Fundamental Limits of Prompt Compression: A Rate-Distortion Framework for Black-Box Language Models

Adway Girish*, Alliot Nagle*

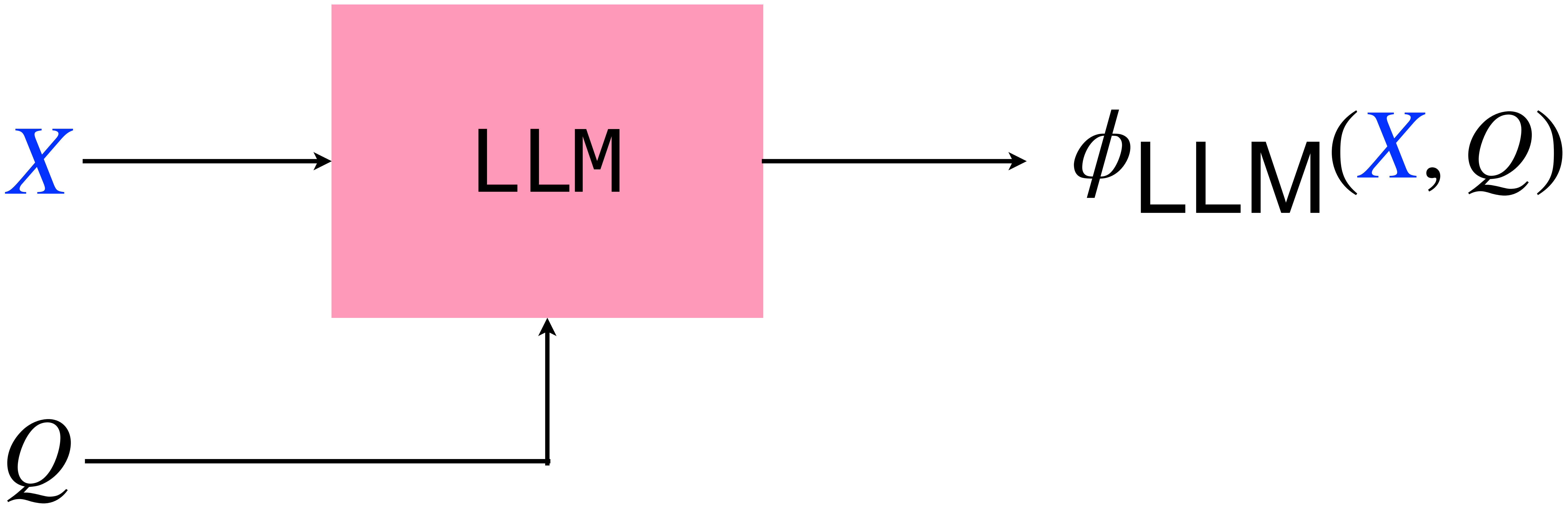
joint work with Marco Bondaschi, Michael Gastpar, Ashok Vardhan Makkuva†, Hyeji Kim†

EPFL

TEXAS
The University of Texas at Austin

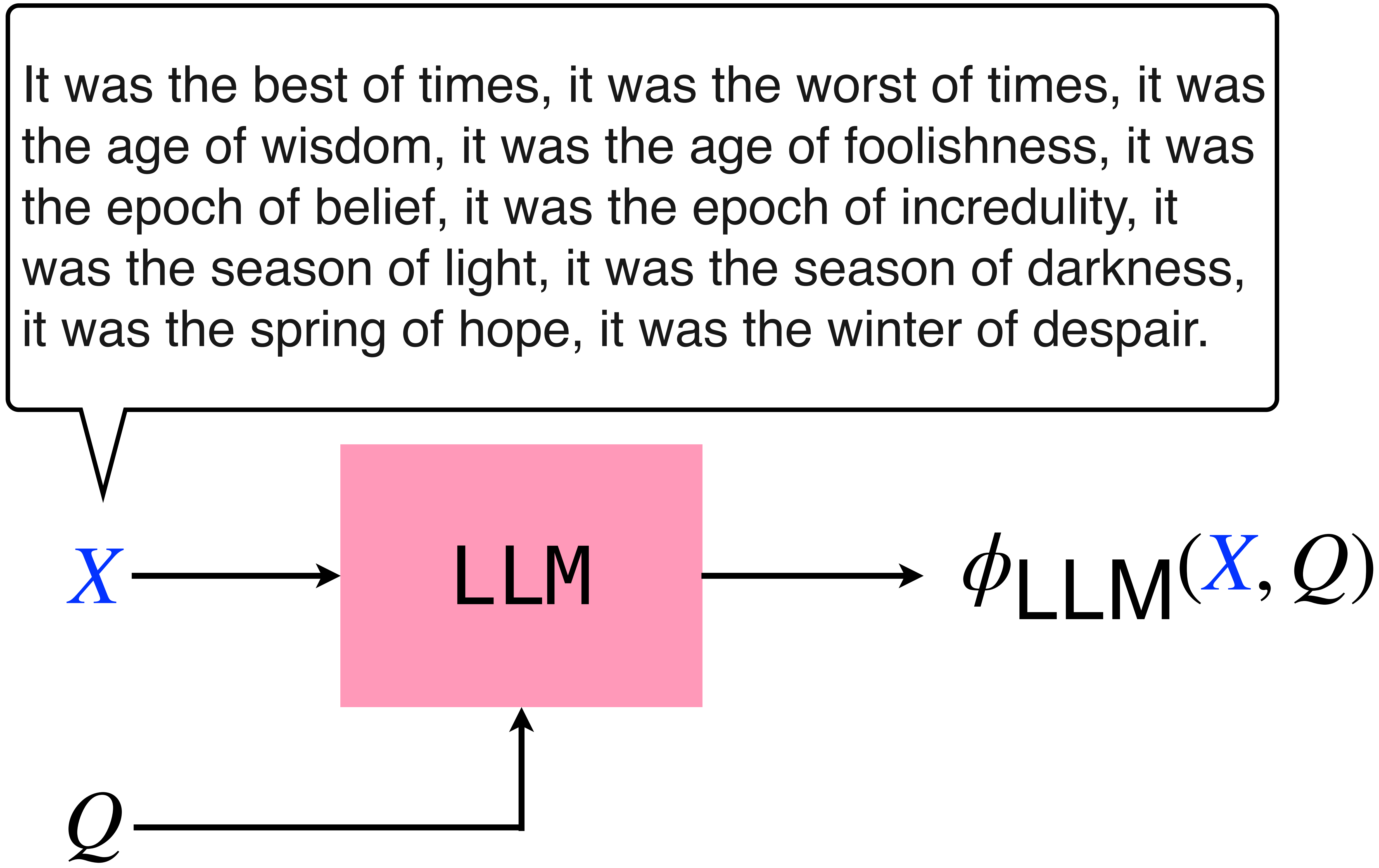
July 27, 2024

A review of prompting

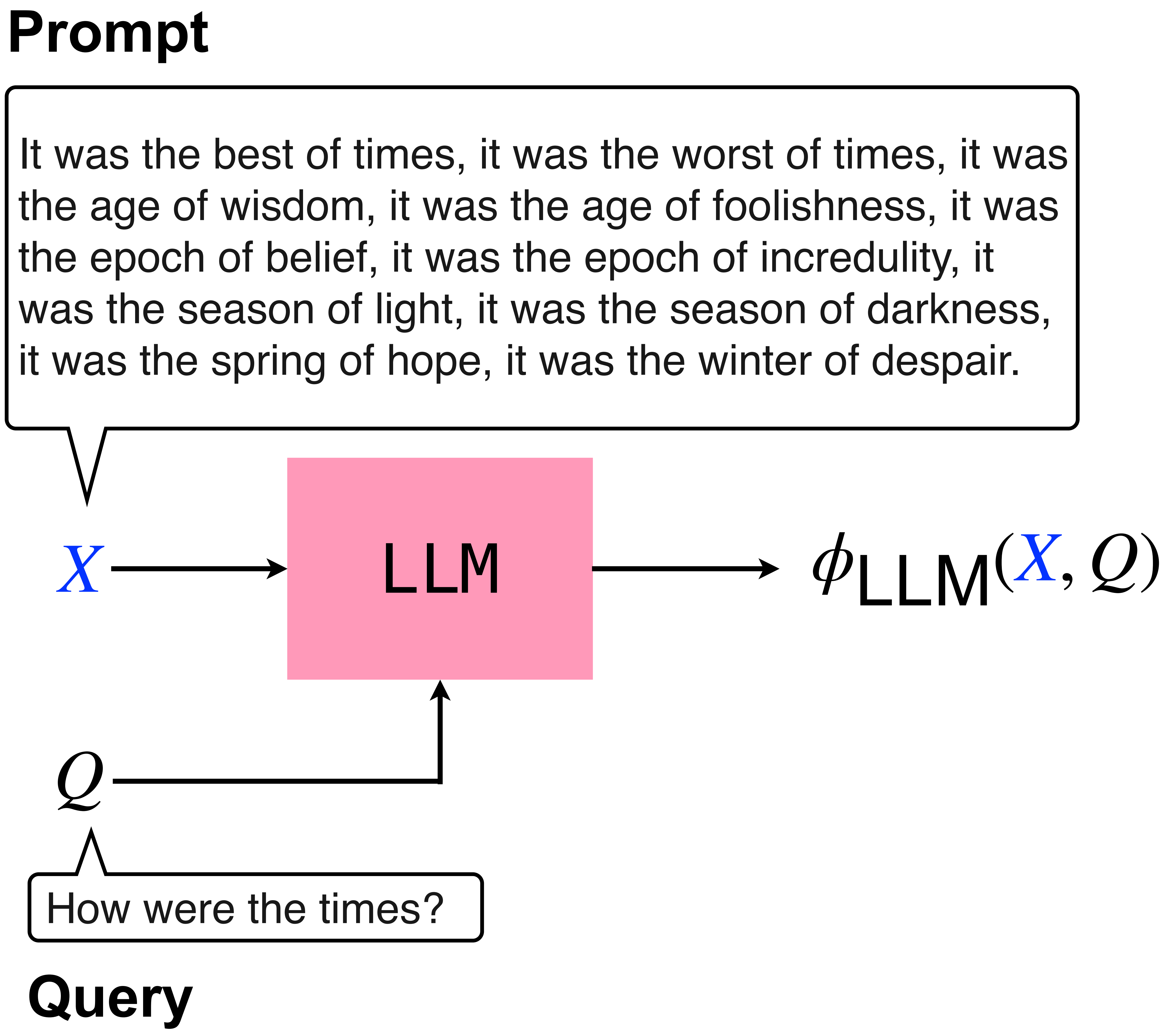


A review of prompting

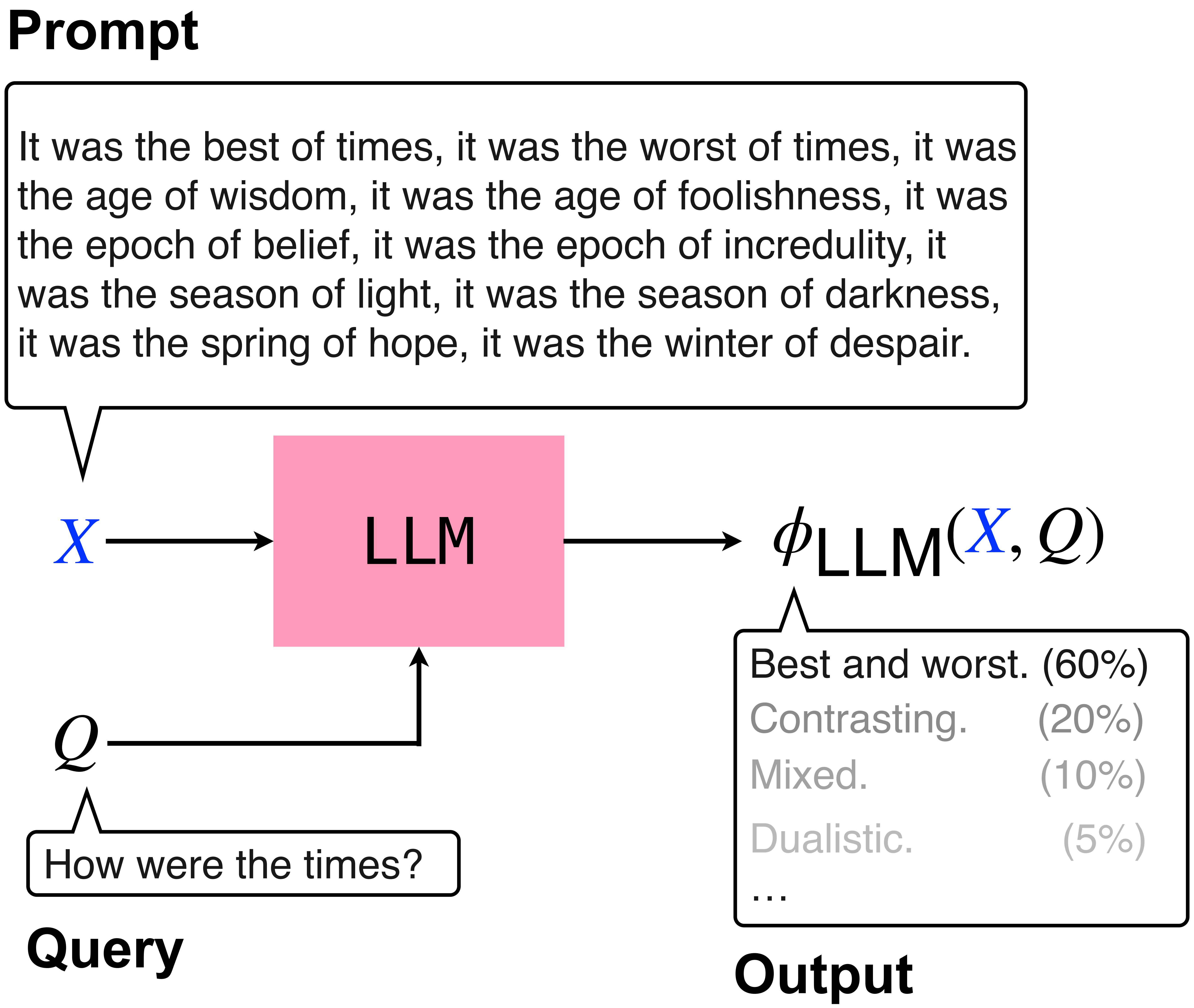
Prompt



A review of prompting



A review of prompting



What is prompt compression?

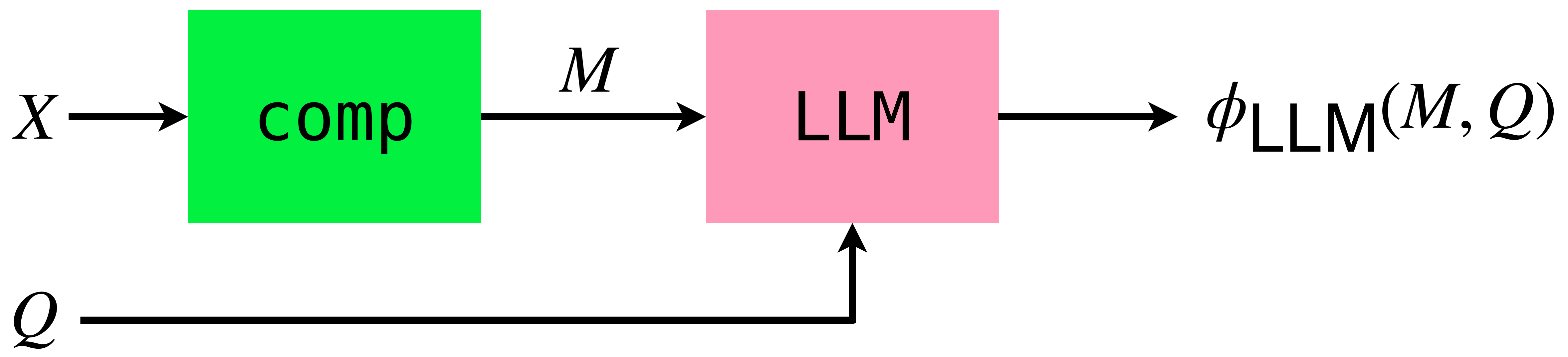
What is prompt compression?

Redundant tokens are removed from the prompt

What is prompt compression?

Redundant tokens are removed from the prompt

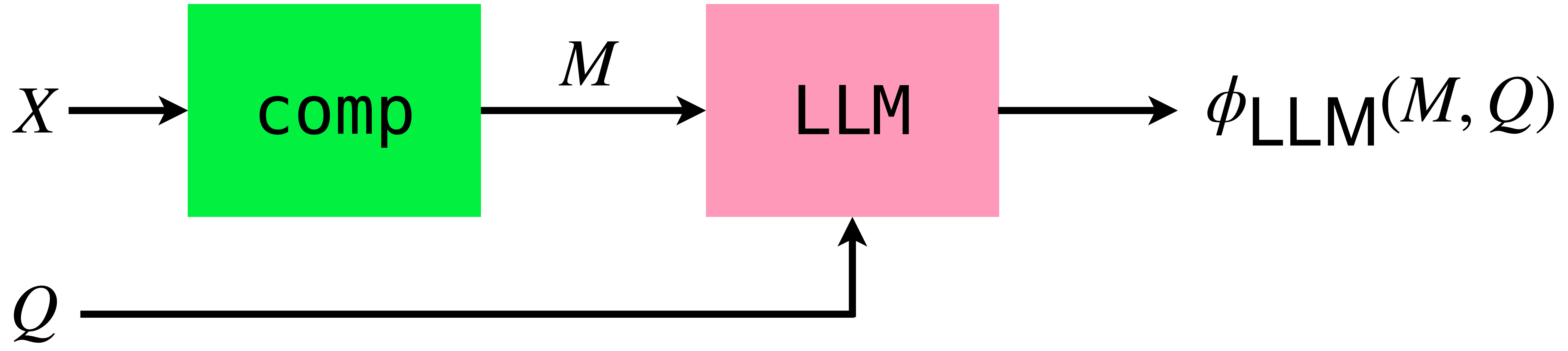
Query-agnostic



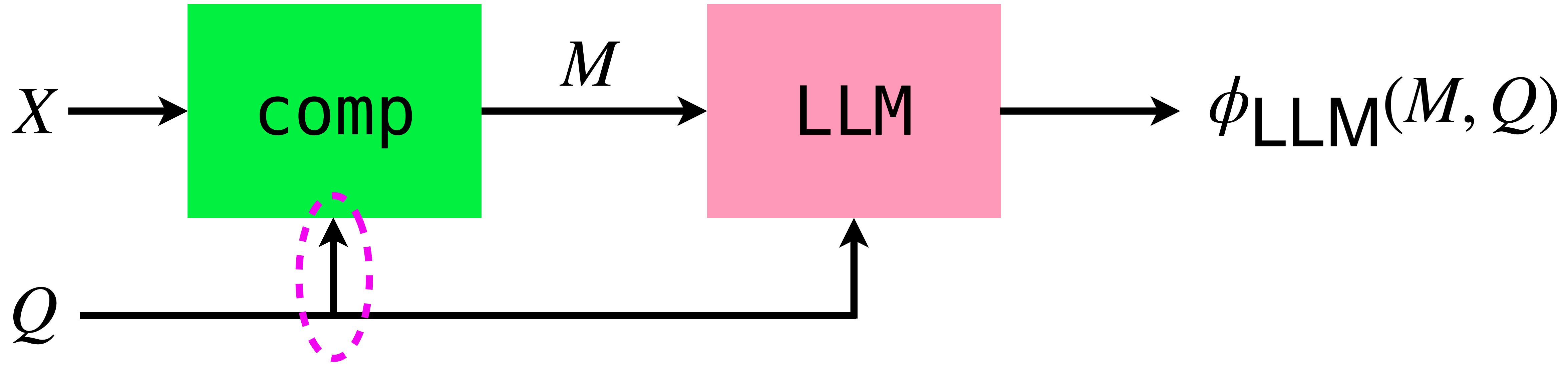
What is prompt compression?

Redundant tokens are removed from the prompt

Query-agnostic

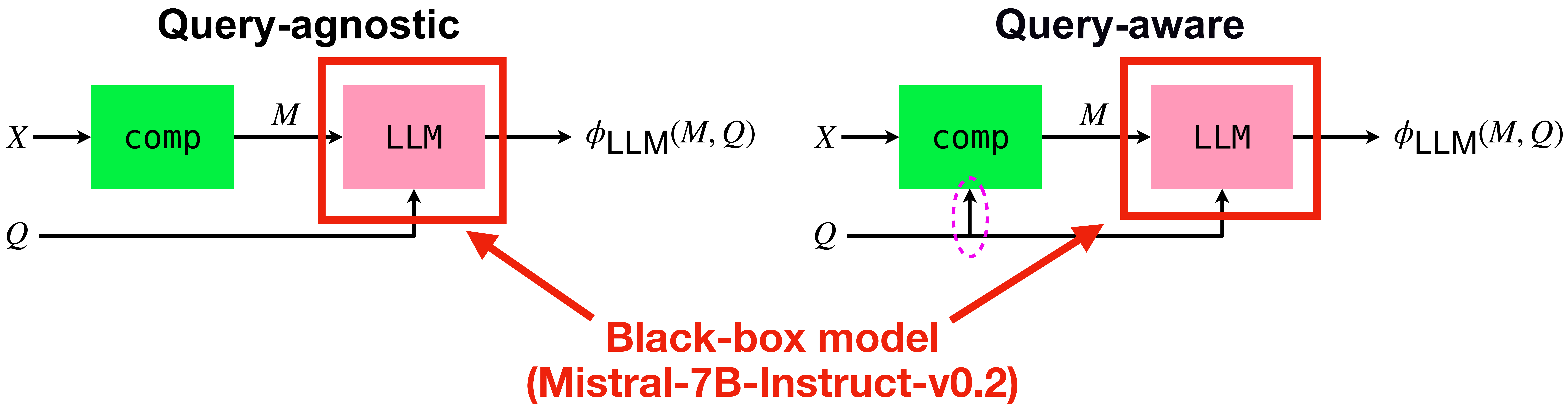


Query-aware



What is prompt compression?

