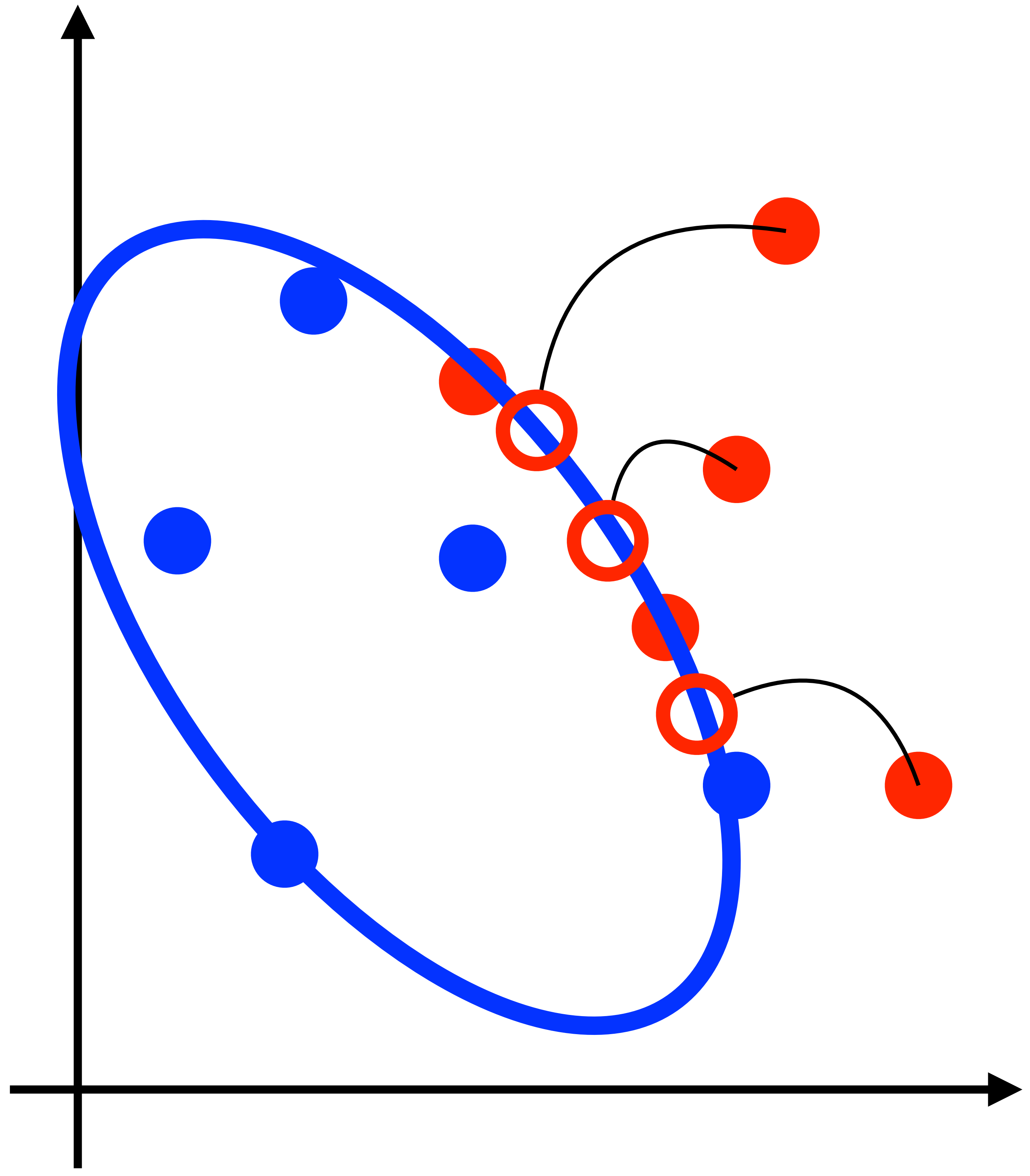


Main Idea



- We work on the high-dimensional vector space
- We establish generalizable intervention methods

Desirable Region: Ellipsoid Model



Activations
dimension 4096

- Desirable region is an ellipsoid:

$$\mathcal{E} = \{a : (a - \hat{\mu})^\top \hat{\Sigma}^{-1} (a - \hat{\mu}) \leq \rho\}$$

- Projection onto the desirable region

$$\mathrm{Proj}_{\mathcal{E}}(x) = \arg\min_{a \in \mathcal{E}} \ (a - x)^\top \widehat{\Sigma}^{-1} (a - x)$$