Redundant tokens are removed from the prompt



Why does prompt compression matter?

Why does prompt compression matter?

1. Reduce input size ➔ reduce time and memory costs

Why does prompt compression matter?

1. **Reduce input size ➔ reduce time and memory costs**

2. **“Lost in the middle” issue is mitigated**

Liu, N., et al. "Lost in the Middle: How Language Models Use Long Contexts," in *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 157–173, 2024.

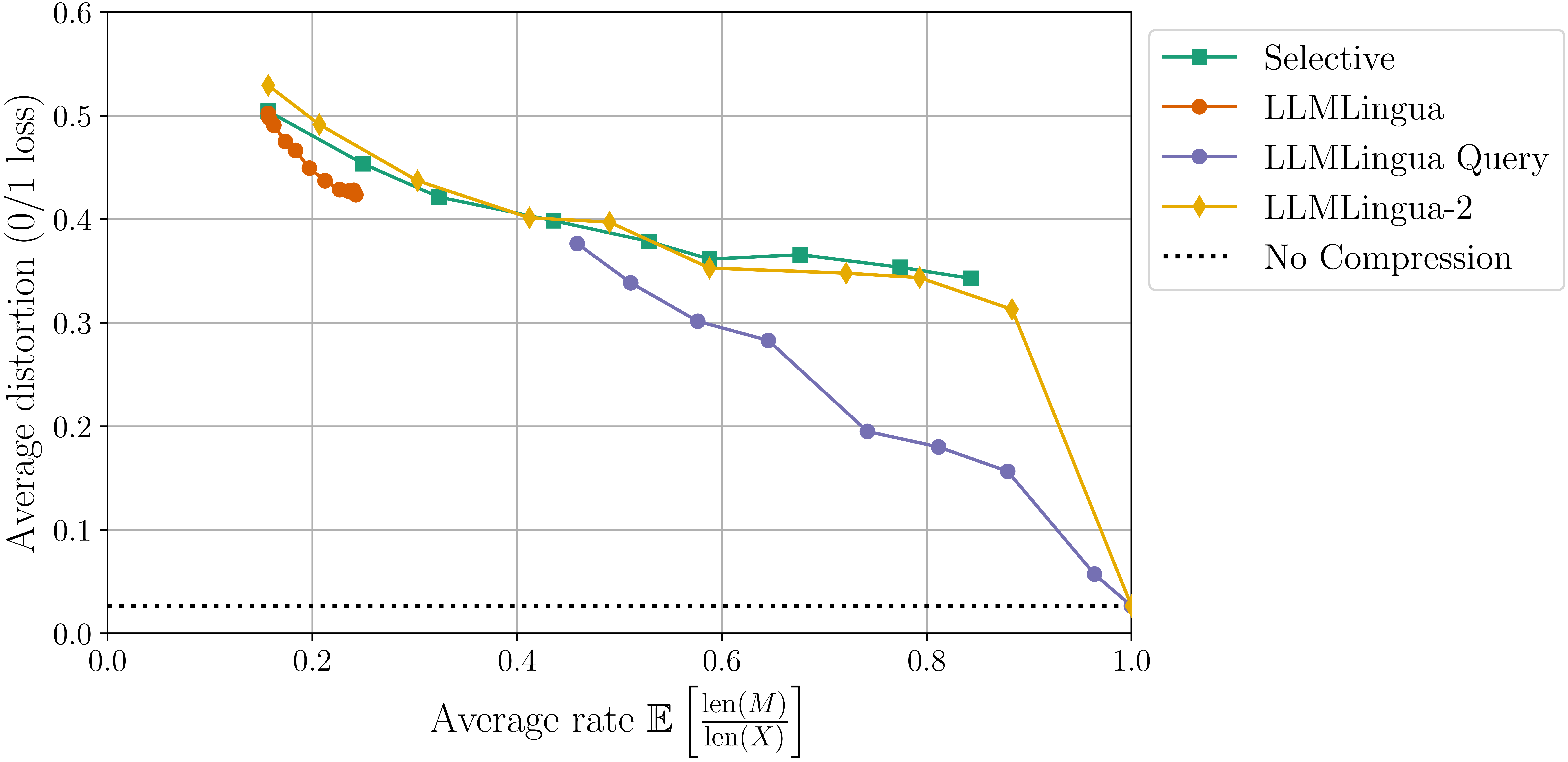
Peng Xu, undefined., et al, "Retrieval meets Long Context Large Language Models," in *The Twelfth International Conference on Learning Representations*, 2024.

Jiang, H., et al, "LongLLMLingua: Accelerating and Enhancing LLMs in Long Context Scenarios via Prompt Compression," in *The 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, 2023.

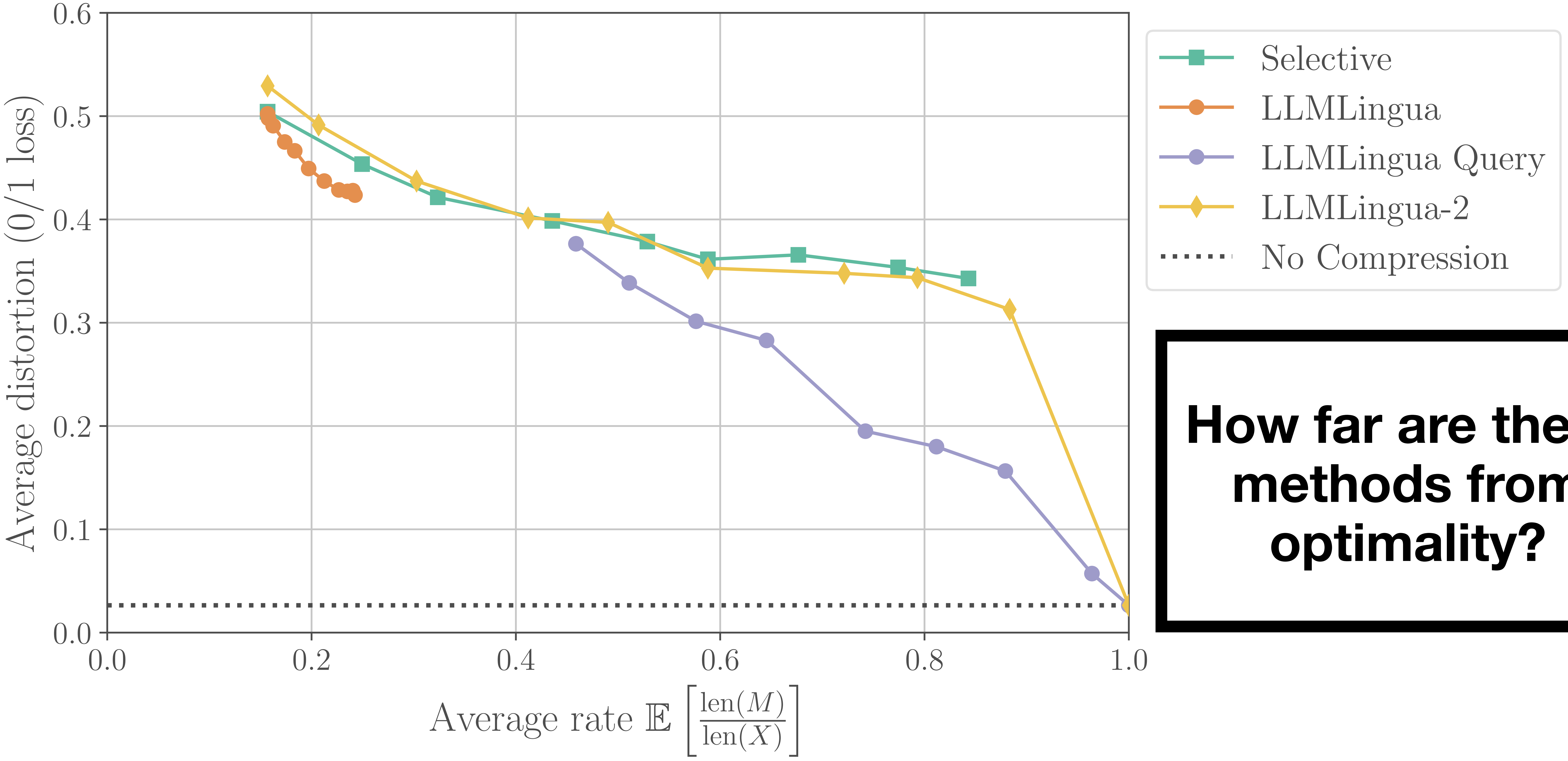
Example

Prompt	**Query**	**Answer**
110011111	Count the number of 1s.	7
11111	Count the number of 0s.	0
00000111	Compute the parity.	1
11011111	What is the length of the longest subsequence of 0s or 1s?	5
0110	Is the binary string a palindrome?	Yes
1100111100	Count the number of transitions from 0 to 1 and 1 to 0.	3
111111	Predict the next bit.	1

Existing compression schemes



Existing compression schemes



How far are these methods from optimality?

Our Main Contributions

Our Main Contributions

1. We introduce a rate-distortion framework to formulate the prompt compression problem

Our Main Contributions

1. We introduce a rate-distortion framework to formulate the prompt compression problem

2. We show a large gap between current methods and optimality

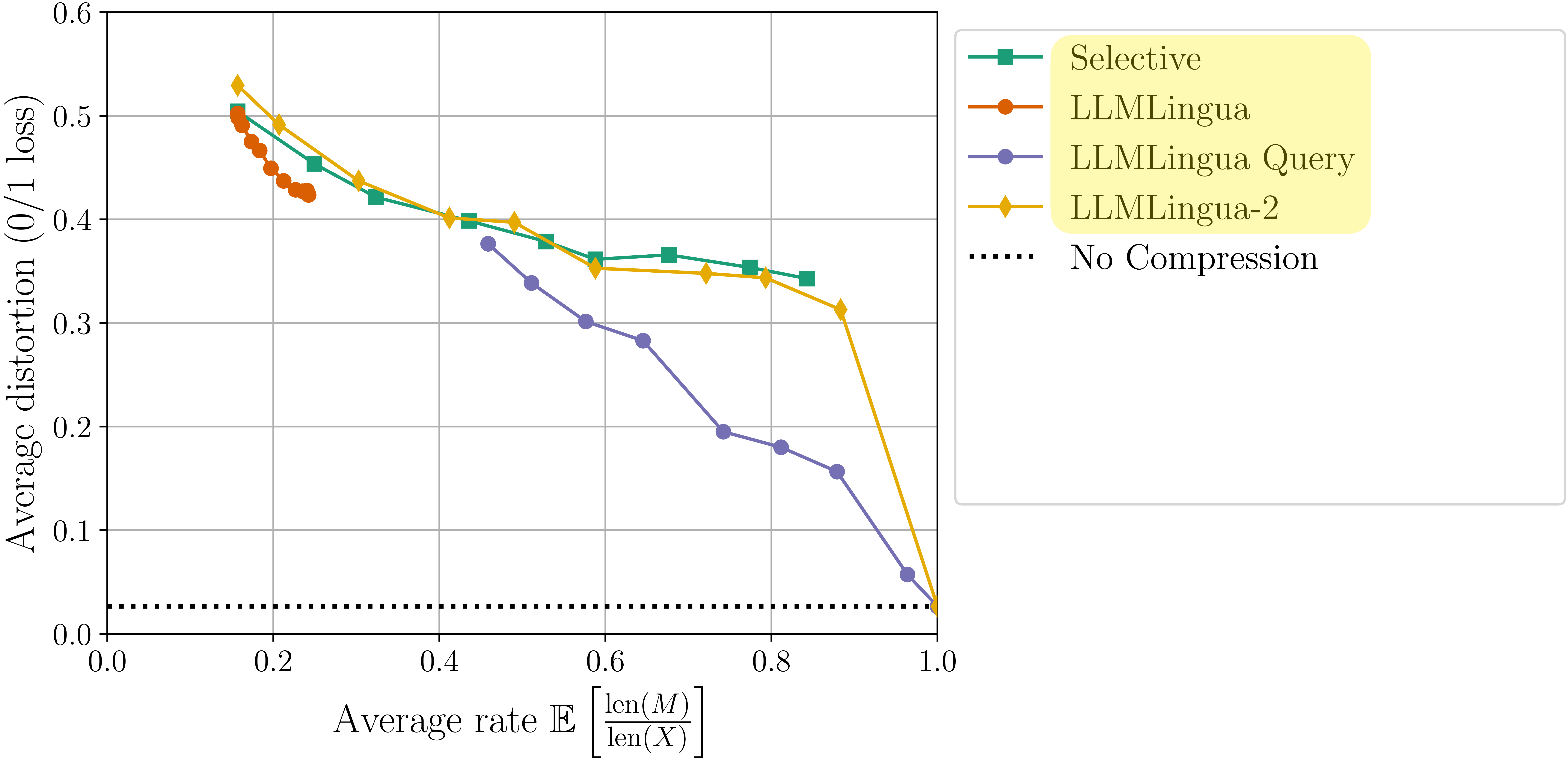
Our Main Contributions

1. We introduce a rate-distortion framework to formulate the prompt compression problem

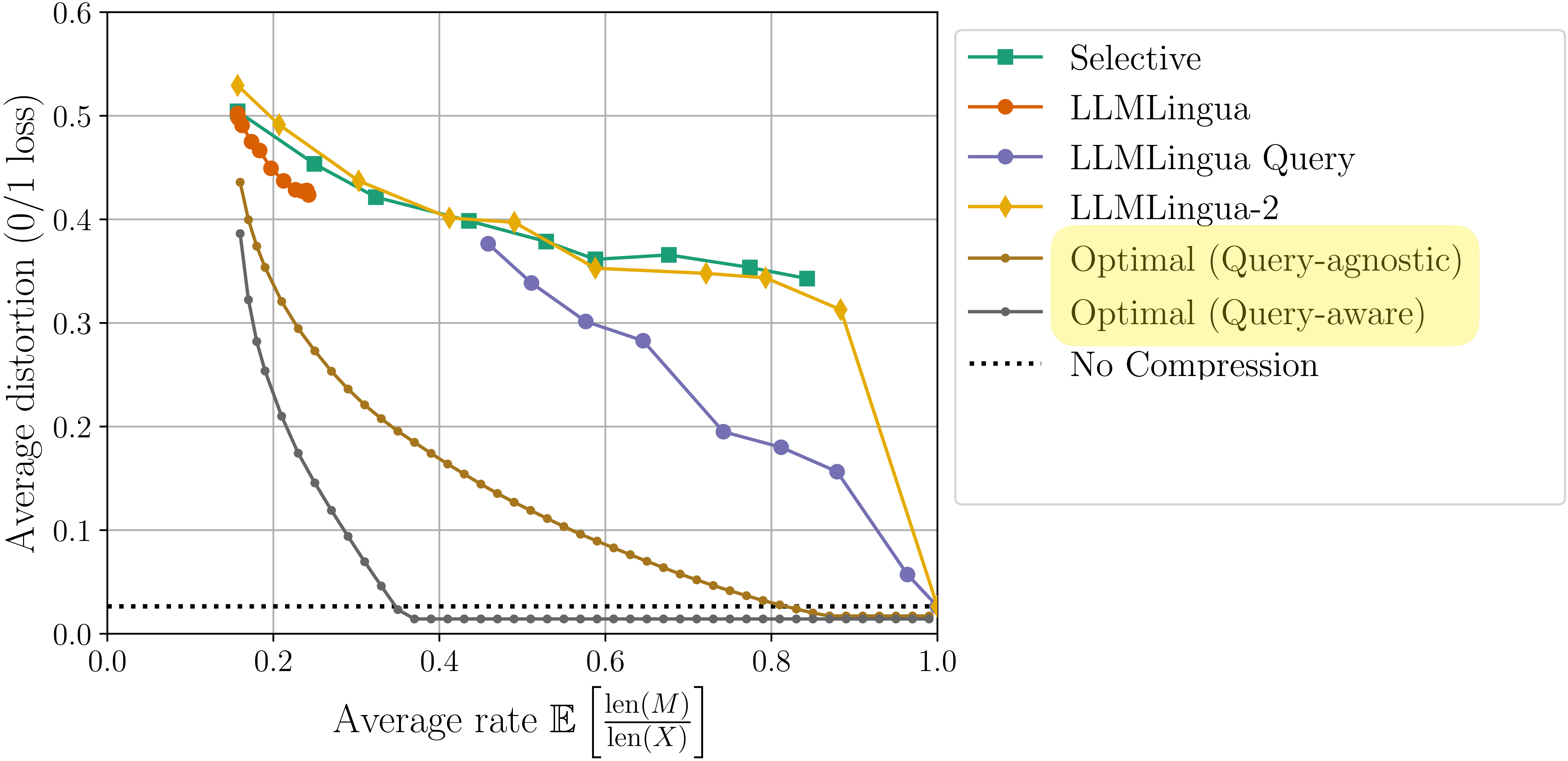
2. We show a large gap between current methods and optimality