- Difficulty: each question has a different good region
- Difficulty: characterize the region with limited information

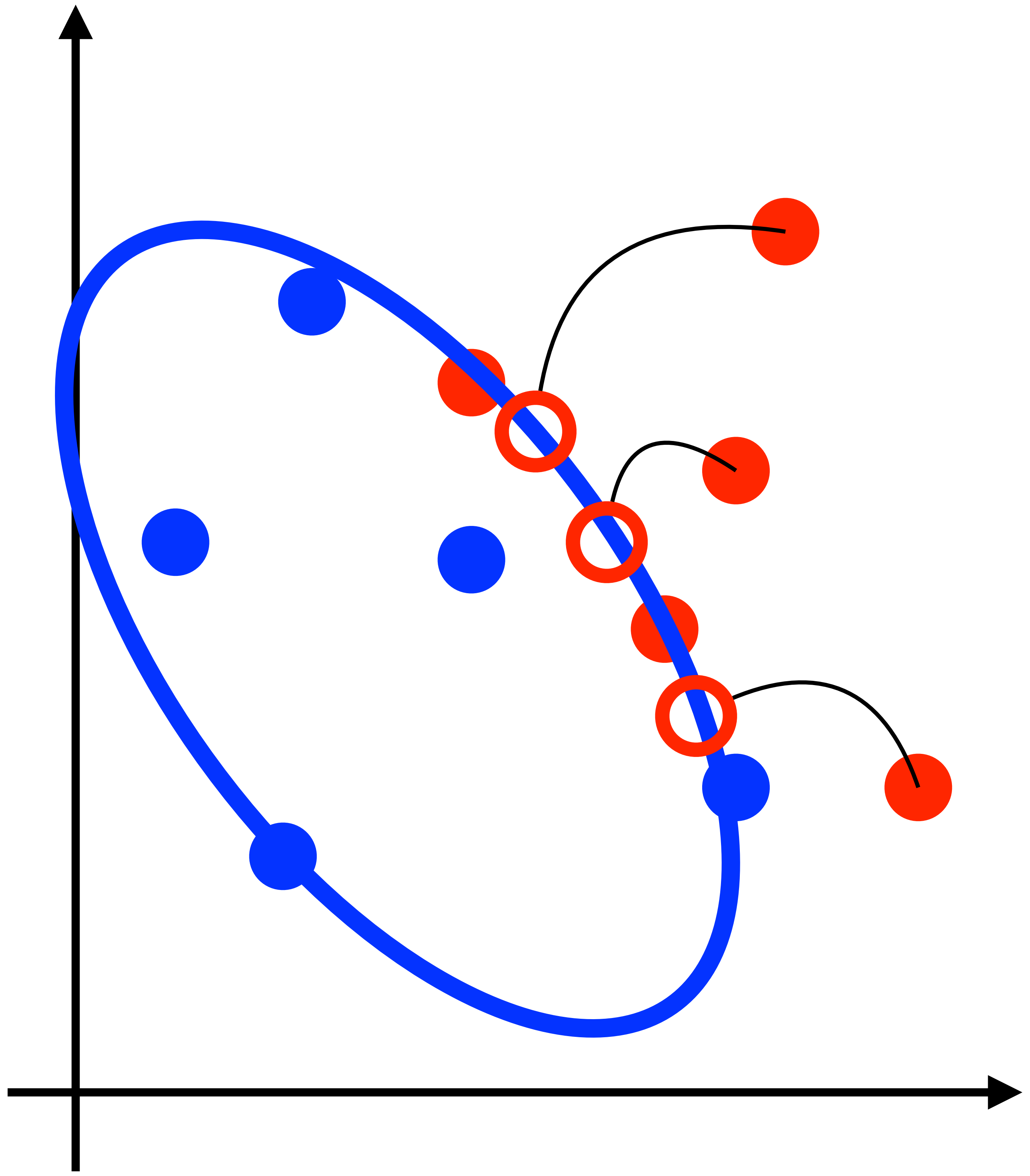
Desirable Region: Ellipsoid Model

- Desirable region is an ellipsoid:

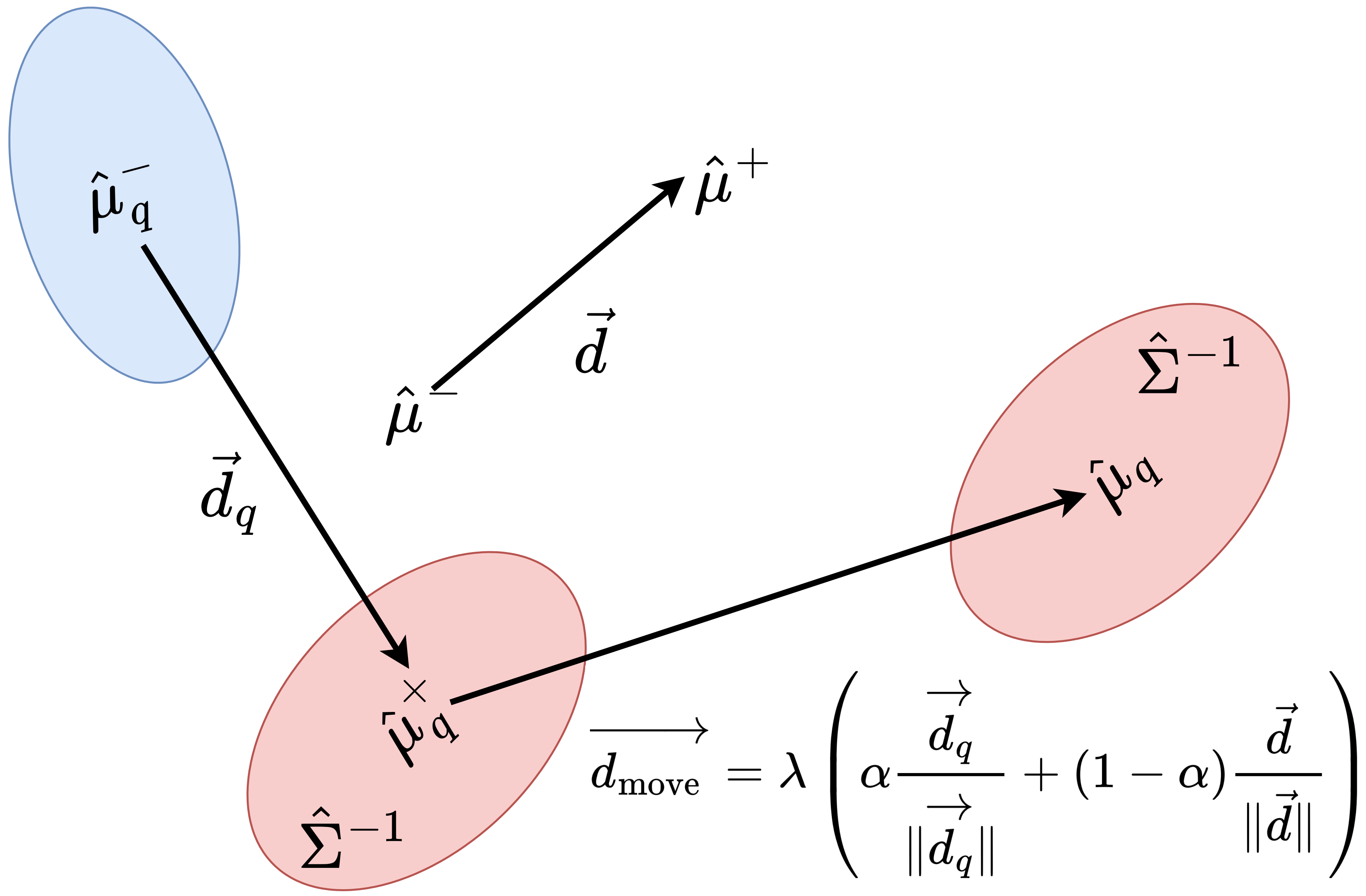
$$\mathcal{E} = \{a : (a - \hat{\mu})^\top \hat{\Sigma}^{-1} (a - \hat{\mu}) \leq \rho\}$$

- Difficulty: each question has a different good region
- Difficulty: characterize the region with limited information

$$\hat{\mu}_q = \hat{\mu}_q^+ + \lambda \left(\alpha \underbrace{\frac{\hat{\mu}_q^+ - \hat{\mu}_q^-}{\|\hat{\mu}_q^+ - \hat{\mu}_q^-\|}} + (1-\alpha) \underbrace{\frac{\hat{\mu}^+ - \hat{\mu}^-}{\|\hat{\mu}^+ - \hat{\mu}^-\|}} \right)$$