3. We adapt an existing method to partially close the gap

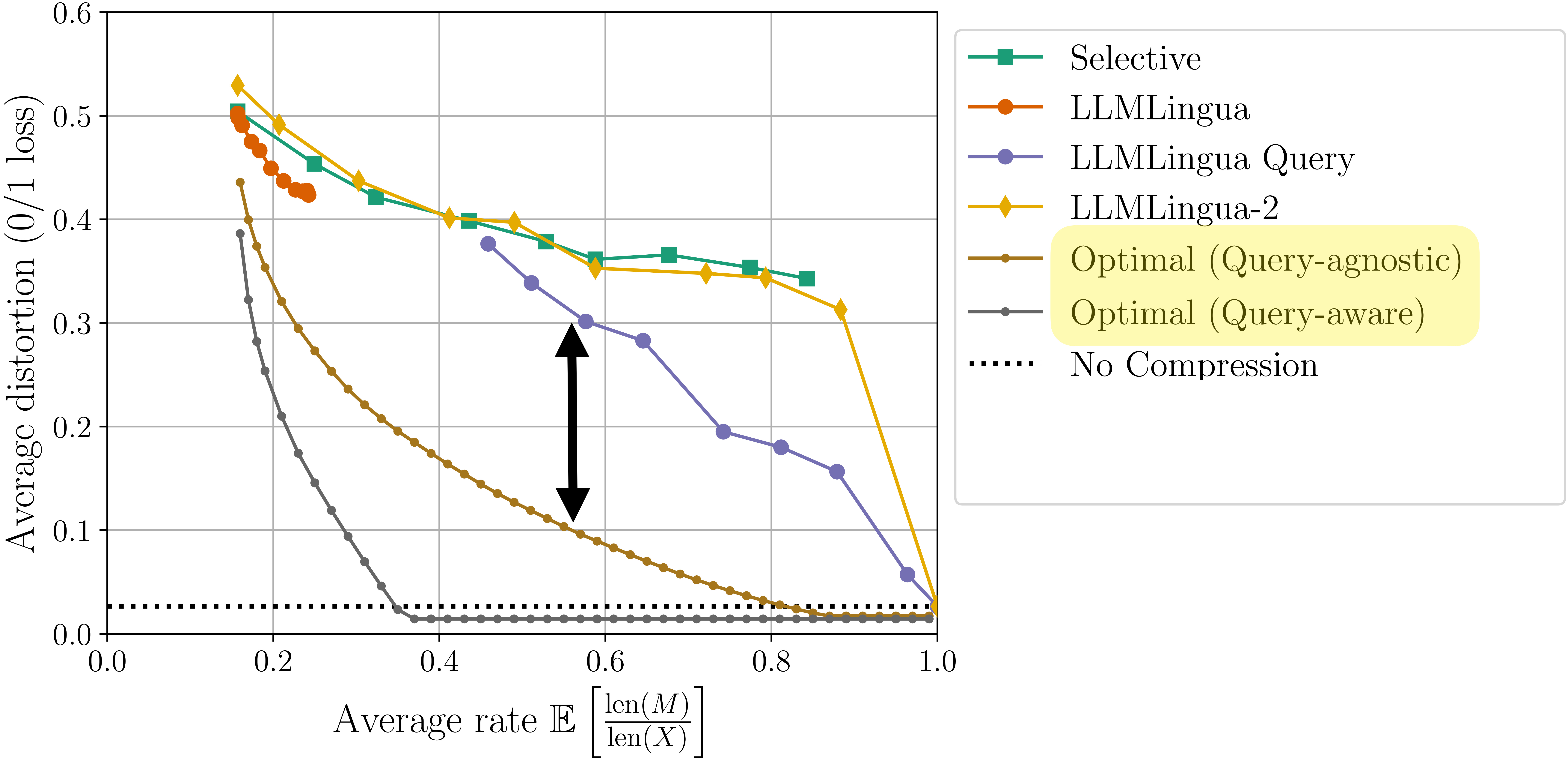
Existing compression schemes



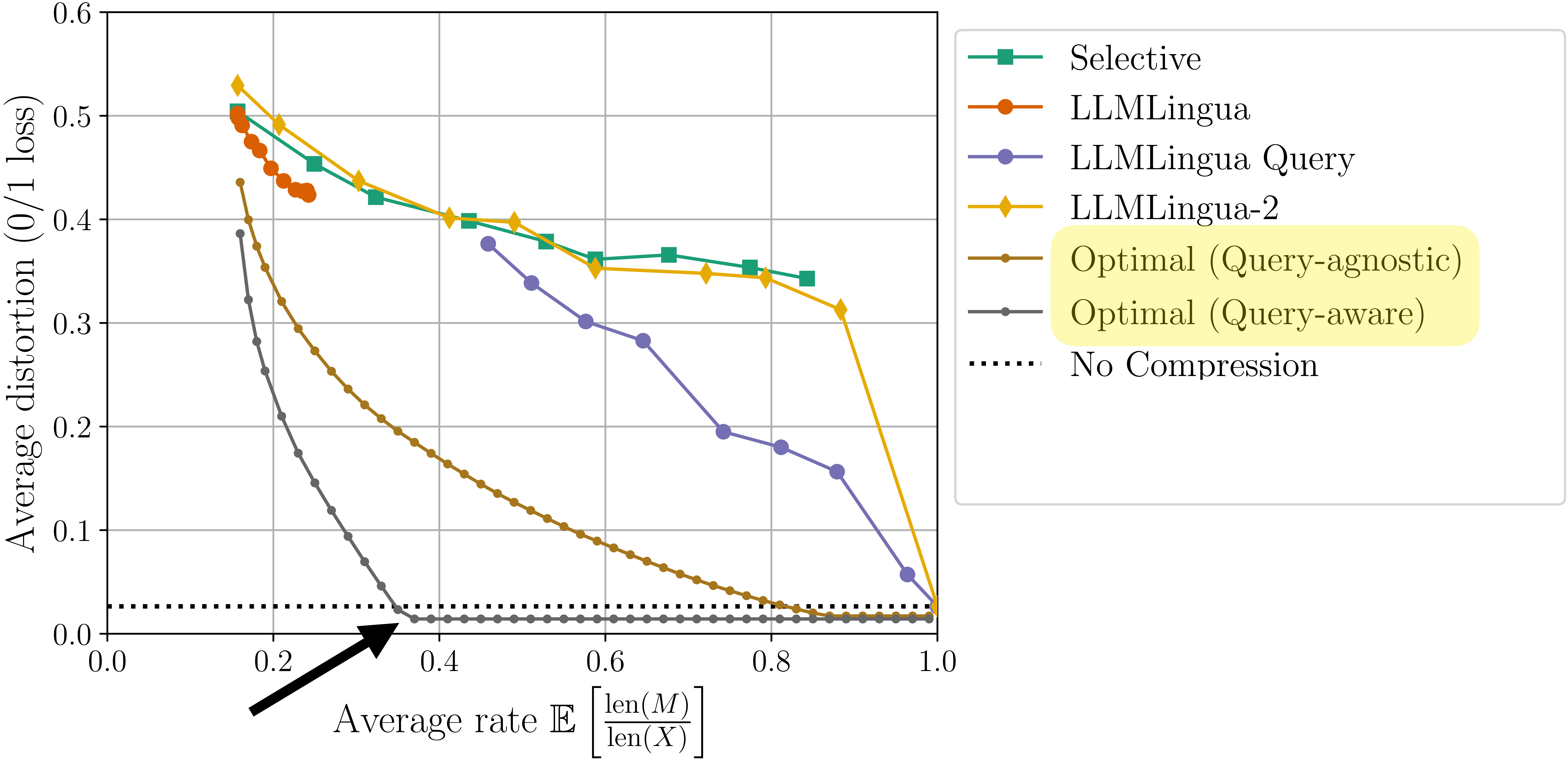
The gap to optimality is large



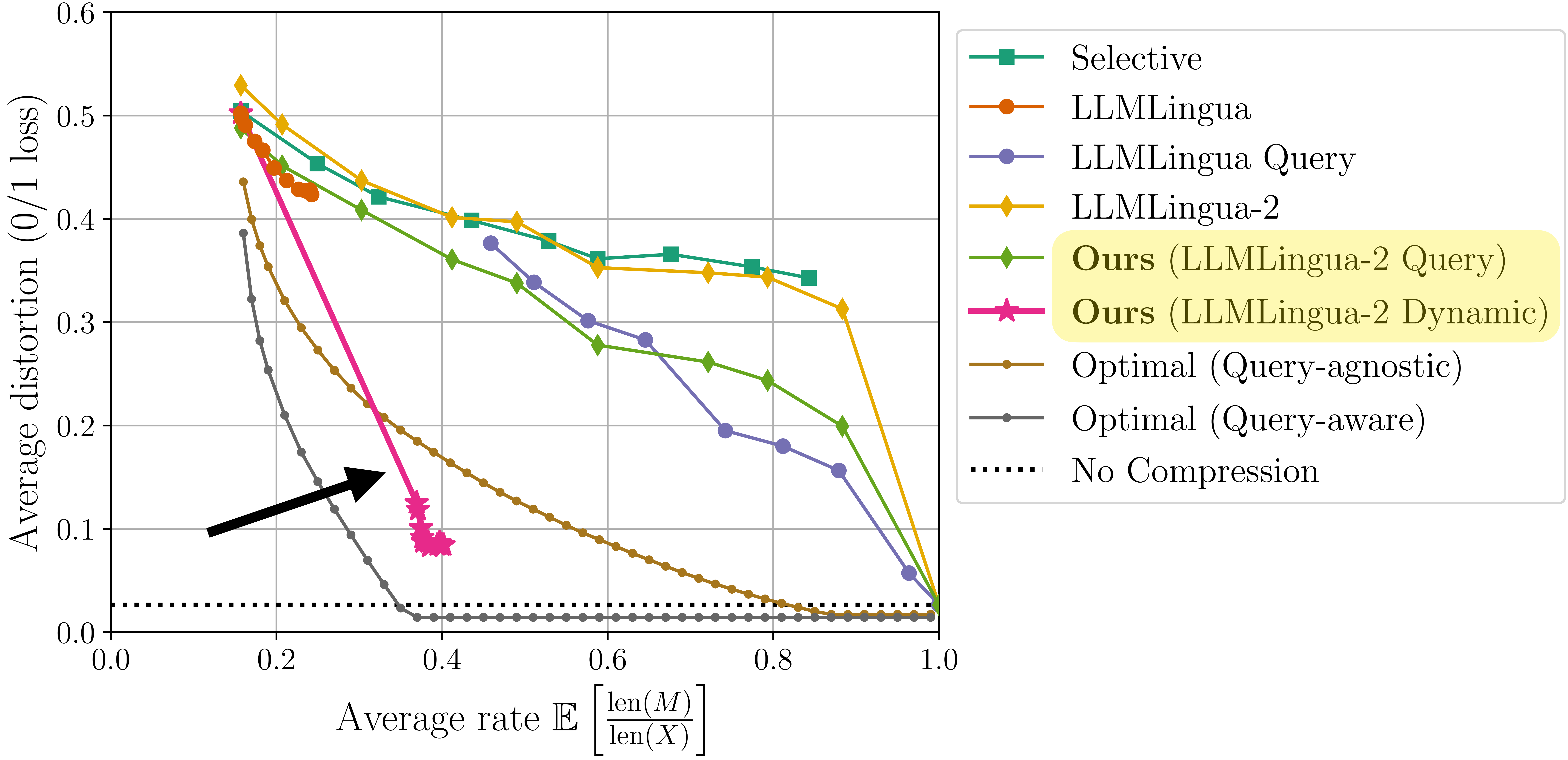
The gap to optimality is large



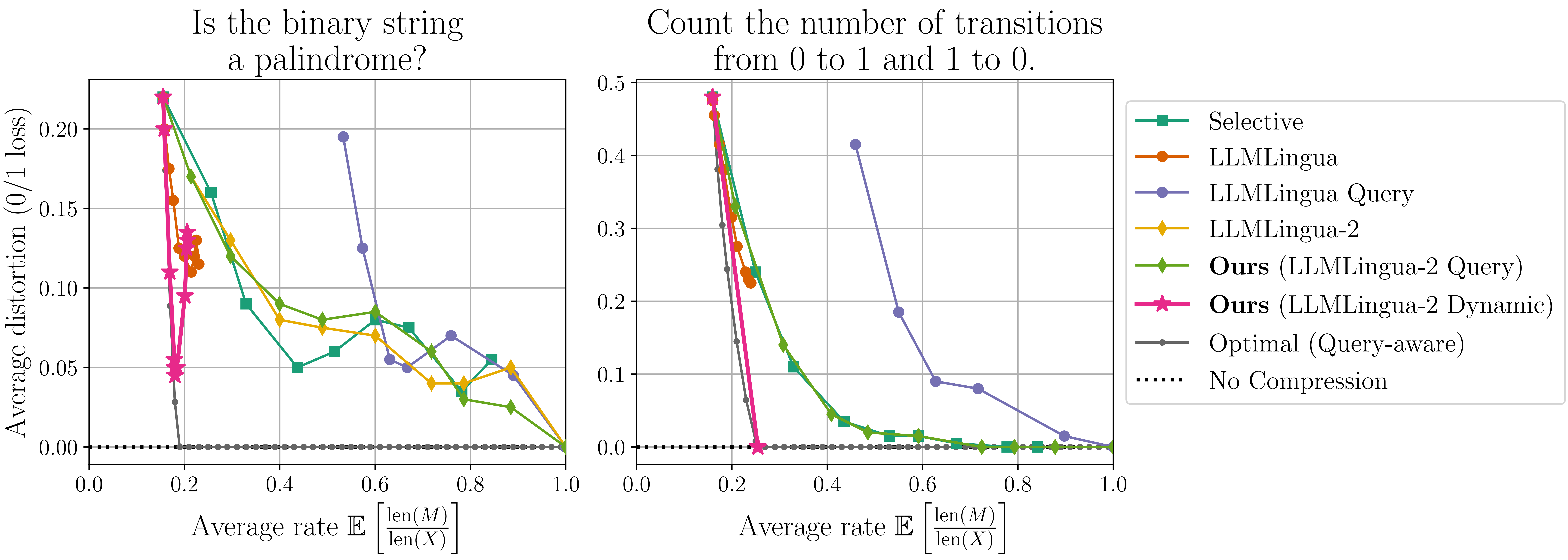
The gap to optimality is large



Significant improvement!



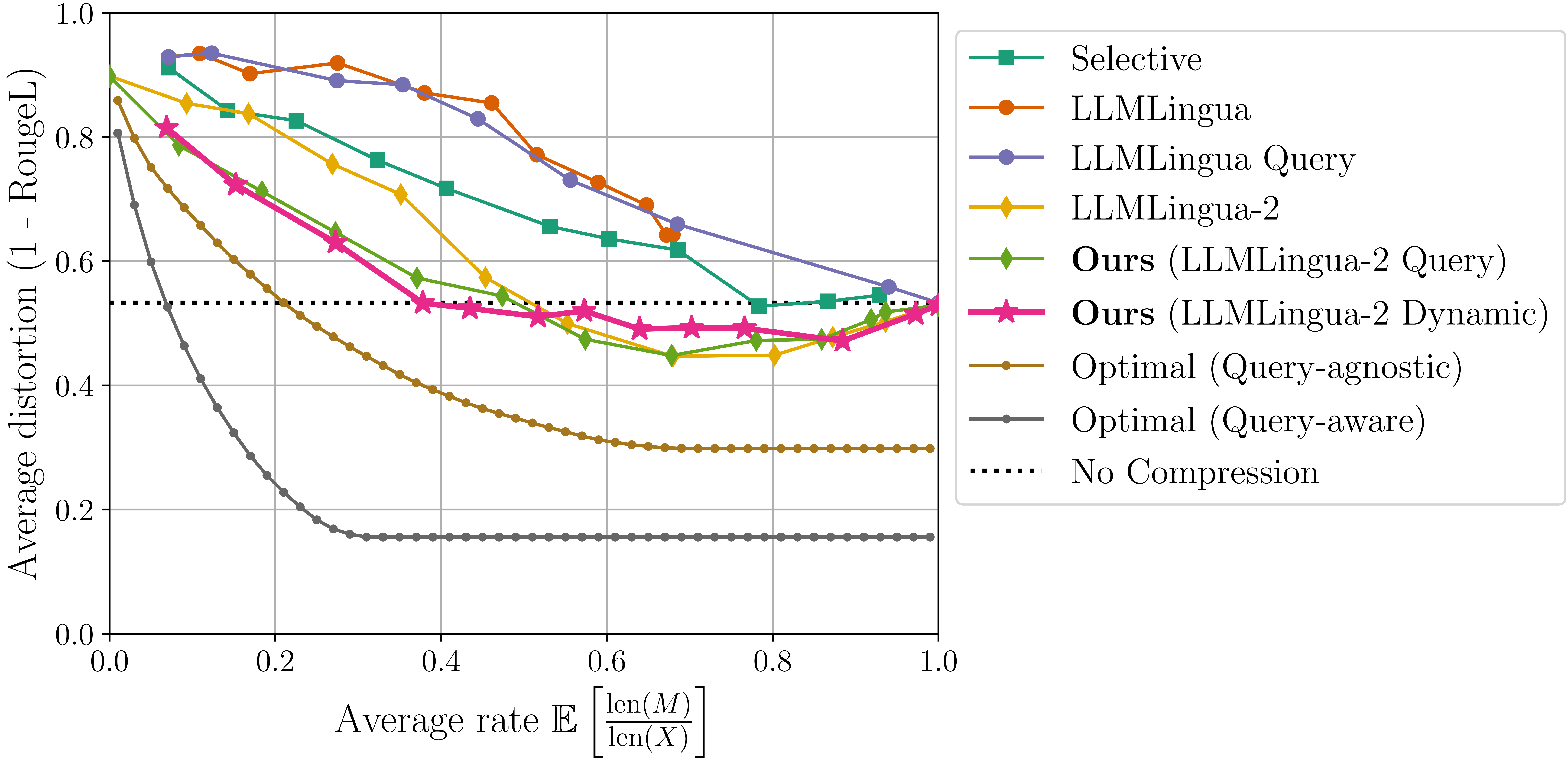
Our method can match optimality



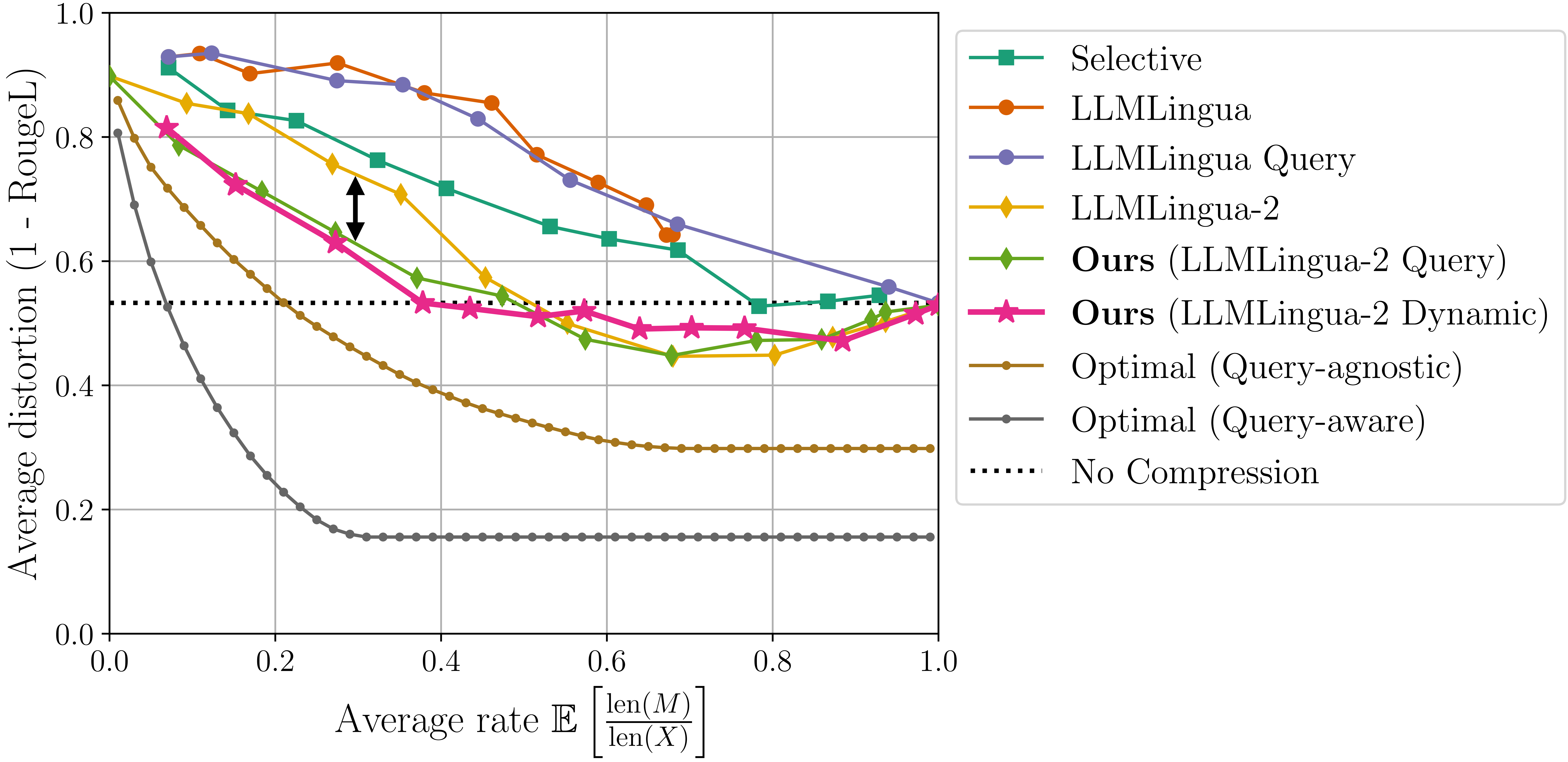
Examples from our natural language dataset

Prompt	Query	Answer
After dinner, the cat chased a mouse around the house.	What was the cat doing?	The cat was chasing a mouse.
The dog barked loudly at the passing mailman on a quiet street.	Where did the barking occur?	On a quiet street.
After school, the child played with toys in the cozy living room.	When was the child playing?	After school.
At the art gallery, the artist painted a colorful mural on the wall.	Where was the painting done?	On the wall at the art gallery.

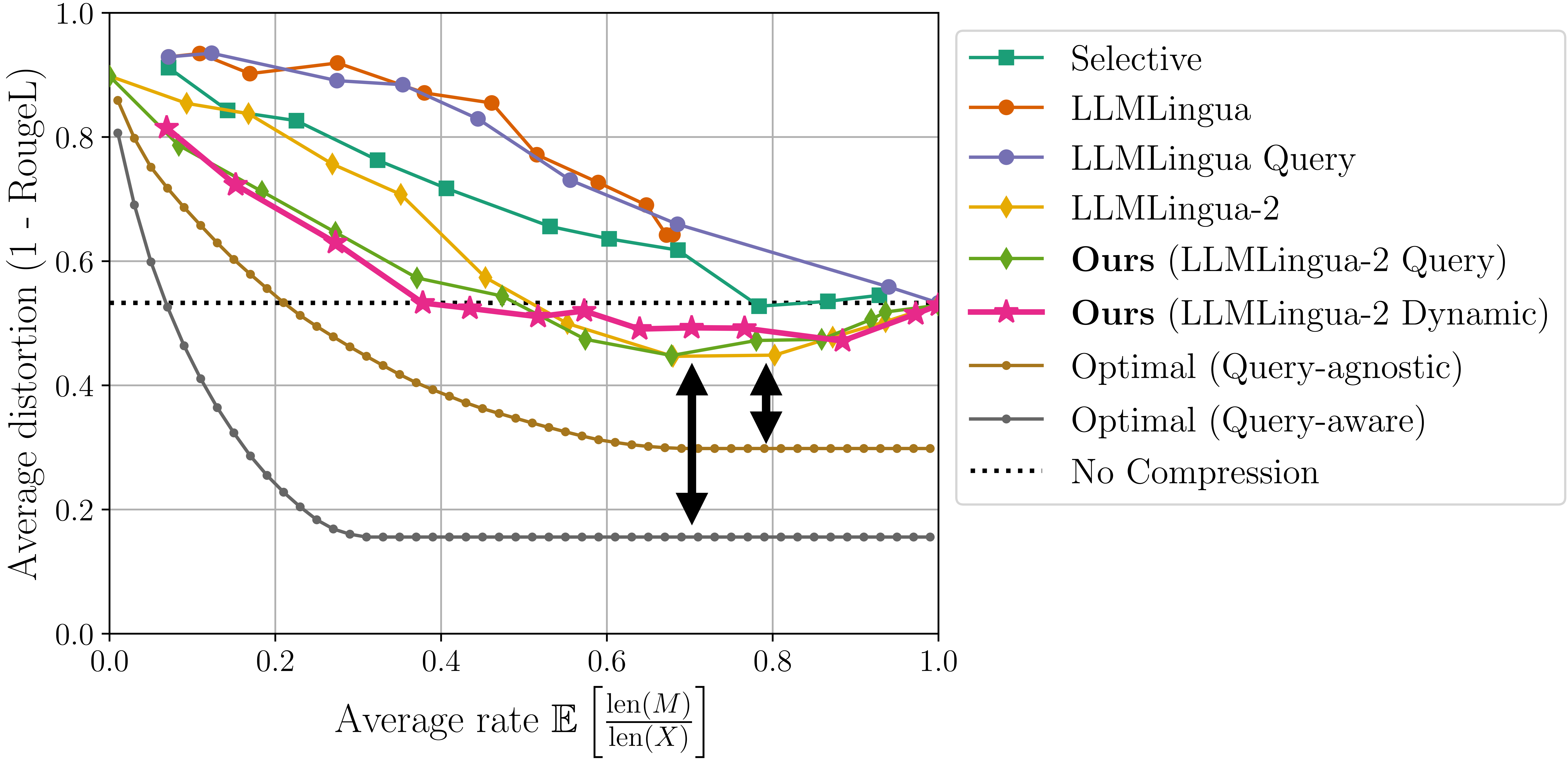
Results on natural language



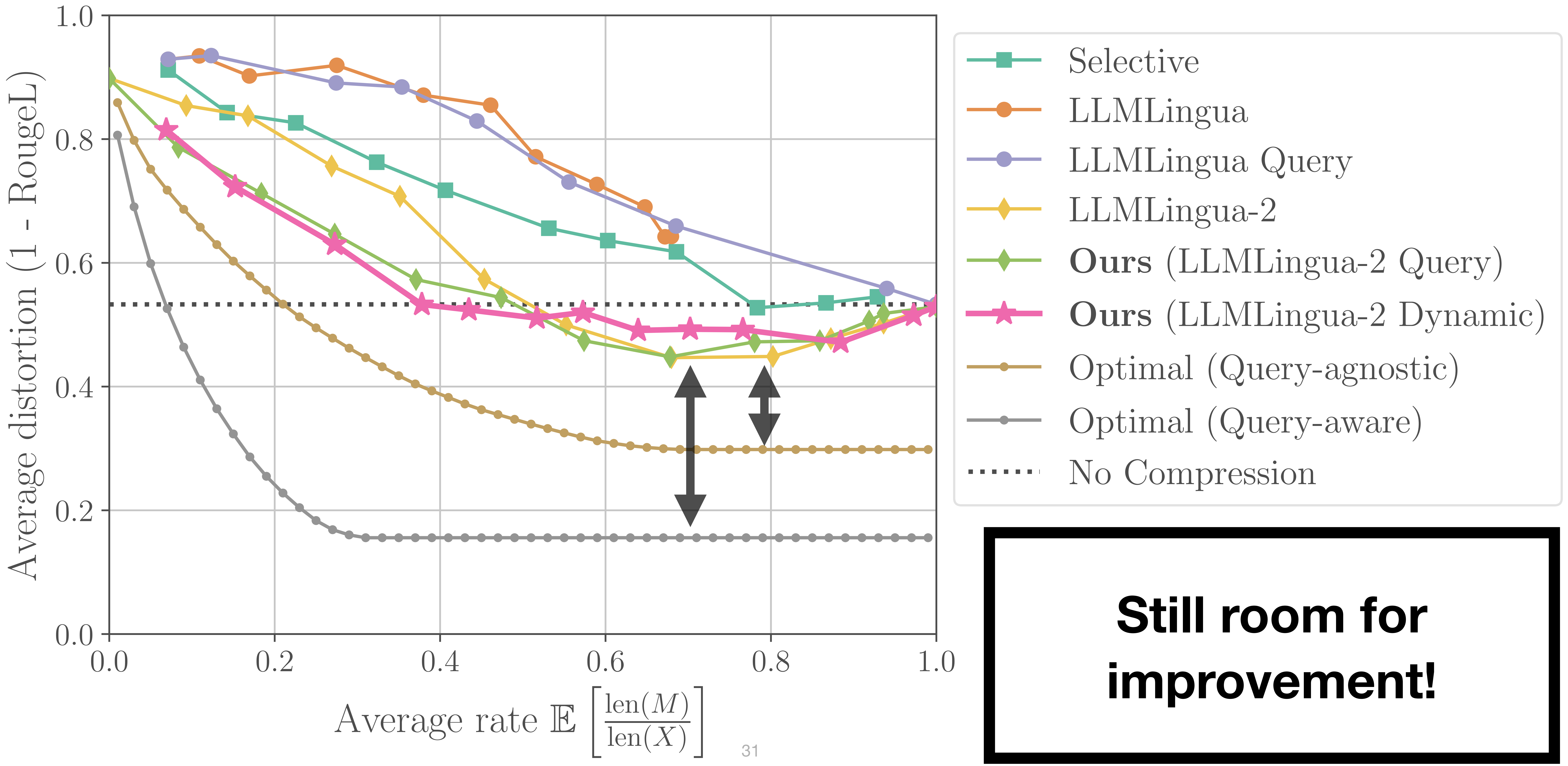
Results on natural language



Results on natural language



Results on natural language



Still room for improvement!

What is our rate-distortion framework?

What is our rate-distortion framework?

How do we compute the optimal RD curves?

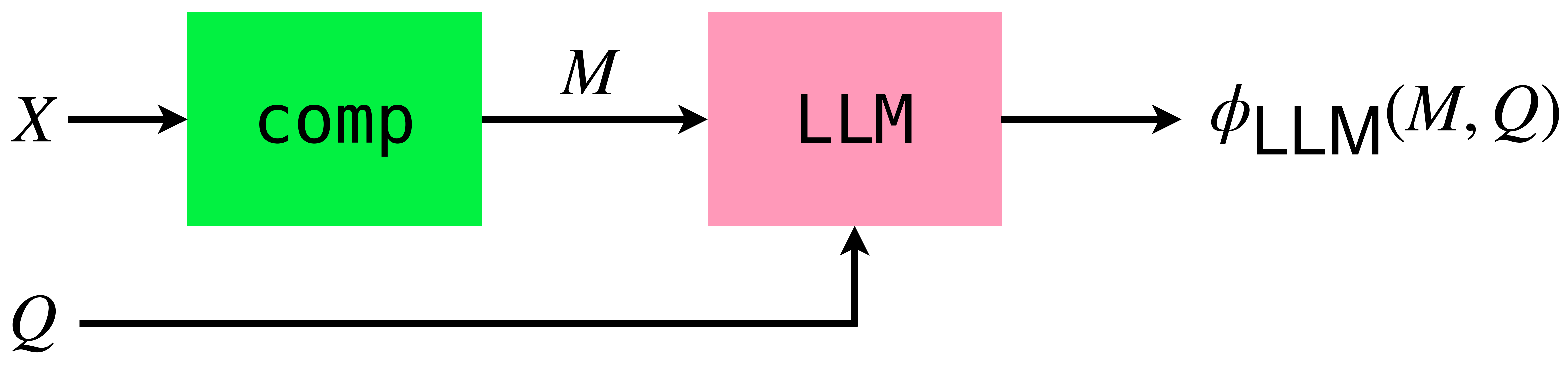
What is our rate-distortion framework?

How do we compute the optimal RD curves?

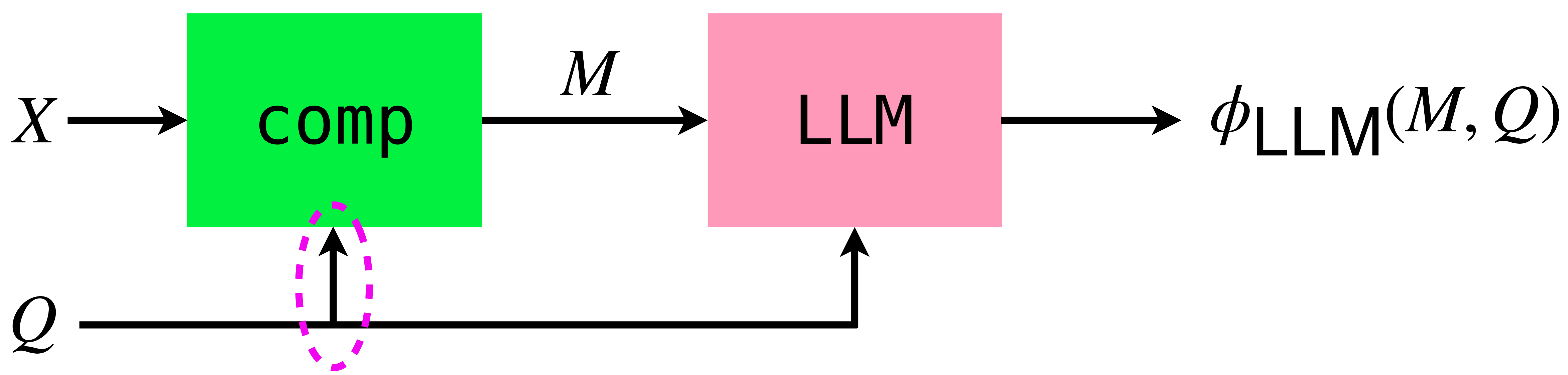
How efficiently can we find the RD curves?