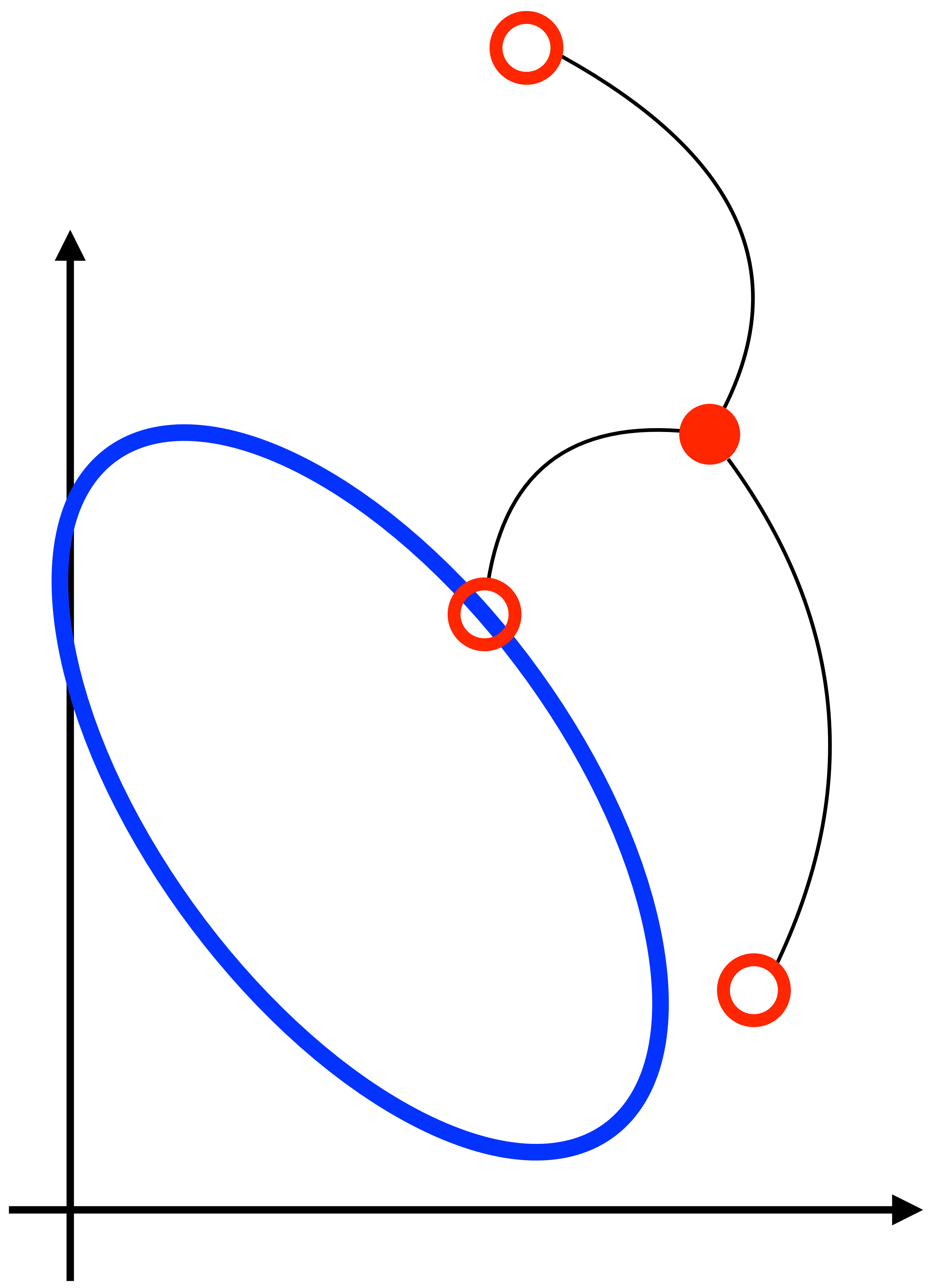


Activations dimension 4096



Ellipsoid Model Parameter

Parametrized Mapping: Low-rank



Different Mappings

- Desirable region is an ellipsoid:

$$\mathcal{E} = \{a : (a - \hat{\mu})^\top \hat{\Sigma}^{-1}(a - \hat{\mu}) \le \rho\}$$