Prompt compression: recap

Query-agnostic

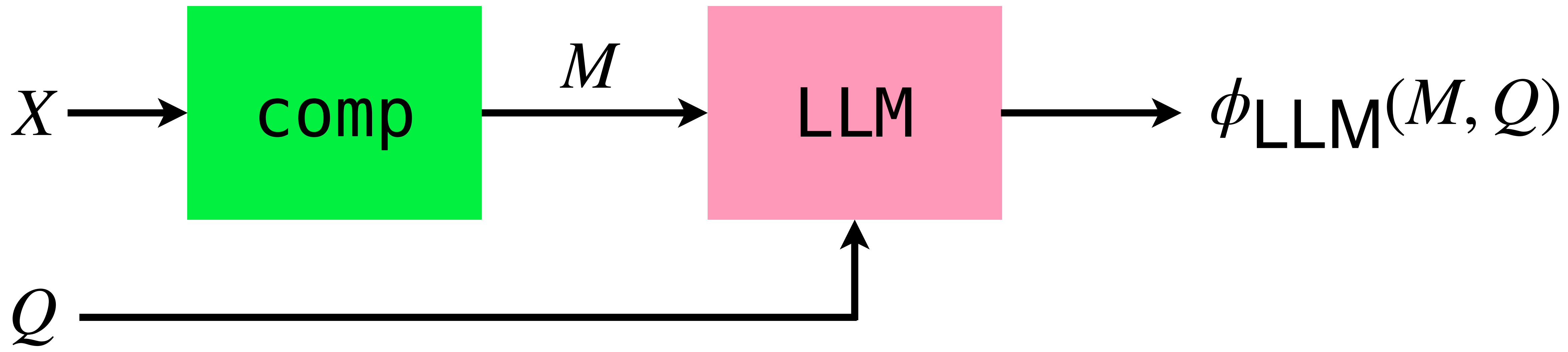


Query-aware

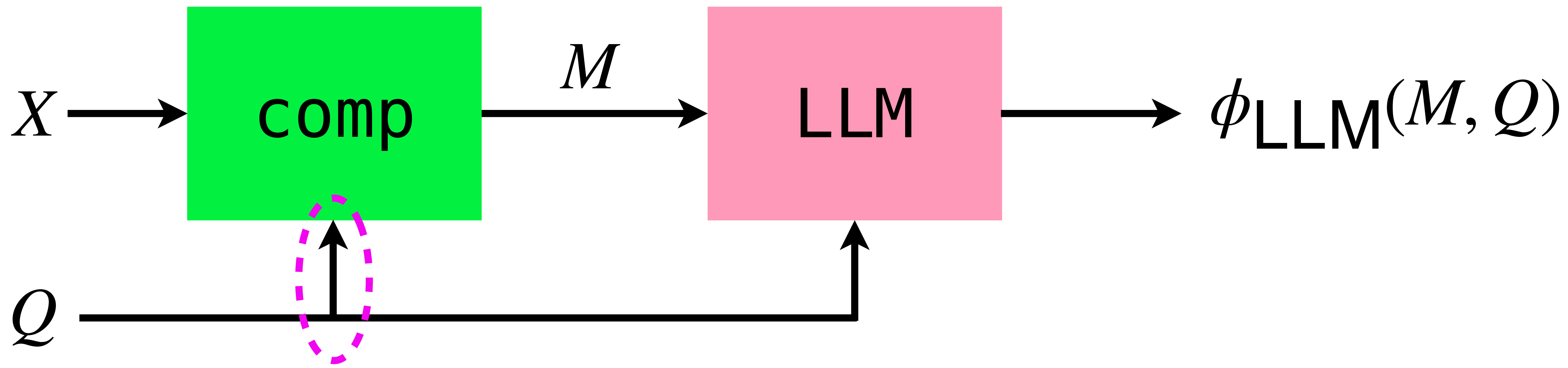


Prompt compression: rate-distortion framework

Query-agnostic



Query-aware

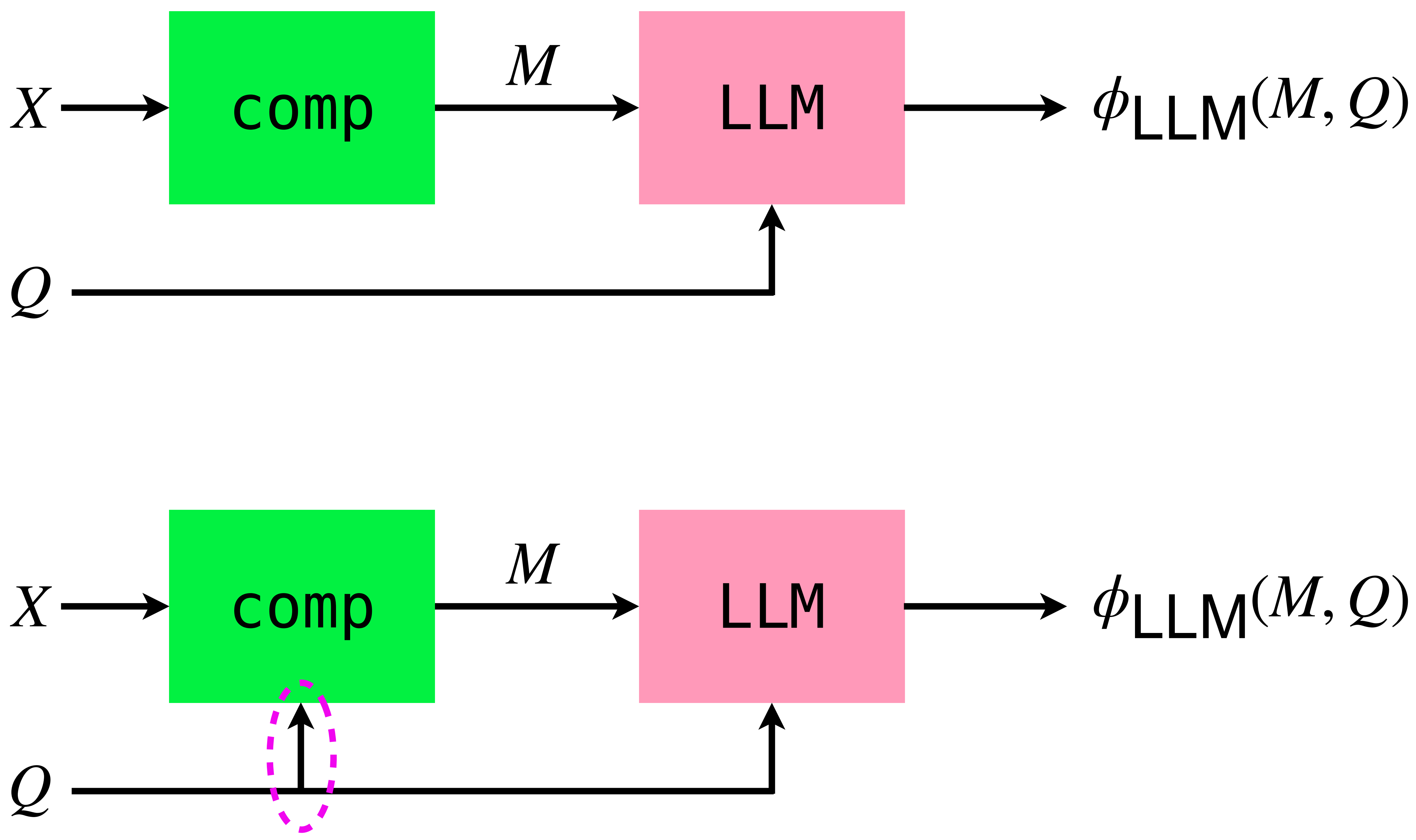


$(X, Q, Y) \sim \mathsf{P}_{XQY}$

Prompt compression: rate-distortion framework

Query-agnostic

Query-aware



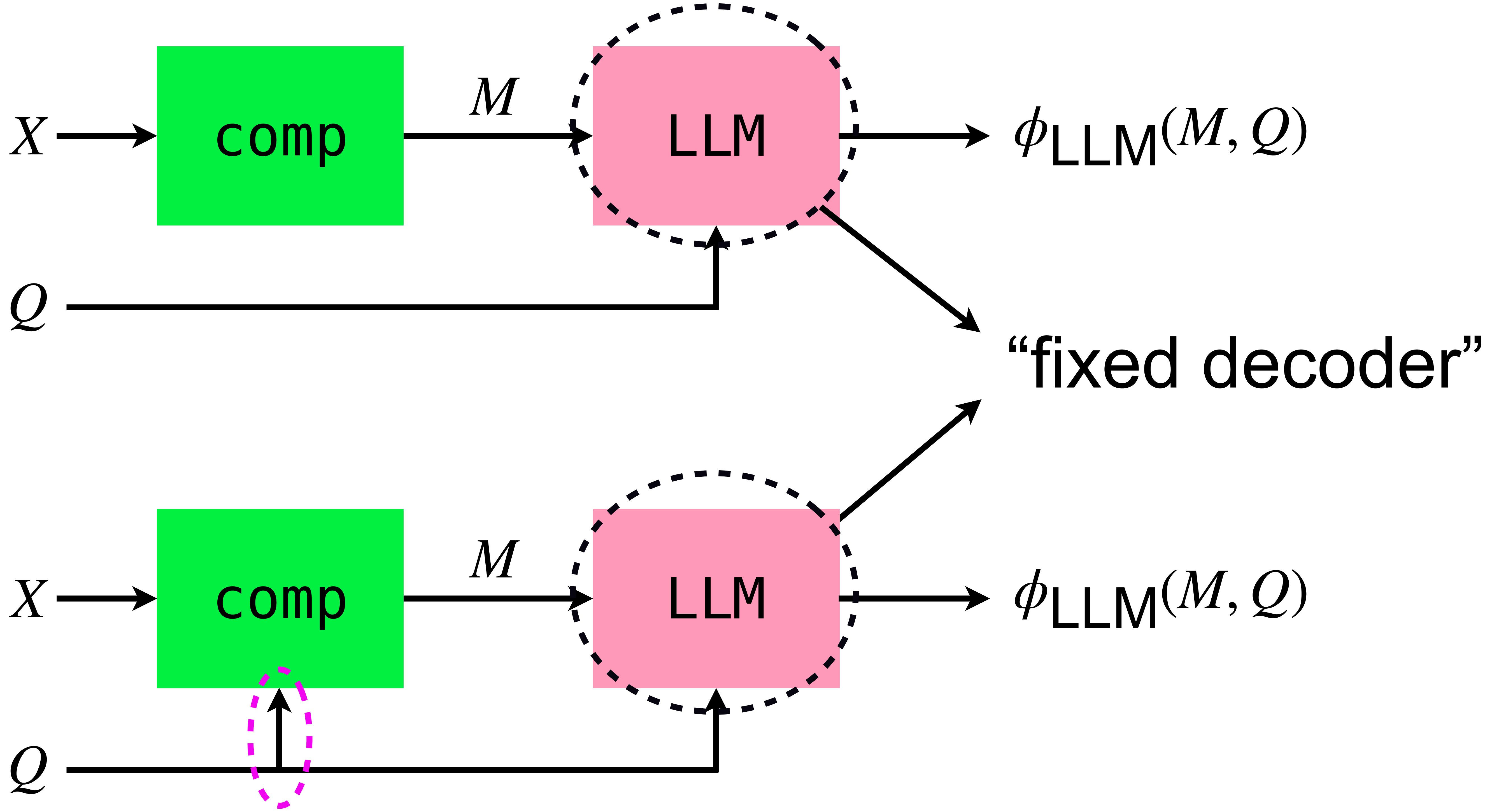
$$(X, Q, Y) \sim \mathsf{P}_{XQY}$$

answer

Prompt compression: rate-distortion framework

Query-agnostic

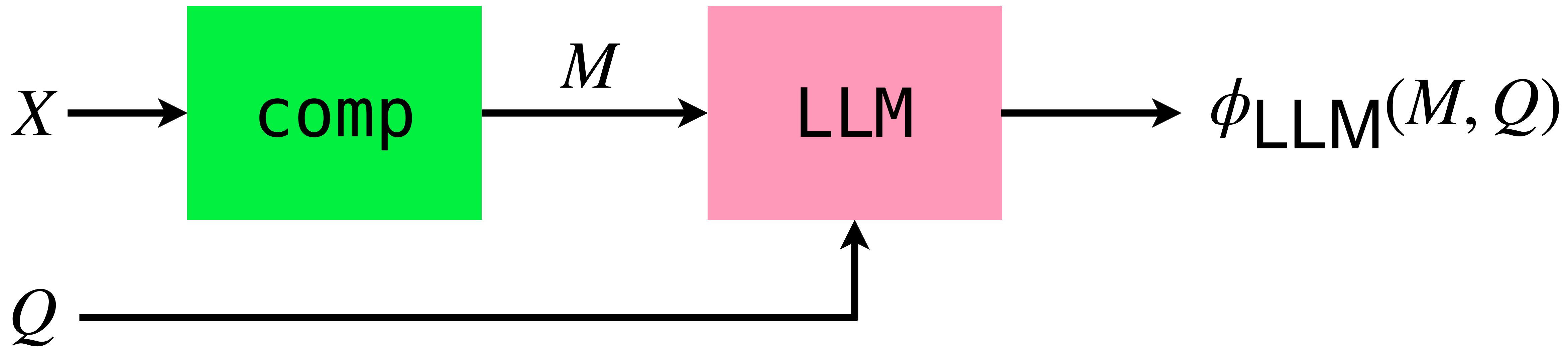
Query-aware



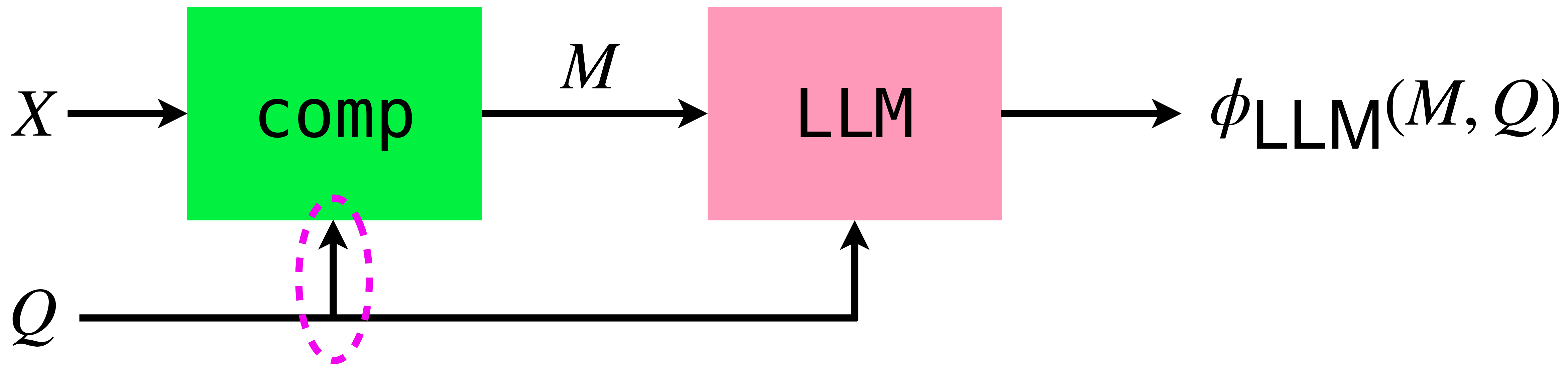
$(X, Q, Y) \sim \mathsf{P}_{XQY}$

Prompt compression: rate-distortion framework

Query-agnostic



Query-aware



$$(X, Q, Y) \sim \mathsf{P}_{XQY}$$

$$\text{Rate} = \mathbb{E}\left[\frac{\text{len}(M)}{\text{len}(X)}\right] \qquad \text{Distortion} = \mathbb{E}\left[\mathsf{d}\left(Y, \phi_{\mathsf{LLM}}(M, Q)\right)\right]$$

Distortion-rate function

$$\text{Rate} = \mathbb{E}\left[\frac{\operatorname{len}(M)}{\operatorname{len}(X)}\right] \qquad \text{Distortion} = \mathbb{E}\left[\mathsf{d}\left(Y, \phi_{\mathsf{LLM}}(M, Q)\right)\right]$$

Distortion-rate function

$$\text{Rate} = \mathbb{E}\left[\frac{\text{len}(M)}{\text{len}(X)}\right] \qquad \text{Distortion} = \mathbb{E}\left[\mathsf{d}\left(Y, \phi_{\mathsf{LLM}}(M, Q)\right)\right]$$

Optimal trade-off $= D^{*}(R) =$

Distortion-rate function

$$\text{Rate} = \mathbb{E}\left[\frac{\text{len}(M)}{\text{len}(X)}\right] \qquad \text{Distortion} = \mathbb{E}\left[\mathsf{d}\left(Y, \phi_{\mathsf{LLM}}(M, Q)\right)\right]$$

Optimal trade-off $= D^*(R) =$ smallest distortion

over all compressors

with rate $\leq R$

Distortion-rate function

$$\text{Rate} = \mathbb{E}\left[\frac{\mathrm{len}(M)}{\mathrm{len}(X)}\right] \qquad \text{Distortion} = \mathbb{E}\left[\mathsf{d}\left(Y, \phi_{\mathsf{LLM}}(M, Q)\right)\right]$$

Optimal trade-off $= D^*(R) =$ smallest distortion

over all compressors

with rate $\leq R$

Distortion-rate function

$$\text{Rate} = \mathbb{E}\left[\frac{\text{len}(M)}{\text{len}(X)}\right] \qquad \text{Distortion} = \mathbb{E}\left[\mathsf{d}\left(Y, \phi_{\mathsf{LLM}}(M, Q)\right)\right]$$

$$\text{Optimal trade-off} = D^*(R) = \inf_{\mathsf{P}_{M|X}} \quad \text{distortion}$$

$$\text{s.t.} \quad \text{rate} \leq R\text{, and}$$

$\mathsf{P}_{M|X}$ "is a compressor"

Distortion-rate function

$$\text{Rate} = \mathbb{E}\left[\frac{\text{len}(M)}{\text{len}(X)}\right] \qquad \text{Distortion} = \mathbb{E}\left[\mathsf{d}\left(Y, \phi_{\mathsf{LLM}}(M, Q)\right)\right]$$

$$\text{Optimal trade-off } = D^*(R) = \inf_{\mathsf{P}_{M|X}} \quad \text{distortion}$$

$$\text{s.t.} \quad \text{rate } \leq R, \text{ and}$$

$$\mathsf{P}_{M|X} \text{ “is a compressor”}$$

Distortion-rate function

$$\text{Rate} = \mathbb{E}\left[\frac{\mathrm{len}(M)}{\mathrm{len}(X)}\right] \qquad \text{Distortion} = \mathbb{E}\left[\mathsf{d}\left(Y, \phi_{\mathsf{LLM}}(M, Q)\right)\right]$$

$$\text{Optimal trade-off} = D^*(R) = \inf_{\mathsf{P}_{M|X}} \text{distortion}$$

$$\text{s.t.} \quad \text{rate} \leq R, \text{ and}$$

$$\mathsf{P}_{M|X} \text{ “is a compressor”}$$

Distortion-rate function

$$\text{Rate} = \mathbb{E}\left[\frac{\text{len}(M)}{\text{len}(X)}\right] \qquad \text{Distortion} = \mathbb{E}\left[\mathsf{d}\left(Y, \phi_{\mathsf{LLM}}(M, Q)\right)\right]$$

$$\text{Optimal trade-off } = D^*(R) = \inf_{\mathsf{P}_{M|X}} \quad \mathbb{E}\left[\mathsf{d}\left(Y, \phi_{\mathsf{LLM}}(M, Q)\right)\right]$$

s.t. rate $\leq R$, and

$\mathsf{P}_{M|X}$ “is a compressor”

Distortion-rate function

$$\text{Rate} = \mathbb{E}\left[\frac{\text{len}(M)}{\text{len}(X)}\right] \qquad \text{Distortion} = \mathbb{E}\left[\mathsf{d}\left(Y, \phi_{\mathsf{LLM}}(M, Q)\right)\right]$$

$$\text{Optimal trade-off } = D^*(R) = \inf_{\mathsf{P}_{M|X}} \quad \mathbb{E}\left[\mathsf{d}\left(Y, \phi_{\mathsf{LLM}}(M, Q)\right)\right]$$

s.t. rate $\leq R$, and

$\mathsf{P}_{M|X}$ "is a compressor"

Distortion-rate function

$$\text{Rate} = \mathbb{E}\left[\frac{\text{len}(M)}{\text{len}(X)}\right] \qquad \text{Distortion} = \mathbb{E}\left[\mathsf{d}\left(Y, \phi_{\mathsf{LLM}}(M, Q)\right)\right]$$

$$\text{Optimal trade-off} = D^*(R) = \inf_{\mathsf{P}_{M|X}} \mathbb{E}\left[\mathsf{d}\left(Y, \phi_{\mathsf{LLM}}(M, Q)\right)\right]$$

s.t. rate $\leq R$, and

$\mathsf{P}_{M|X}$ “is a compressor”

Distortion-rate function

$$\text{Rate} = \mathbb{E}\left[\frac{\text{len}(M)}{\text{len}(X)}\right] \qquad \text{Distortion} = \mathbb{E}\left[\mathsf{d}\left(Y, \phi_{\mathsf{LLM}}(M, Q)\right)\right]$$

$$\text{Optimal trade-off} = D^*(R) = \inf_{\mathsf{P}_{M|X}} \mathbb{E}\left[\mathsf{d}\left(Y, \phi_{\mathsf{LLM}}(M, Q)\right)\right]$$

$$\text{s.t.} \quad \mathbb{E}\left[\frac{\text{len}(M)}{\text{len}(X)}\right] \leq R, \text{ and}$$

$\mathsf{P}_{M|X}$ "is a compressor"

Distortion-rate function

$$\text{Rate} = \mathbb{E}\left[\frac{\text{len}(M)}{\text{len}(X)}\right] \qquad \text{Distortion} = \mathbb{E}\left[\mathsf{d}\left(Y, \phi_{\mathsf{LLM}}(M, Q)\right)\right]$$

$$\text{Optimal trade-off } = D^*(R) = \inf_{\mathsf{P}_{M|X}} \quad \mathbb{E}\left[\mathsf{d}\left(Y, \phi_{\mathsf{LLM}}(M, Q)\right)\right]$$

$$\text{s.t.} \quad \mathbb{E}\left[\frac{\text{len}(M)}{\text{len}(X)}\right] \leq R, \text{ and}$$

$$\mathsf{P}_{M|X} \text{ “is a compressor”}$$

Distortion-rate function: linear program

$$\text{Rate} = \mathbb{E}\left[\frac{\text{len}(M)}{\text{len}(X)}\right] \qquad \text{Distortion} = \mathbb{E}\left[\mathsf{d}\left(Y, \phi_{\mathsf{LLM}}(M, Q)\right)\right]$$

$$\text{Optimal trade-off } = D^*(R) = \inf_{\mathsf{P}_{M|X}} \mathbb{E}\left[\mathsf{d}\left(Y, \phi_{\mathsf{LLM}}(M, Q)\right)\right]$$

linear program ✅

$$\text{s.t.} \quad \mathbb{E}\left[\frac{\text{len}(M)}{\text{len}(X)}\right] \le R, \text{ and}$$

$\mathsf{P}_{M|X}$ “is a compressor”

Distortion-rate function: linear program

$$\text{Rate} = \mathbb{E}\left[\frac{\mathrm{len}(M)}{\mathrm{len}(X)}\right] \qquad \text{Distortion} = \mathbb{E}\left[\mathsf{d}\left(Y, \phi_{\mathsf{LLM}}(M, Q)\right)\right]$$

$$\text{Optimal trade-off } = D^*(R) = \inf_{\mathsf{P}_{M|X}} \mathbb{E}\left[\mathsf{d}\left(Y, \phi_{\mathsf{LLM}}(M, Q)\right)\right]$$

$$\text{s.t.} \quad \mathbb{E}\left[\frac{\mathrm{len}(M)}{\mathrm{len}(X)}\right] \leq R, \text{ and}$$

$\mathsf{P}_{M|X}$ "is a compressor"

linear program ✅

Problem solved!

Distortion-rate function: linear program

$$\text{Rate} = \mathbb{E}\left[\frac{\text{len}(M)}{\text{len}(X)}\right] \qquad \text{Distortion} = \mathbb{E}\left[\mathsf{d}\left(Y, \phi_{\mathsf{LLM}}(M, Q)\right)\right]$$

$$\text{Optimal trade-off } = D^*(R) = \inf_{\mathsf{P}_{M|X}} \quad \mathbb{E}\left[\mathsf{d}\left(Y, \phi_{\mathsf{LLM}}(M, Q)\right)\right]$$

$$\text{s.t.} \quad \mathbb{E}\left[\frac{\text{len}(M)}{\text{len}(X)}\right] \le R, \text{ and}$$

$\mathsf{P}_{M|X}$ “is a compressor”

linear program ✅

Problem solved?

impossible

Distortion-rate function: linear program

$$\text{Rate} = \mathbb{E}\left[\frac{\mathrm{len}(M)}{\mathrm{len}(X)}\right] \qquad \text{Distortion} = \mathbb{E}\left[\mathsf{d}\left(Y, \phi_{\mathsf{LLM}}(M, Q)\right)\right]$$

$$\text{Optimal trade-off} = D^*(R) = \inf_{\mathsf{P}_{M|X}} \mathbb{E}\left[\mathsf{d}\left(Y, \phi_{\mathsf{LLM}}(M, Q)\right)\right]$$

$$\text{s.t.} \quad \mathbb{E}\left[\frac{\mathrm{len}(M)}{\mathrm{len}(X)}\right] \leq R, \text{ and}$$

$$\mathsf{P}_{M|X} \text{ “is a compressor”}$$

linear program ✅

large dimension ❌

impossible

Distortion-rate function: linear program

$$\text{Rate} = \mathbb{E}\left[\frac{\text{len}(M)}{\text{len}(X)}\right] \qquad \text{Distortion} = \mathbb{E}\left[\mathsf{d}\left(Y, \phi_{\mathsf{LLM}}(M, Q)\right)\right]$$

$$\text{Optimal trade-off} = D^*(R) = \inf_{\mathsf{P}_{M|X}} \mathbb{E}\left[\mathsf{d}\left(Y, \phi_{\mathsf{LLM}}(M, Q)\right)\right]$$

$$\text{s.t.} \quad \mathbb{E}\left[\frac{\text{len}(M)}{\text{len}(X)}\right] \leq R, \text{ and}$$

$\mathsf{P}_{M|X}$ "is a compressor"

linear program ✔

large dimension ✘

$\approx 32{,}000^{100}$

Distortion-rate function: $^\wedge$impossible? linear program

$$\text{Rate} = \mathbb{E}\left[\frac{\text{len}(M)}{\text{len}(X)}\right] \qquad \text{Distortion} = \mathbb{E}\left[\mathsf{d}\left(Y, \phi_{\mathsf{LLM}}(M, Q)\right)\right]$$

$$\text{Optimal trade-off } = D^*(R) = \inf_{\mathsf{P}_{M|X}} \mathbb{E}\left[\mathsf{d}\left(Y, \phi_{\mathsf{LLM}}(M, Q)\right)\right]$$

$$\text{s.t.} \quad \mathbb{E}\left[\frac{\text{len}(M)}{\text{len}(X)}\right] \leq R, \text{ and}$$

$\mathsf{P}_{M|X}$ "is a compressor"

linear program ✅

large dimension ❌

$\approx 32{,}000^{100}$

Can we solve it?

Distortion-rate function: $\wedge$ linear program
impossible?

$$\text{Rate} = \mathbb{E}\left[\frac{\mathrm{len}(M)}{\mathrm{len}(X)}\right] \qquad \text{Distortion} = \mathbb{E}\left[\mathsf{d}\left(Y, \phi_{\mathsf{LLM}}(M, Q)\right)\right]$$

$$\text{Optimal trade-off } = D^*(R) = \inf_{\mathsf{P}_{M|X}} \mathbb{E}\left[\mathsf{d}\left(Y, \phi_{\mathsf{LLM}}(M, Q)\right)\right]$$

$$\text{s.t.} \quad \mathbb{E}\left[\frac{\mathrm{len}(M)}{\mathrm{len}(X)}\right] \le R, \text{ and}$$

$\mathsf{P}_{M|X}$ "is a compressor"

linear program ✅

large dimension ❌

$\approx 32{,}000^{100}$

Can we solve it?

⟶ Yes, via dual!

Distortion-rate function: dual linear program

$$\text{Rate} = \mathbb{E}\left[\frac{\text{len}(M)}{\text{len}(X)}\right] \qquad \text{Distortion} = \mathbb{E}\left[\mathrm{d}\left(Y, \phi_{\text{LLM}}(M, Q)\right)\right]$$

$$\text{Optimal trade-off } = D^*(R) = \inf_{\mathrm{P}_{M|X}} \mathbb{E}\left[\mathrm{d}\left(Y, \phi_{\text{LLM}}(M, Q)\right)\right]$$

$$\text{s.t.} \quad \mathbb{E}\left[\frac{\text{len}(M)}{\text{len}(X)}\right] \leq R, \text{ and}$$

$\mathrm{P}_{M|X}$ "is a compressor"

linear program ✓

large dimension ✗

$\approx 32{,}000^{100}$

Can we solve it?

⟶ Yes, via dual!

$$\text{“} = \sup_{\lambda \geq 0} \left\{ -\lambda R + \sum_{x \in \mathscr{X}} \min_{m \in \mathscr{M}_x} \left[\boldsymbol{D}_{x,m} + \lambda \boldsymbol{R}_{x,m}\right] \right\} \text{”}$$

Distortion-rate function: dual linear program

$$\text{Rate} = \mathbb{E}\left[\frac{\text{len}(M)}{\text{len}(X)}\right] \qquad \text{Distortion} = \mathbb{E}\left[\mathsf{d}\left(Y, \phi_{\mathsf{LLM}}(M, Q)\right)\right]$$

$$\text{Optimal trade-off } = D^*(R) = \inf_{\mathsf{P}_{M|X}} \mathbb{E}\left[\mathsf{d}\left(Y, \phi_{\mathsf{LLM}}(M, Q)\right)\right]$$

$$\text{s.t.} \quad \mathbb{E}\left[\frac{\text{len}(M)}{\text{len}(X)}\right] \leq R, \text{ and}$$

$\mathsf{P}_{M|X}$ “is a compressor”

linear program

large dimension

$\approx 32{,}000^{100}$

Can we solve it?

→ Yes, via dual!

$$\text{“} = \sup_{\lambda \geq 0} \left\{ -\lambda R + \sum_{x \in \mathcal{X}} \min_{m \in \boxed{\mathcal{M}_x}} \left[\boxed{\boldsymbol{D}_{x,m}} + \lambda \boxed{\boldsymbol{R}_{x,m}}\right] \right\} \text{”}$$

Dual linear program

$$D^*(R) = \sup_{\lambda \geq 0} \left\{ -\lambda R + \sum_{x \in \mathcal{X}} \min_{m \in \mathcal{M}_x} \left[\boldsymbol{D}_{x,m} + \lambda \boldsymbol{R}_{x,m} \right] \right\}$$

Dual linear program

$$D^*(R) = \boxed{\sup_{\lambda \geq 0}} \left\{ -\lambda R + \boxed{\sum_{x \in \mathscr{X}}} \boxed{\min_{m \in \mathscr{M}_x} \left[\boldsymbol{D}_{x,m} + \lambda \boldsymbol{R}_{x,m} \right]} \right\}$$

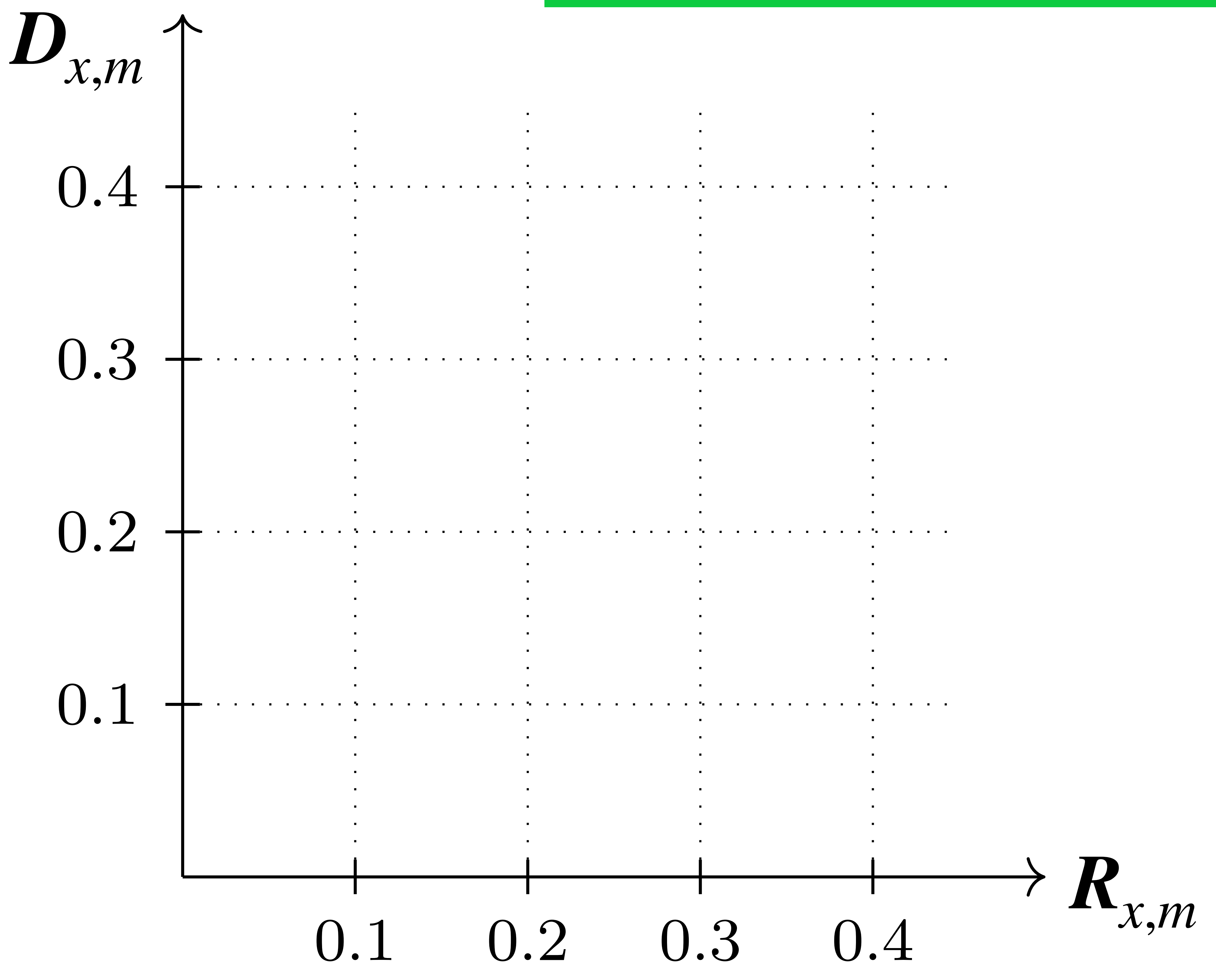
Dual linear program

$$D^*(R) = \sup_{\lambda \geq 0} \left\{ -\lambda R + \sum_{x \in \mathscr{X}} \boxed{\min_{m \in \mathscr{M}_x} \left[\boldsymbol{D}_{x,m} + \lambda \boldsymbol{R}_{x,m} \right]} \right\}$$

Dual linear program: geometric solution?

$$D^*(R) = \sup_{\lambda \geq 0} \left\{ -\lambda R + \sum_{x \in \mathscr{X}} \boxed{\min_{m \in \mathscr{M}_x} \left[\boldsymbol{D}_{x,m} + \lambda \boldsymbol{R}_{x,m} \right]} \right\}$$

— Fix $x \in \mathscr{X}$

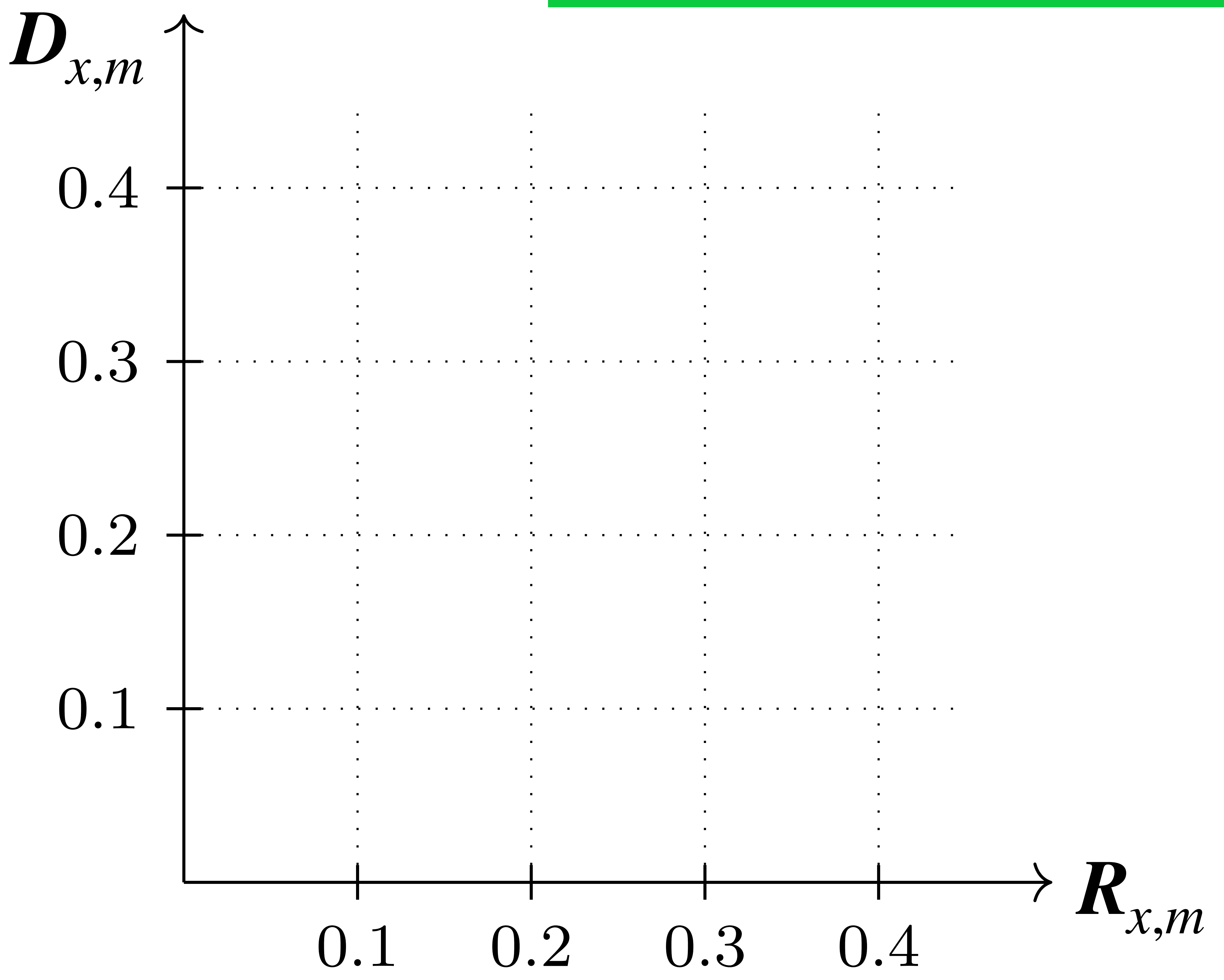


Dual linear program: geometric solution?

$$D^*(R) = \sup_{\lambda \geq 0} \left\{ -\lambda R + \sum_{x \in \mathscr{X}} \boxed{\min_{m \in \mathscr{M}_x} \left[\boldsymbol{D}_{x,m} + \lambda \boldsymbol{R}_{x,m} \right]} \right\}$$

— Fix $x \in \mathscr{X}$

— Plot $\left\{ \left(\boldsymbol{R}_{x,m}, \boldsymbol{D}_{x,m} \right) \right\}_{m \in \mathscr{M}_x}$



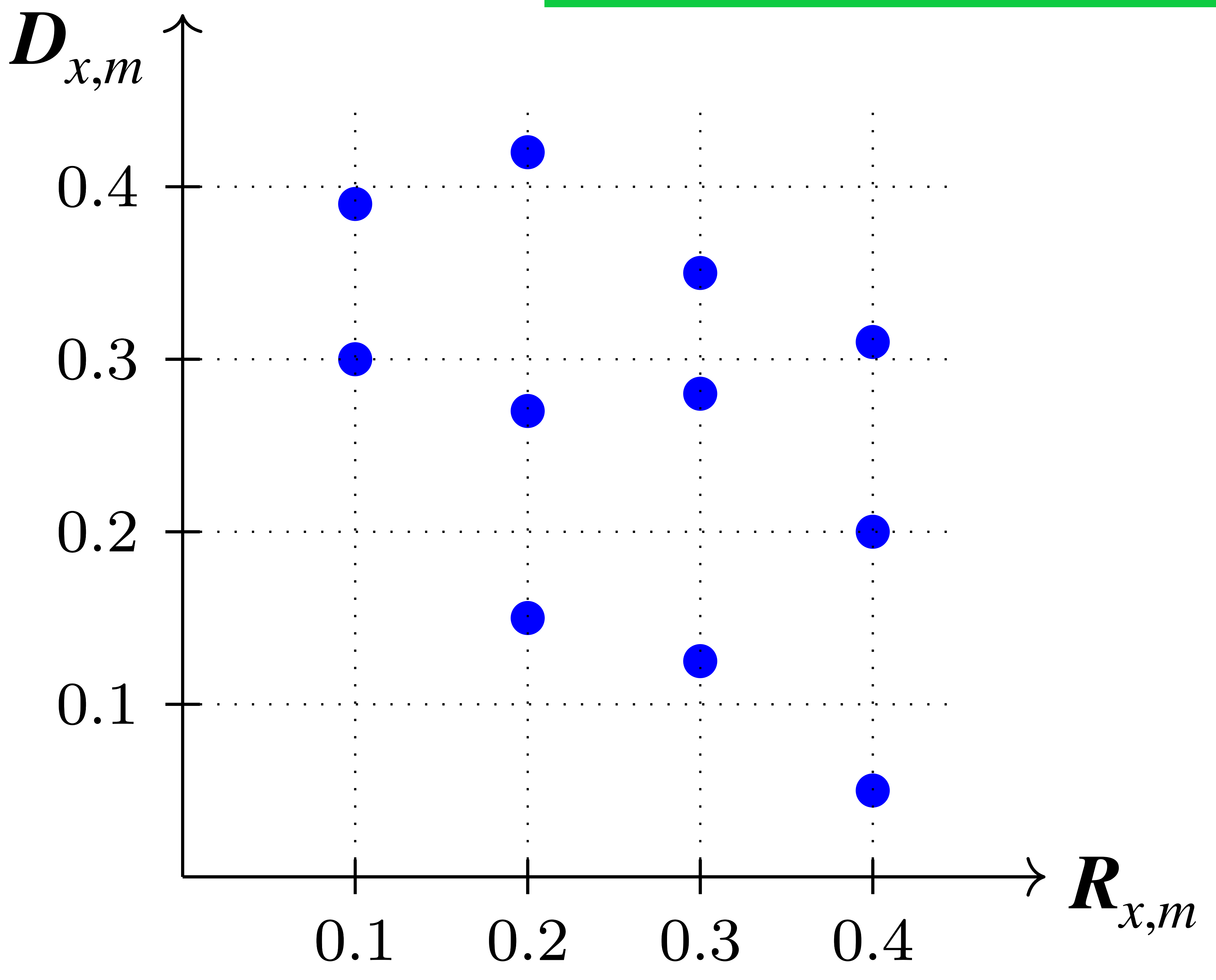
Dual linear program: geometric solution?

$$D^*(R) = \sup_{\lambda \geq 0} \left\{ -\lambda R + \sum_{x \in \mathcal{X}} \boxed{\min_{m \in \mathcal{M}_x} \left[D_{x,m} + \lambda R_{x,m} \right]} \right\}$$

— Fix $x \in \mathcal{X}$

$|\mathcal{M}_x| = 11$ here

— Plot $\left\{ \left(R_{x,m}, D_{x,m} \right) \right\}_{m \in \mathcal{M}_x}$



Dual linear program: geometric solution?

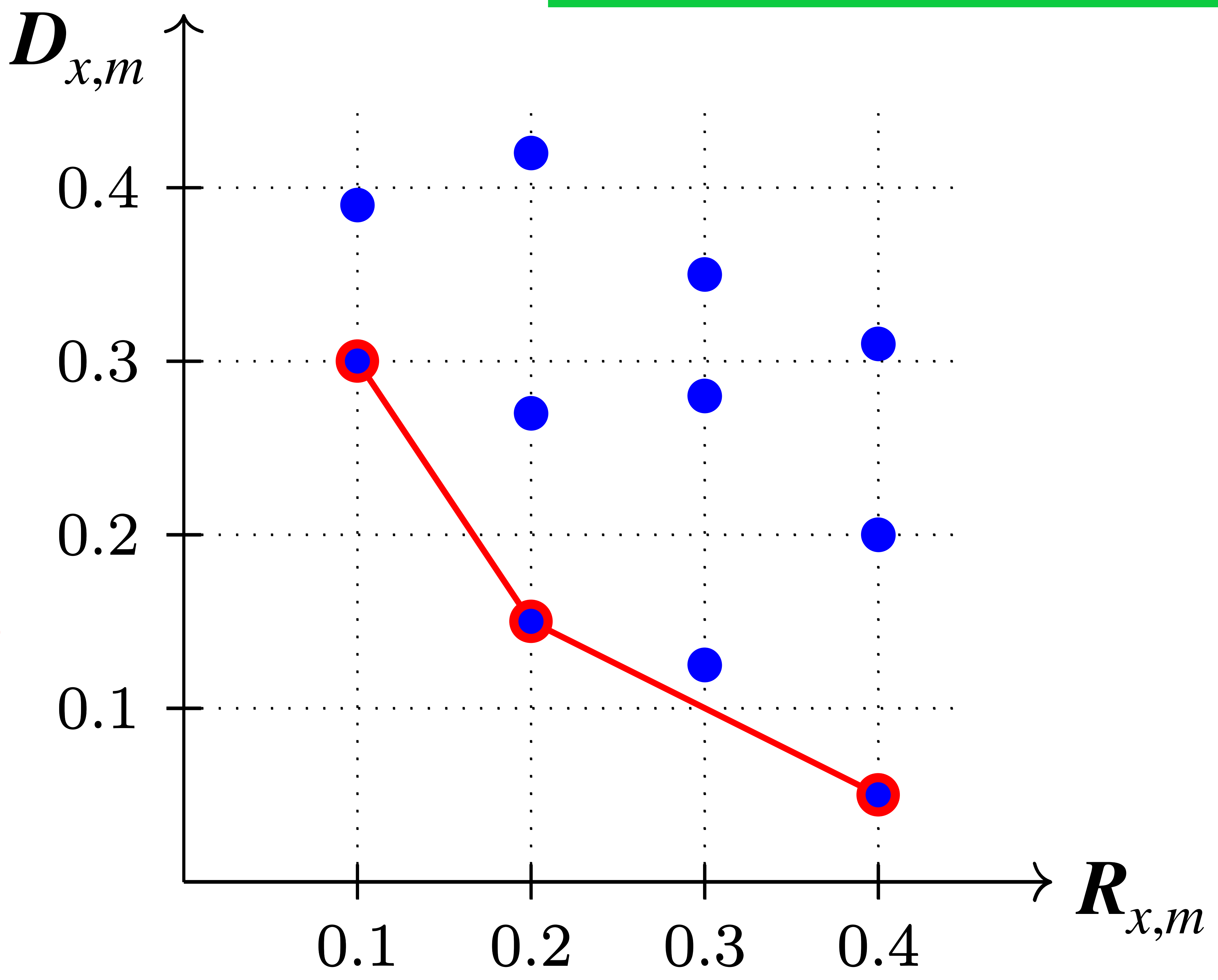
$$D^*(R) = \sup_{\lambda \geq 0} \left\{ -\lambda R + \sum_{x \in \mathcal{X}} \boxed{\min_{m \in \mathcal{M}_x} \left[\boldsymbol{D}_{x,m} + \lambda \boldsymbol{R}_{x,m} \right]} \right\}$$

— Fix $x \in \mathcal{X}$

$|\mathcal{M}_x| = 11$ here

— Plot $\left\{ (\boldsymbol{R}_{x,m}, \boldsymbol{D}_{x,m}) \right\}_{m \in \mathcal{M}_x}$

— lower-left concave envelope



Dual linear program: geometric solution?

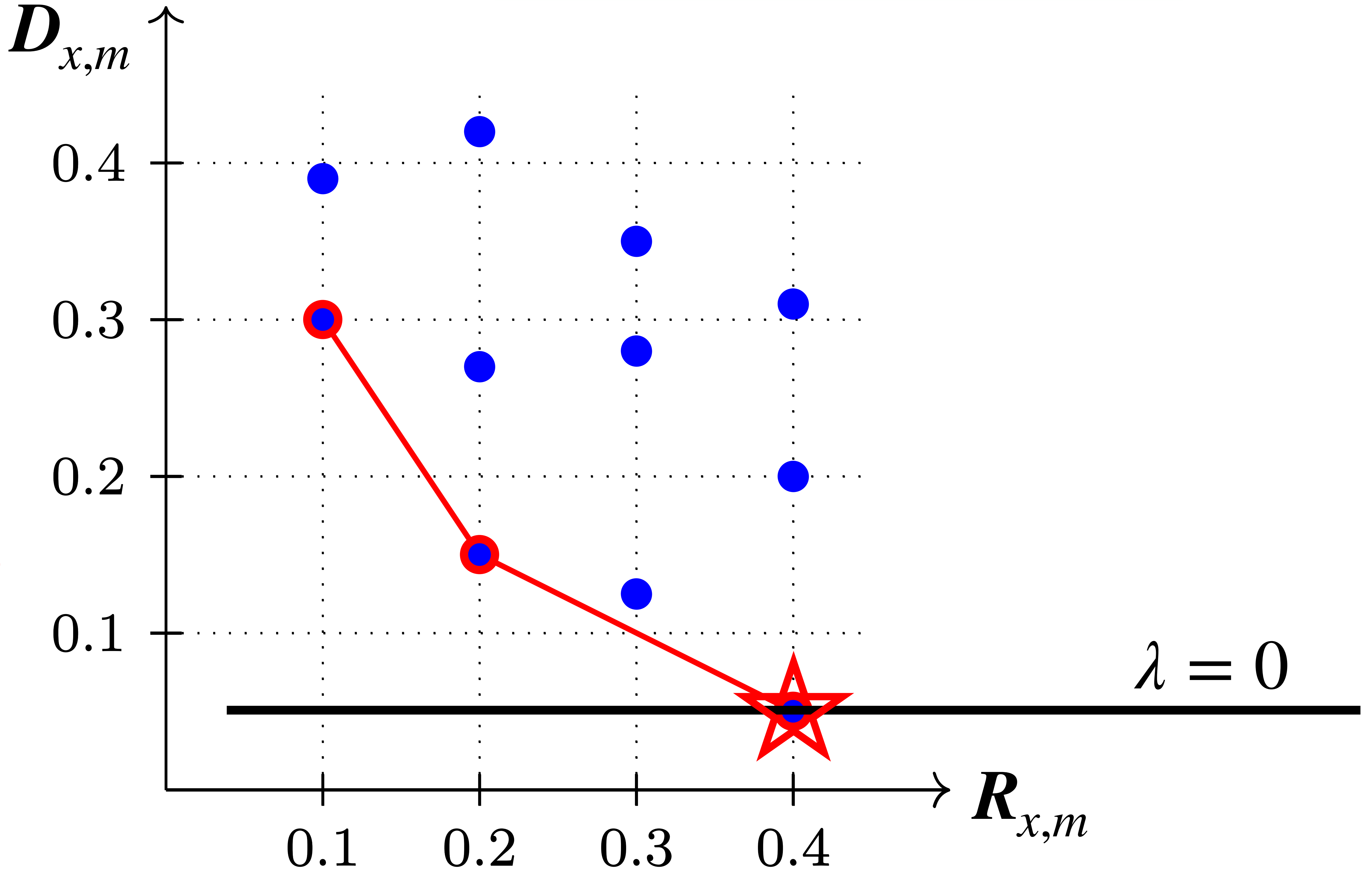
$$D^*(R) = \sup_{\lambda \geq 0} \left\{ -\lambda R + \sum_{x \in \mathcal{X}} \boxed{\min_{m \in \mathcal{M}_x} \left[\boldsymbol{D}_{x,m} + \lambda \boldsymbol{R}_{x,m} \right]} \right\}$$

— Fix $x \in \mathcal{X}$

$|\mathcal{M}_x| = 11$ here

— Plot $\left\{ \left(\boldsymbol{R}_{x,m}, \boldsymbol{D}_{x,m} \right) \right\}_{m \in \mathcal{M}_x}$

— lower-left concave envelope



Dual linear program: geometric solution?