- Projection onto the desirable region

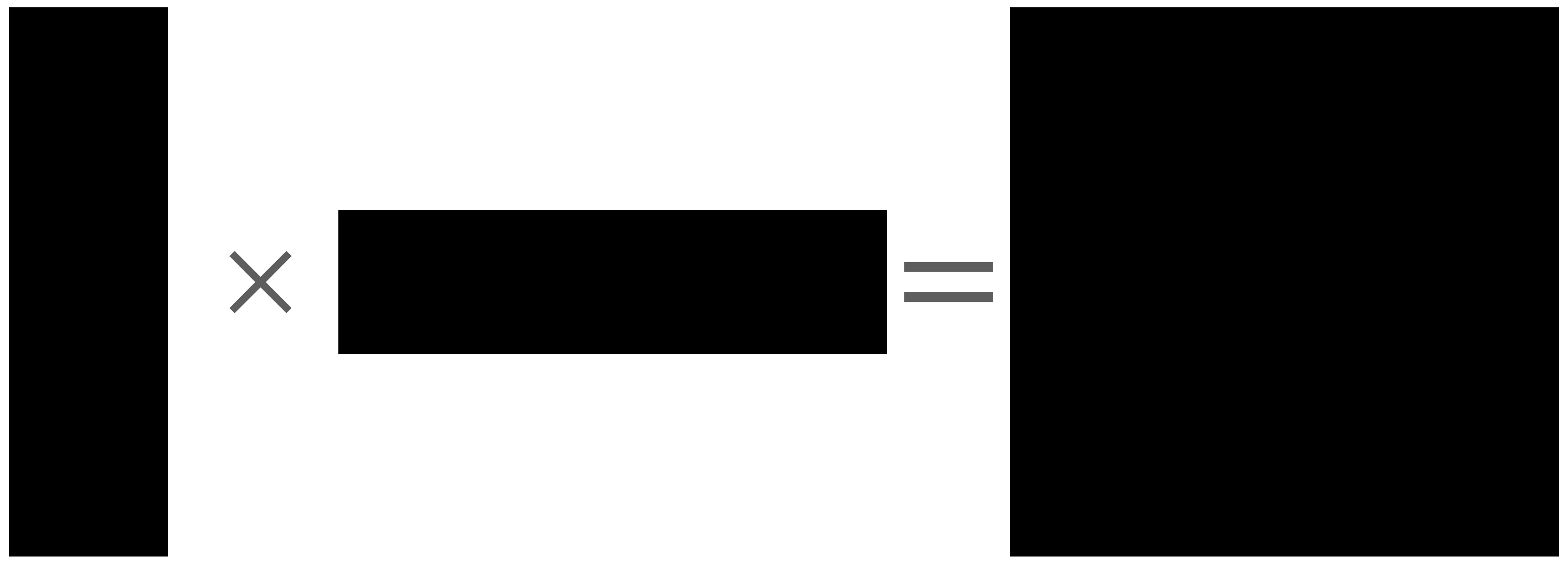
$$\mathrm{Proj}_{\mathcal{E}}(x) = \arg\min_{a \in \mathcal{E}} \ (a - x)^\top \widehat{\Sigma}^{-1}(a - x)$$

- Learn a parametrized mapping:

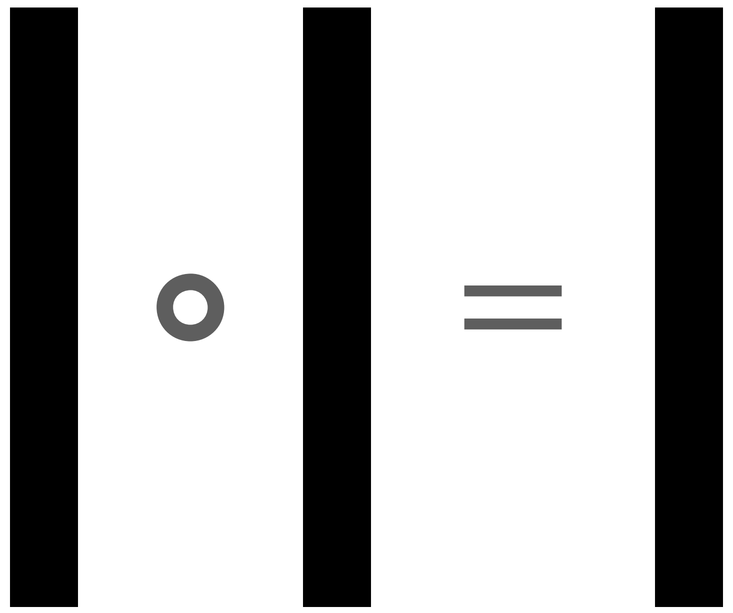
$$f : a \mapsto (I + L(a)R^\top)a + s$$

Low-rank matrices

Parametrized Mapping: Low-rank



$L(a) \in \mathbb{R}^{D\times k},\ R \in \mathbb{R}^{D\times k},\ s \in \mathbb{R}^{D\times 1}.$



$$L_i(a) = \tanh(W_i \circ a + b_i), \forall i \in [k]$$

Parametrized Mapping

- Desirable region is an ellipsoid:

$$\mathcal{E} = \{a : (a-\hat{\mu})^\top \hat{\Sigma}^{-1}(a-\hat{\mu}) \le \rho\}$$