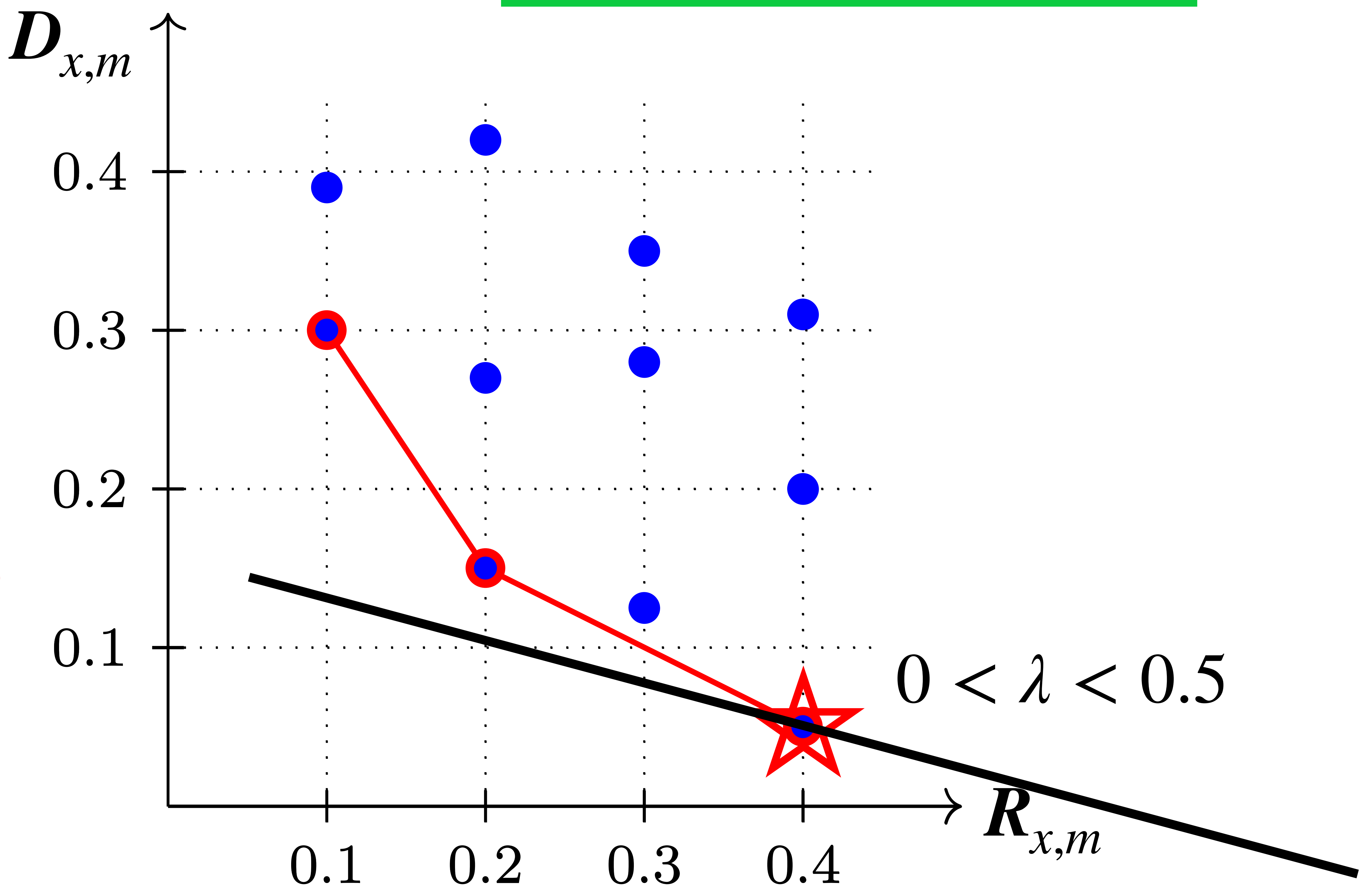
$$D^*(R) = \sup_{\lambda \geq 0} \left\{ -\lambda R + \sum_{x \in \mathcal{X}} \boxed{\min_{m \in \mathcal{M}_x} \left[\boldsymbol{D}_{x,m} + \lambda \boldsymbol{R}_{x,m} \right]} \right\}$$

— Fix $x \in \mathcal{X}$

$|\mathcal{M}_x| = 11$ here

— Plot $\left\{ \left(\boldsymbol{R}_{x,m}, \boldsymbol{D}_{x,m} \right) \right\}_{m \in \mathcal{M}_x}$

— lower-left concave envelope



Dual linear program: geometric solution?

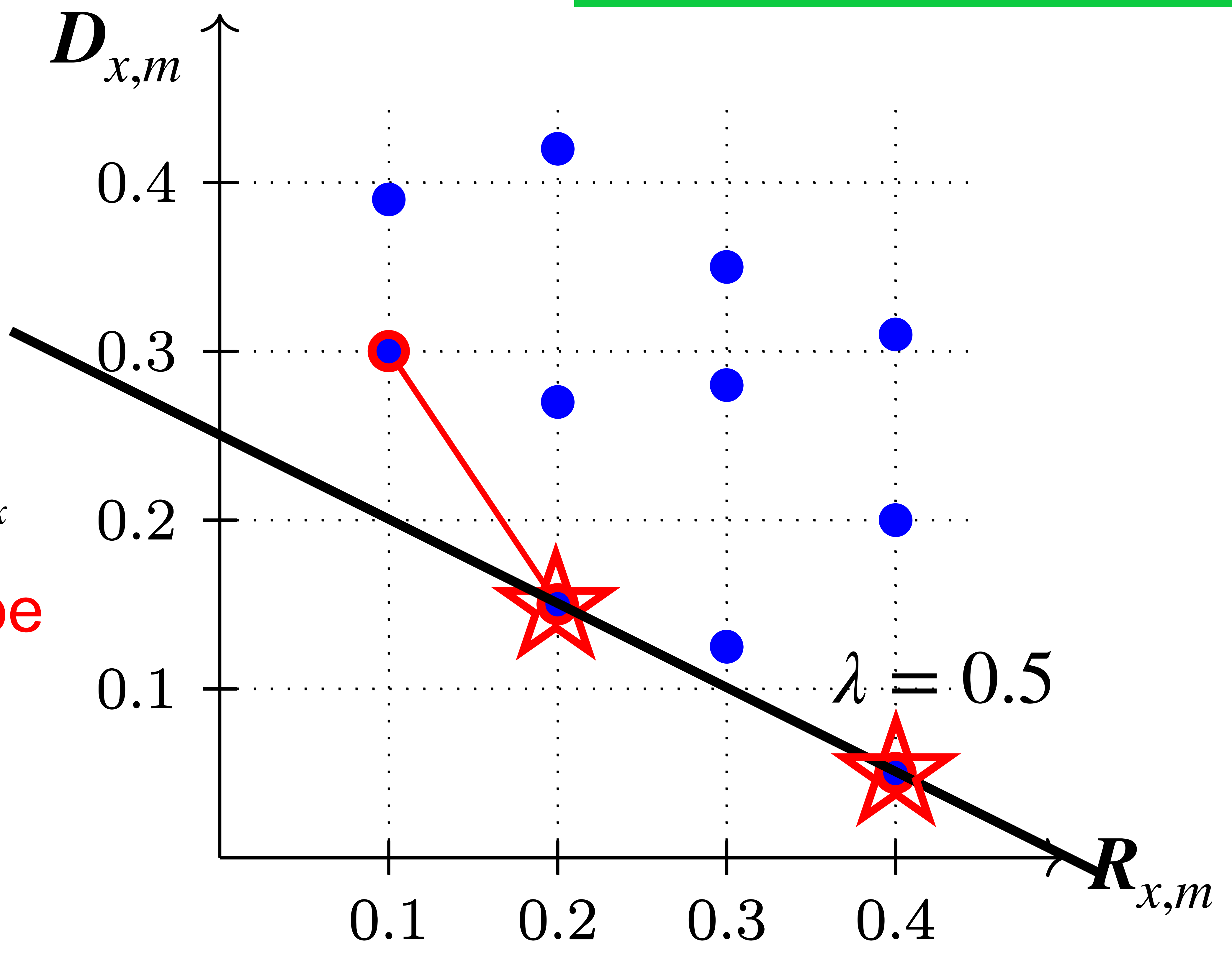
$$D^*(R) = \sup_{\lambda \geq 0} \left\{ -\lambda R + \sum_{x \in \mathcal{X}} \boxed{\min_{m \in \mathcal{M}_x} \left[\boldsymbol{D}_{x,m} + \lambda \boldsymbol{R}_{x,m} \right]} \right\}$$

— Fix $x \in \mathcal{X}$

$|\mathcal{M}_x| = 11$ here

— Plot $\left\{ \left(\boldsymbol{R}_{x,m}, \boldsymbol{D}_{x,m} \right) \right\}_{m \in \mathcal{M}_x}$

— lower-left concave envelope



Dual linear program: geometric solution?

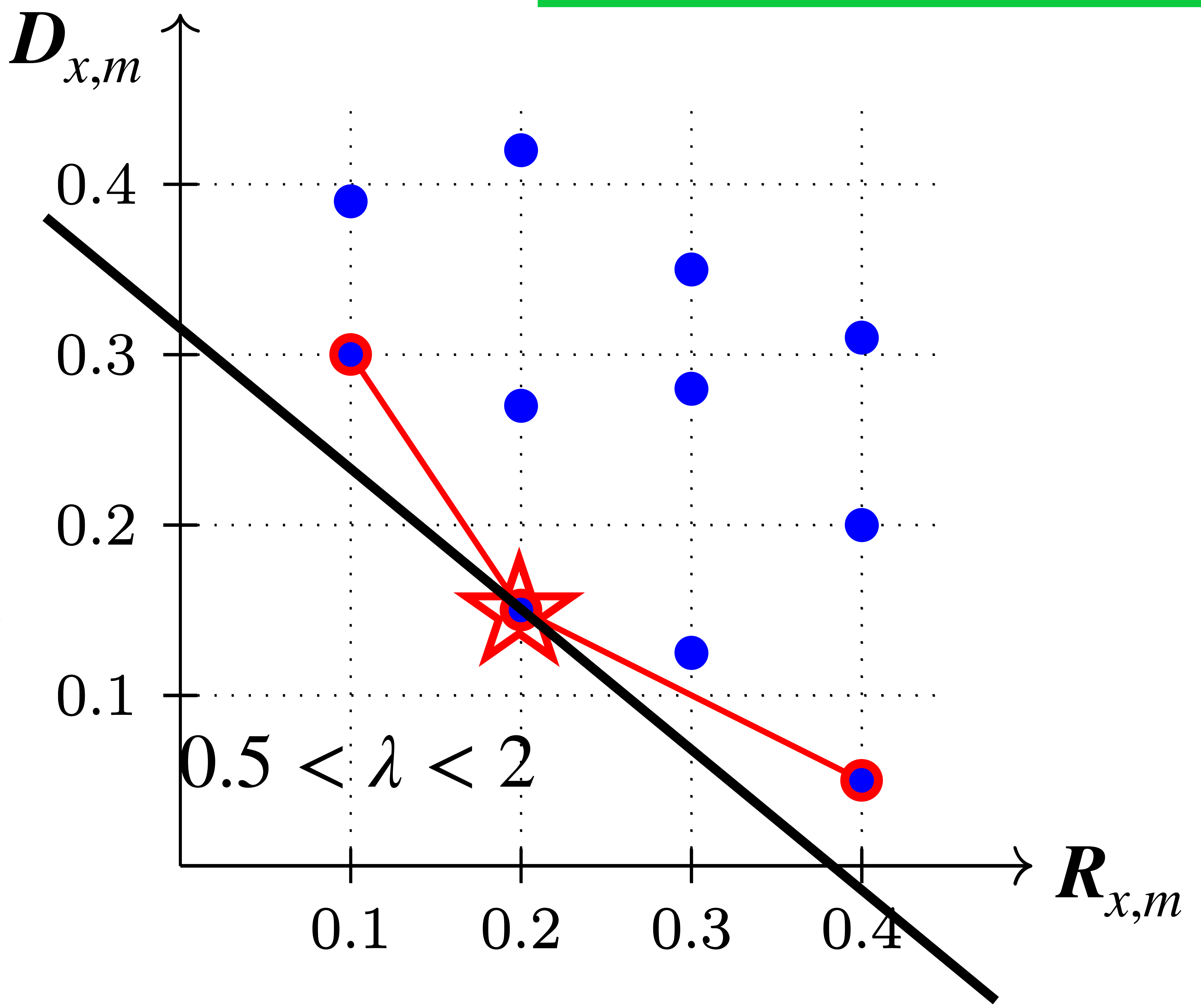
$$D^*(R) = \sup_{\lambda \geq 0} \left\{ -\lambda R + \sum_{x \in \mathcal{X}} \boxed{\min_{m \in \mathcal{M}_x} \left[\boldsymbol{D}_{x,m} + \lambda \boldsymbol{R}_{x,m} \right]} \right\}$$

— Fix $x \in \mathcal{X}$

$|\mathcal{M}_x| = 11$ here

— Plot $\left\{ \left(\boldsymbol{R}_{x,m}, \boldsymbol{D}_{x,m} \right) \right\}_{m \in \mathcal{M}_x}$

— lower-left concave envelope



Dual linear program: geometric solution?

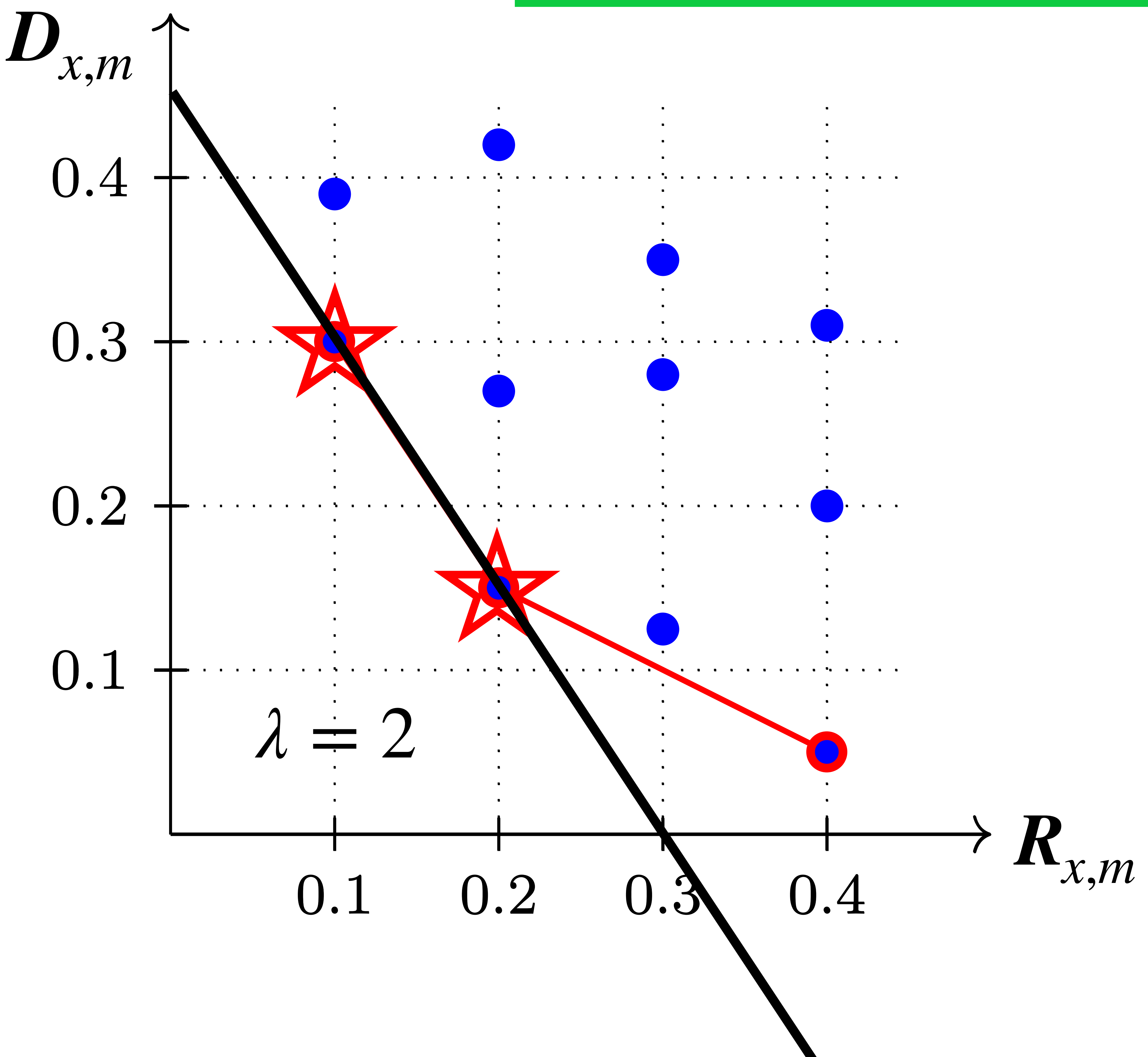
$$D^*(R) = \sup_{\lambda \geq 0} \left\{ -\lambda R + \sum_{x \in \mathscr{X}} \boxed{\min_{m \in \mathscr{M}_x} \left[\boldsymbol{D}_{x,m} + \lambda \boldsymbol{R}_{x,m} \right]} \right\}$$

— Fix $x \in \mathscr{X}$

$|\mathscr{M}_x| = 11$ here

— Plot $\left\{ \left(\boldsymbol{R}_{x,m}, \boldsymbol{D}_{x,m} \right) \right\}_{m \in \mathscr{M}_x}$

— lower-left concave envelope



Dual linear program: geometric solution?

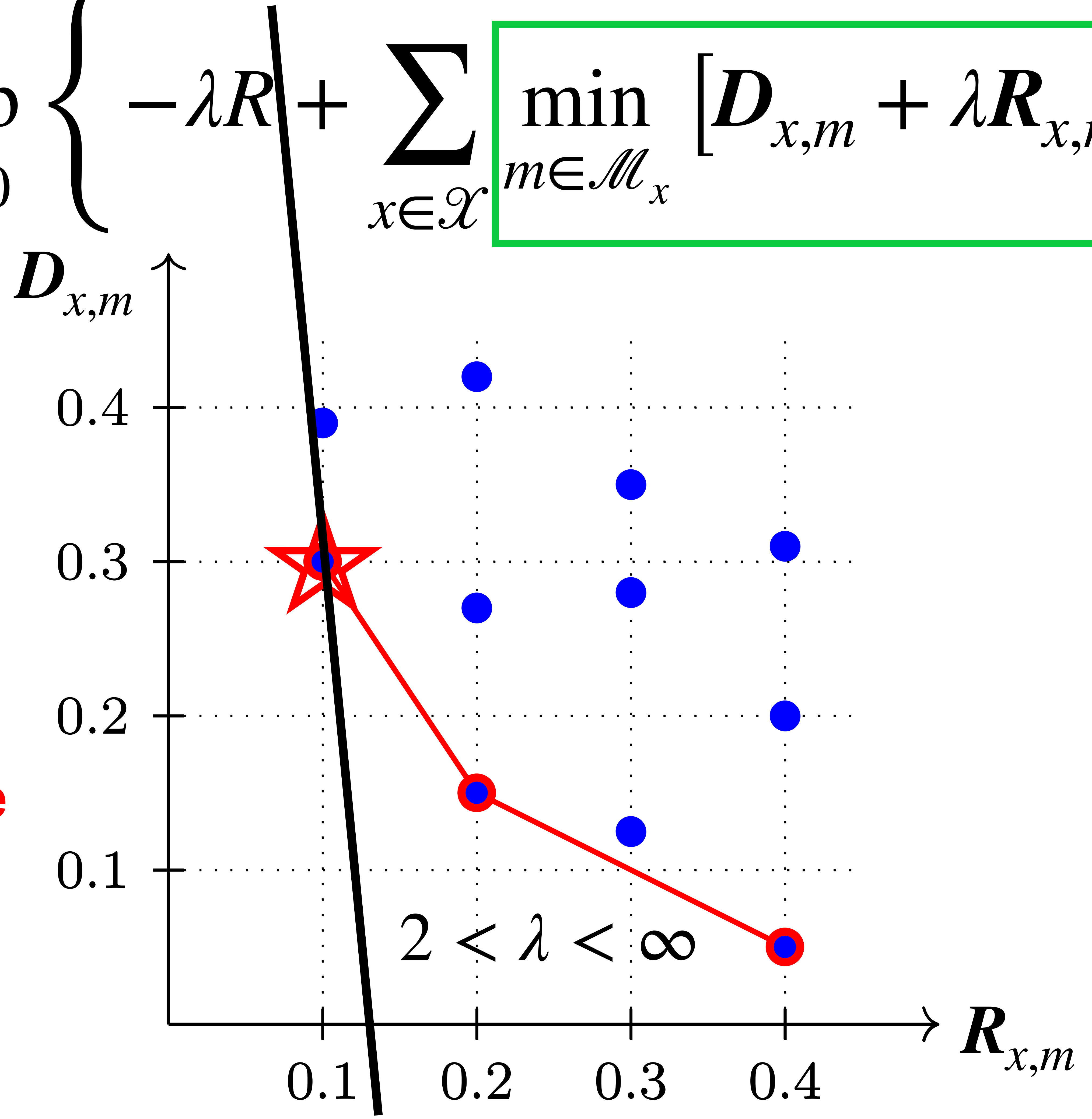
$$D^*(R) = \sup_{\lambda \geq 0} \left\{ -\lambda R + \sum_{x \in \mathcal{X}} \boxed{\min_{m \in \mathcal{M}_x} \left[\boldsymbol{D}_{x,m} + \lambda \boldsymbol{R}_{x,m} \right]} \right\}$$

— Fix $x \in \mathcal{X}$

$|\mathcal{M}_x| = 11$ here

— Plot $\left\{ \left(\boldsymbol{R}_{x,m}, \boldsymbol{D}_{x,m} \right) \right\}_{m \in \mathcal{M}_x}$