- Projection onto the desirable region

$$\mathrm{Proj}_{\mathcal{E}}(x) = \arg\min_{a\in\mathcal{E}}\ (a-x)^\top \widehat{\Sigma}^{-1}(a-x)$$

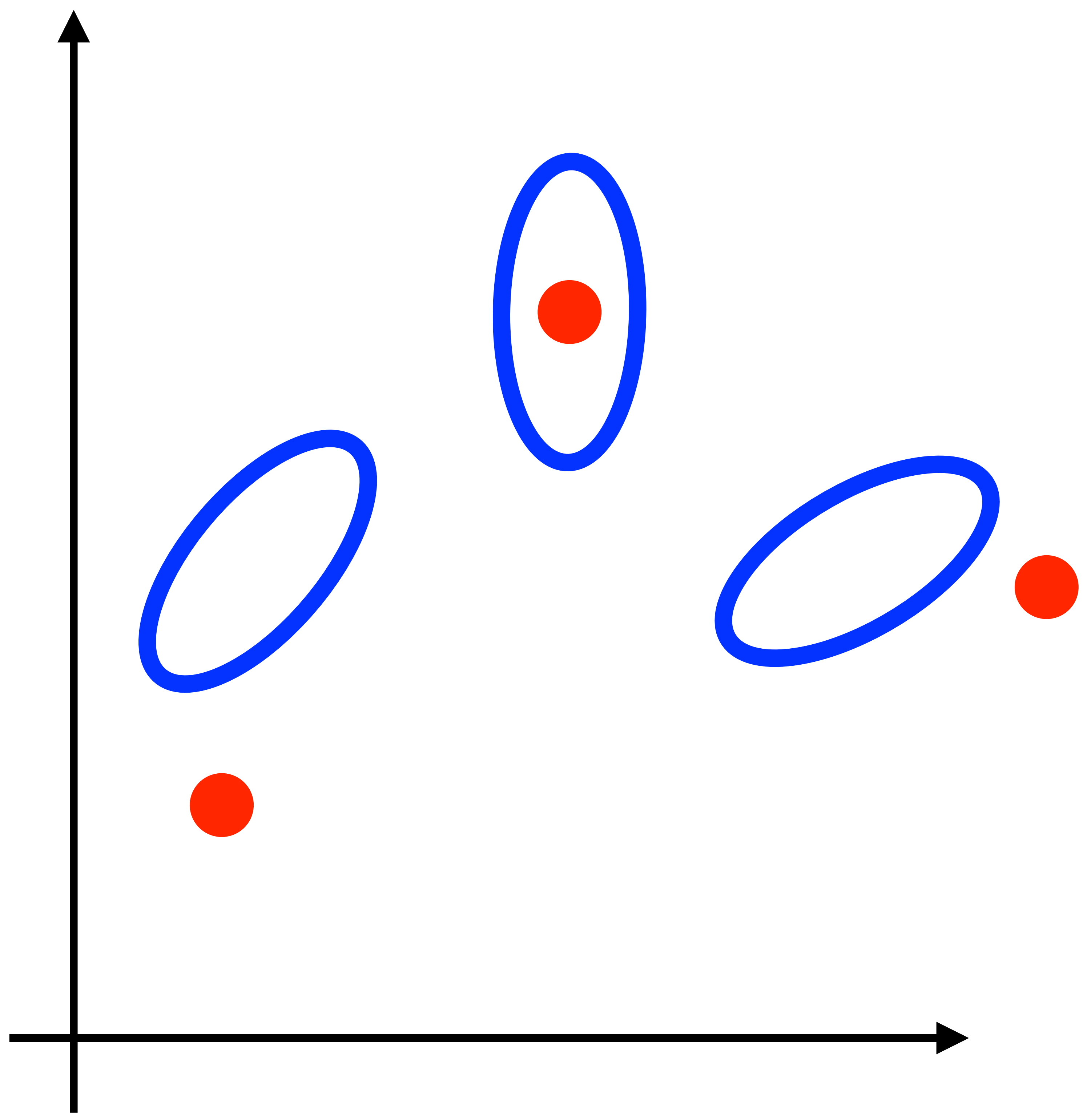
- Learn a parametrized mapping:

$$f : a \mapsto (I + L(a)R^\top)a + s$$

Low-rank matrices

- Low-rank, nonlinear, stable training

From Model to Optimization Problem



Activations for different questions

- Desirable region is an ellipsoid:

$$\mathcal{E} = \{a : (a - \hat{\mu})^\top \hat{\Sigma}^{-1}(a - \hat{\mu}) \leq \rho\}$$

- Projection onto the desirable region

$$\mathrm{Proj}_{\mathcal{E}}(x) = \arg\min_{a \in \mathcal{E}} \; (a - x)^\top \widehat{\Sigma}^{-1}(a - x)$$

- Learn a parametrized mapping:

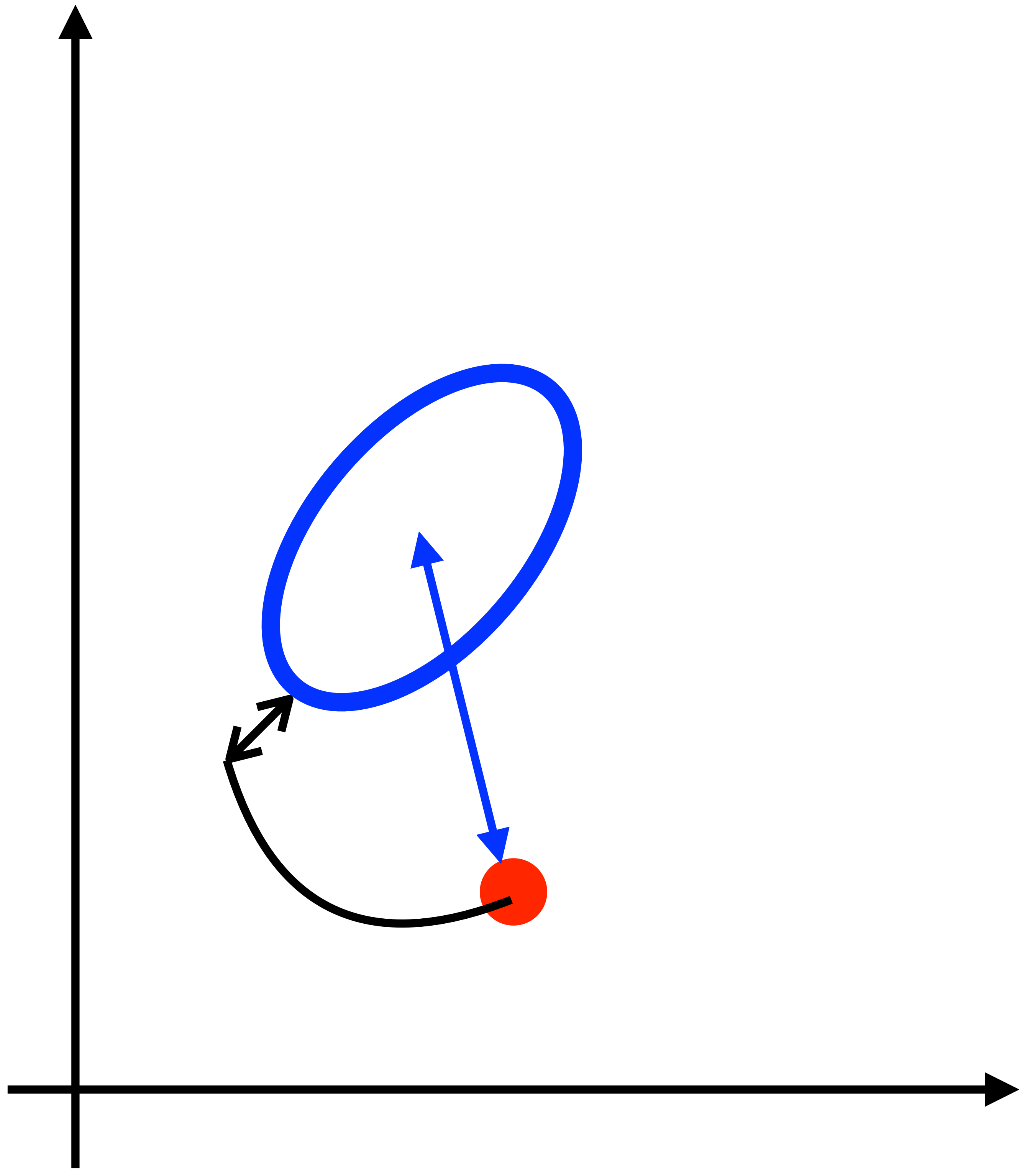
$$f : a \mapsto (I + L(a)R^\top)a + s$$

Low-rank matrices

- Learn a parametrized mapping:

$$\min_f \sum_q \sum_{i \in \mathcal{B}(q) \cup \mathcal{G}(q)} c_q(f(a_i), \mathrm{Proj}_{\mathcal{E}_q}(f(a_i)))$$

Optimization Problem: Discussion



Optimization Problem

- Optimization problem:

$$\min_f \sum_q \sum_{i \in \mathcal{B}(q) \cup \mathcal{G}(q)} c_q(f(a_i), \mathrm{Proj}_{\mathcal{E}_q}(f(a_i)))$$

- Why ellipsoid?
- Why such loss function?
- Potential challenges in solving the problem?

- Equivalent formulation:

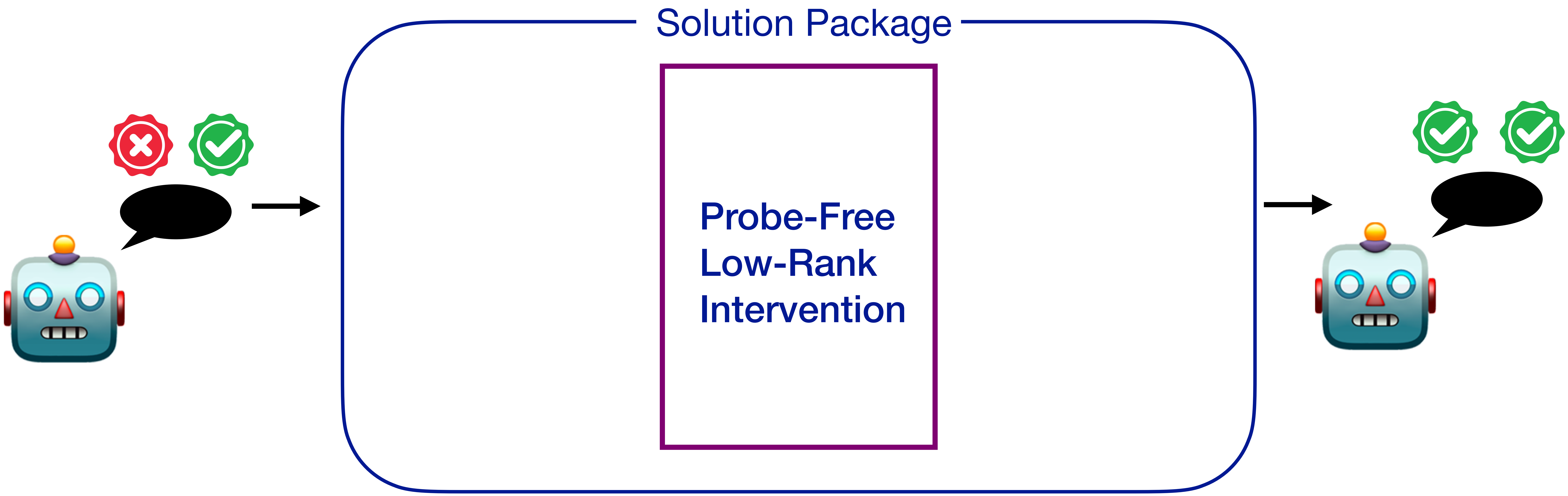
$$\min_f \sum_q \sum_{i \in \mathcal{B}(q) \cup \mathcal{G}(q)} \left[\left(\sqrt{c_q(f(a_i), \hat{\mu}_q)} - \sqrt{\rho_q} \right)_+ \right]^2$$

Performance

Methods	True * Info (%) ↑	True (%) ↑	MC1 ↑	MC2 ↑	CE ↓	KL ↓
Unintervened	51.87	59.86	35.38	53.32	2.31	0.00
ITI	57.02	63.04	37.46	55.59	2.32	0.17
FLORAIN (ours)	**60.68**	**67.70**	**39.65**	**59.57**	2.35	0.18
FSP	55.97	58.63	40.76	57.84	2.31	0.00
FSP + ITI	56.78	59.24	41.50	59.01	2.33	0.13
FSP + FLORAIN (ours)	**61.14**	**62.45**	**44.52**	**61.48**	2.37	0.16

(c) Llama2-chat-13B

Take away



- Keywords: ellipsoid model + low-rank mapping + optimization problem
- Future directions: region modeling in LM