— lower-left concave envelope



Dual linear program: $\overset{\text{efficient}}{\wedge}$ geometric solution

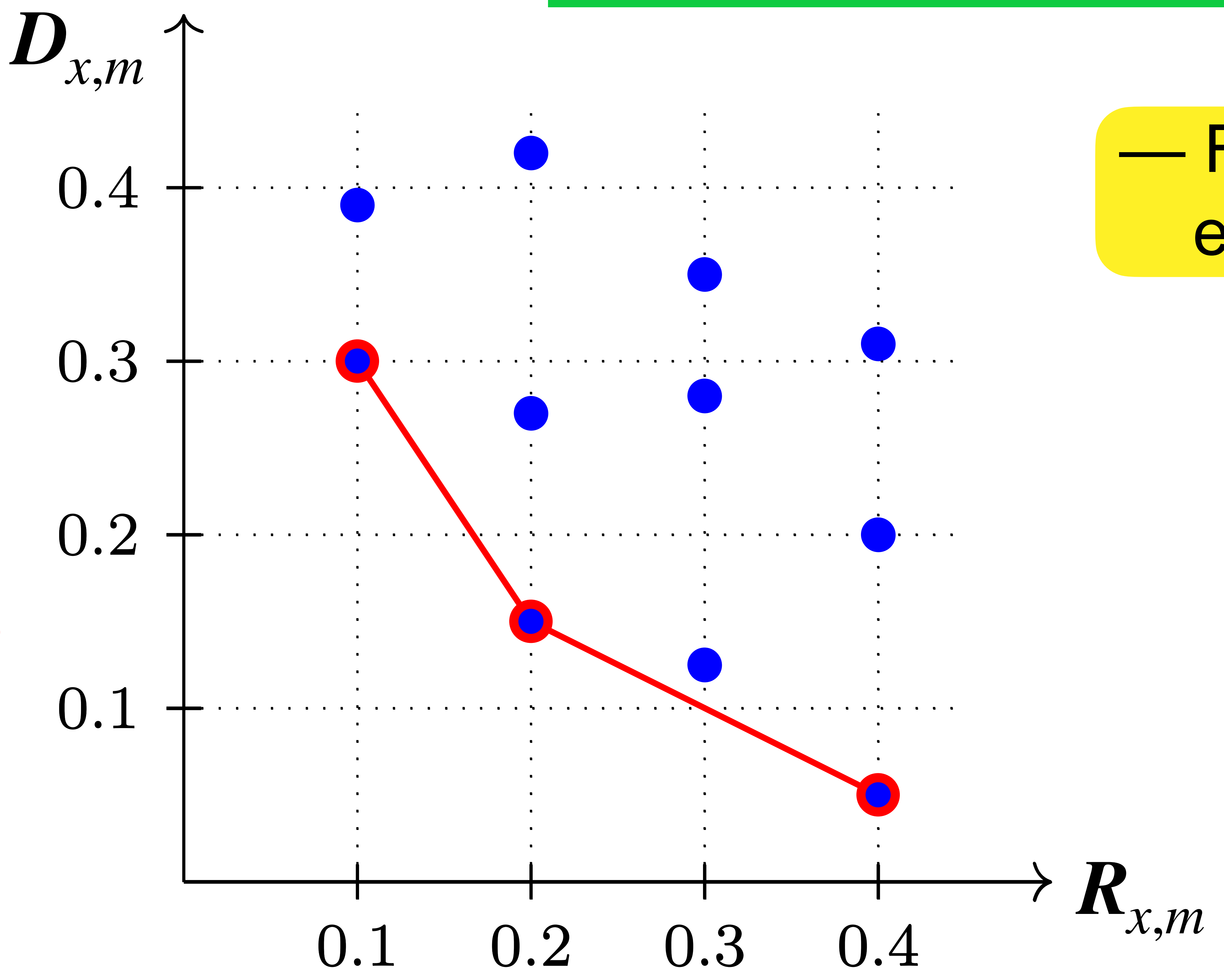
$$D^*(R) = \sup_{\lambda \geq 0} \left\{ -\lambda R + \sum_{x \in \mathcal{X}} \boxed{\min_{m \in \mathcal{M}_x} \left[\boldsymbol{D}_{x,m} + \lambda \boldsymbol{R}_{x,m} \right]} \right\}$$

— Fix $x \in \mathcal{X}$

$|\mathcal{M}_x| = 11$ here

— Plot $\left\{ \left(\boldsymbol{R}_{x,m}, \boldsymbol{D}_{x,m} \right) \right\}_{m \in \mathcal{M}_x}$

— lower-left concave envelope



— Find minimizing m for each $(\lambda, x) \longrightarrow$ easy!

Dual linear program: efficient geometric solution

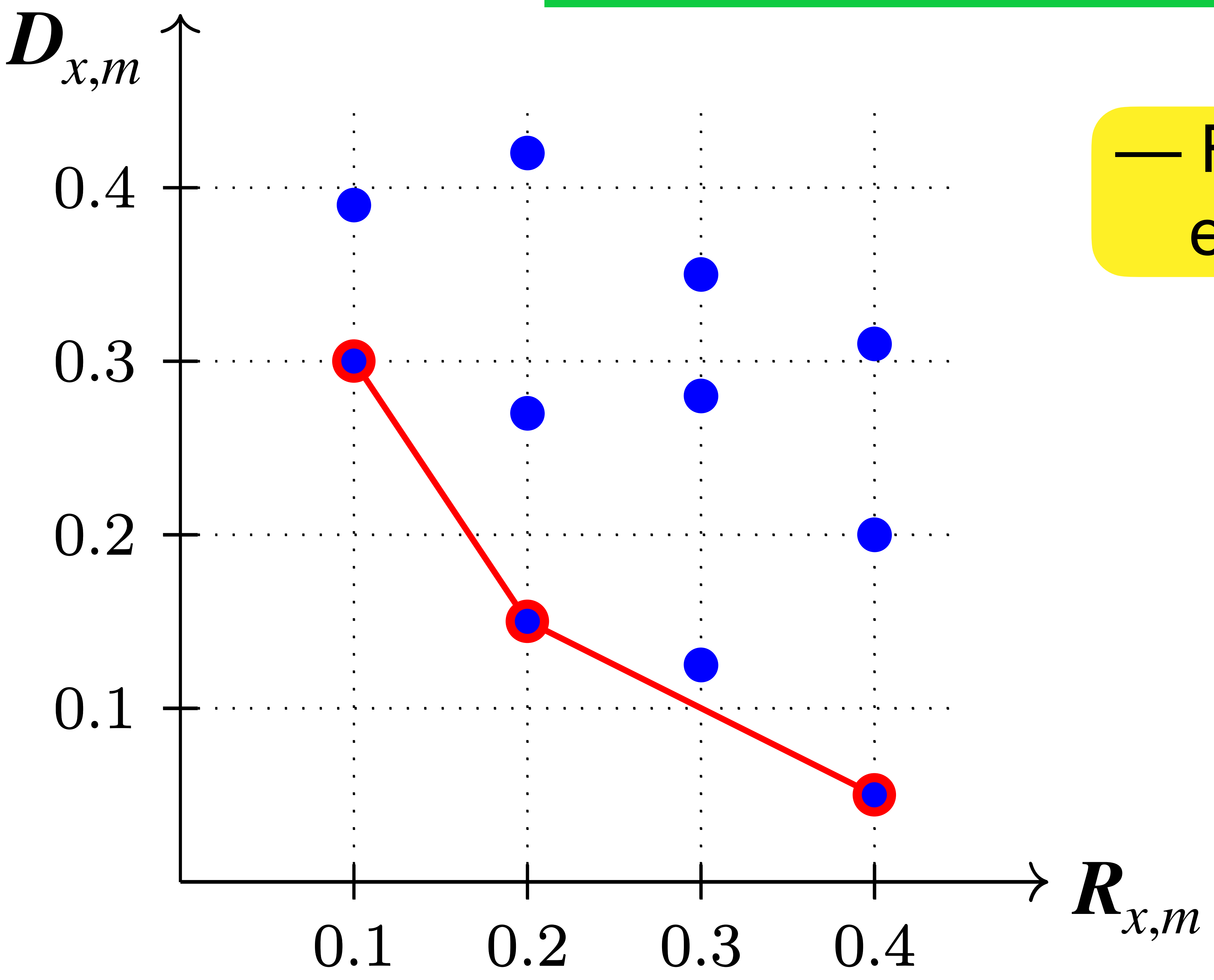
$$D^*(R) = \sup_{\lambda \geq 0} \left\{ -\lambda R + \sum_{x \in \mathcal{X}} \boxed{\min_{m \in \mathcal{M}_x} \left[\boldsymbol{D}_{x,m} + \lambda \boldsymbol{R}_{x,m} \right]} \right\}$$

— Fix $x \in \mathcal{X}$

$|\mathcal{M}_x| = 11$ here

— Plot $\left\{ \left(\boldsymbol{R}_{x,m}, \boldsymbol{D}_{x,m} \right) \right\}_{m \in \mathcal{M}_x}$

— lower-left concave envelope



— Find minimizing m for each $(\lambda, x) \longrightarrow$ easy!

Dual linear program: efficient geometric solution

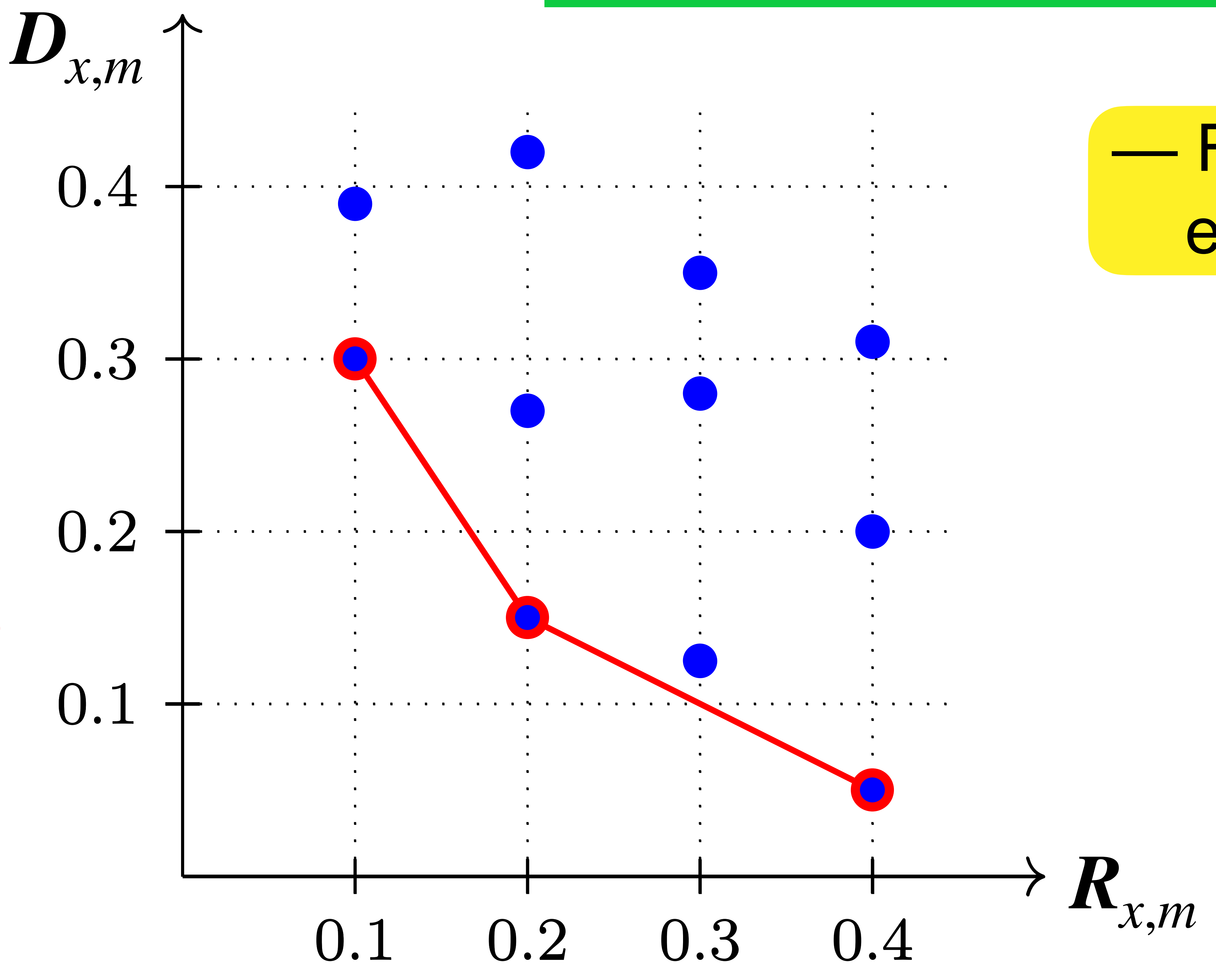
$$D^*(R) = \sup_{\lambda \geq 0} \left\{ -\lambda R + \sum_{x \in \mathcal{X}} \boxed{\min_{m \in \mathcal{M}_x} \left[\boldsymbol{D}_{x,m} + \lambda \boldsymbol{R}_{x,m} \right]} \right\}$$

— Fix $x \in \mathcal{X}$

$|\mathcal{M}_x| = 11$ here

— Plot $\left\{ \left(\boldsymbol{R}_{x,m}, \boldsymbol{D}_{x,m} \right) \right\}_{m \in \mathcal{M}_x}$

— lower-left concave envelope



— Find minimizing m for each $(\lambda, x) \longrightarrow$ easy!

$\approx 32{,}000^{100}$

Dual linear program: efficient geometric solution

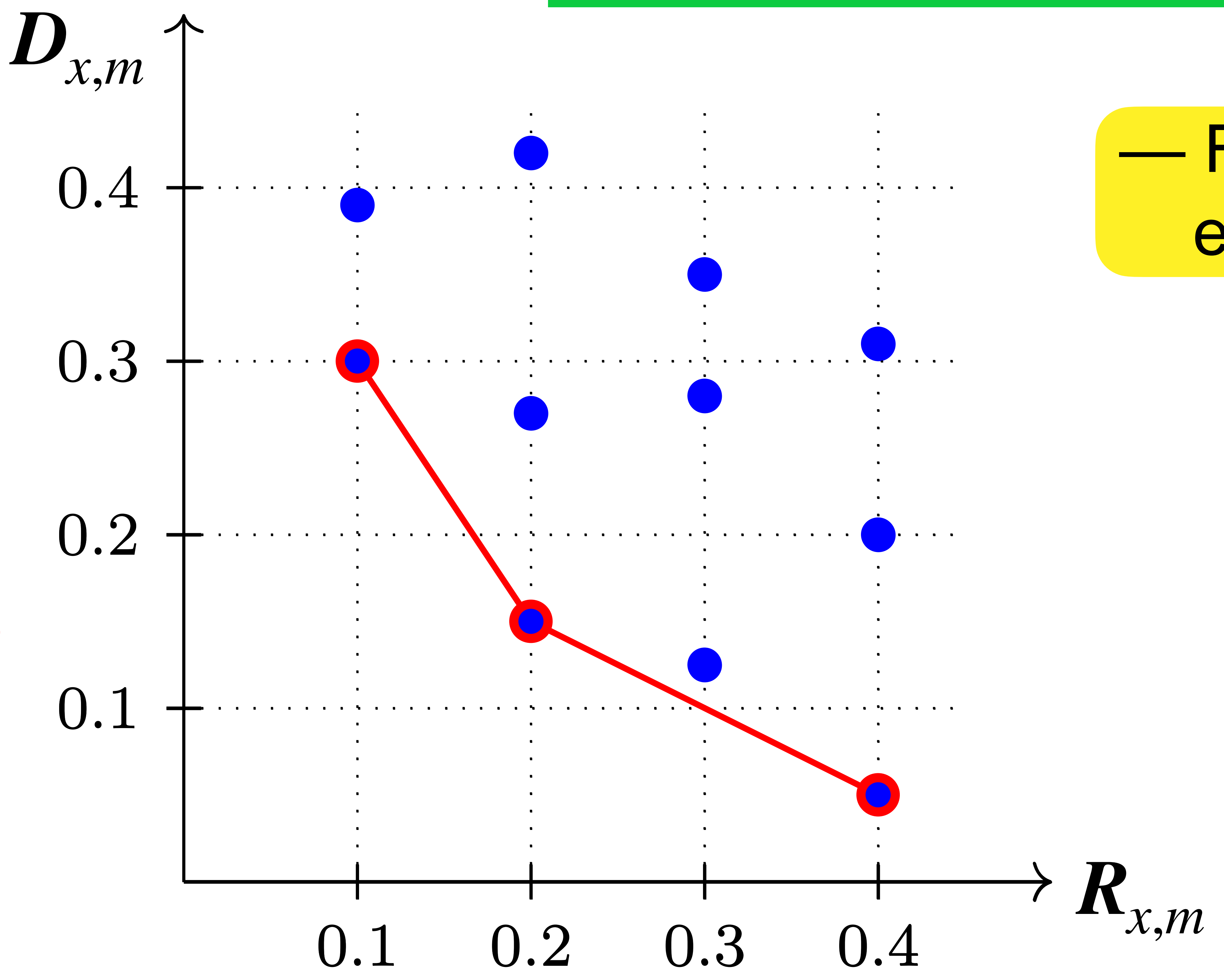
$$D^*(R) = \sup_{\lambda\ge 0}\left\{-\lambda R + \sum_{x\in\mathcal{X}} \boxed{\min_{m\in\mathcal{M}_x}\left[\boldsymbol{D}_{x,m}+\lambda \boldsymbol{R}_{x,m}\right]}\right\}$$

— Fix $x \in \mathcal{X}$

$|\mathcal{M}_x| = 11$ here

— Plot $\left\{ \left(\boldsymbol{R}_{x,m}, \boldsymbol{D}_{x,m}\right) \right\}_{m\in\mathcal{M}_x}$

— lower-left concave envelope



— Find minimizing m for each $(\lambda, x) \longrightarrow$ easy!

$\approx 32{,}000^{100}$

Dual linear program: efficient geometric solution

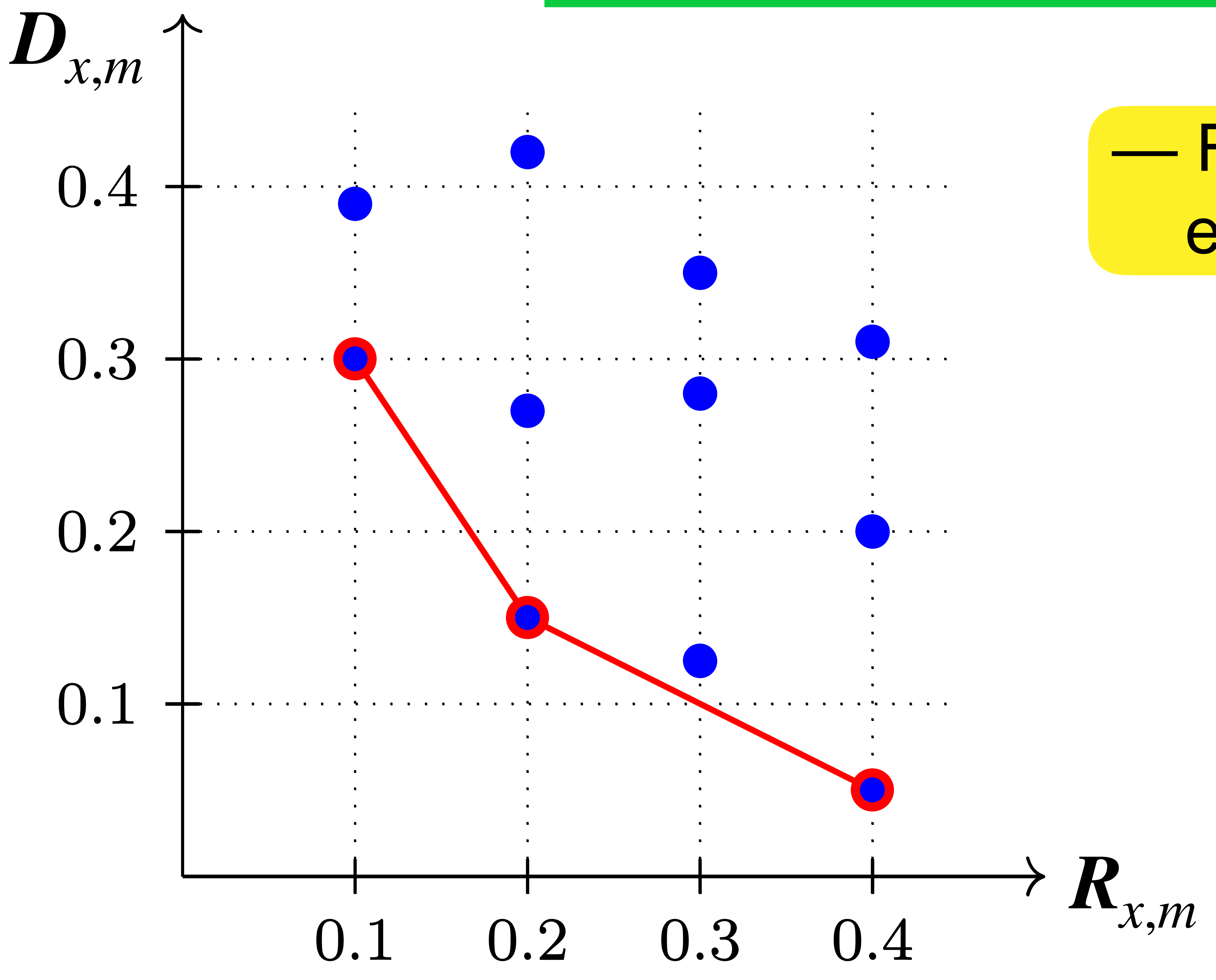
$$D^*(R) = \sup_{\lambda \geq 0} \left\{ -\lambda R + \sum_{x \in \mathcal{X}} \boxed{\min_{m \in \mathcal{M}_x} \left[\boldsymbol{D}_{x,m} + \lambda \boldsymbol{R}_{x,m} \right]} \right\}$$

— Fix $x \in \mathcal{X}$

$|\mathcal{M}_x| = 11$ here

— Plot $\left\{ \left(\boldsymbol{R}_{x,m}, \boldsymbol{D}_{x,m} \right) \right\}_{m \in \mathcal{M}_x}$

— lower-left concave envelope



— Find minimizing m for each $(\lambda, x) \longrightarrow$ easy!

$\approx 32{,}000^{100}$

$\downarrow$

≈ 100

Dual linear program: efficient geometric solution

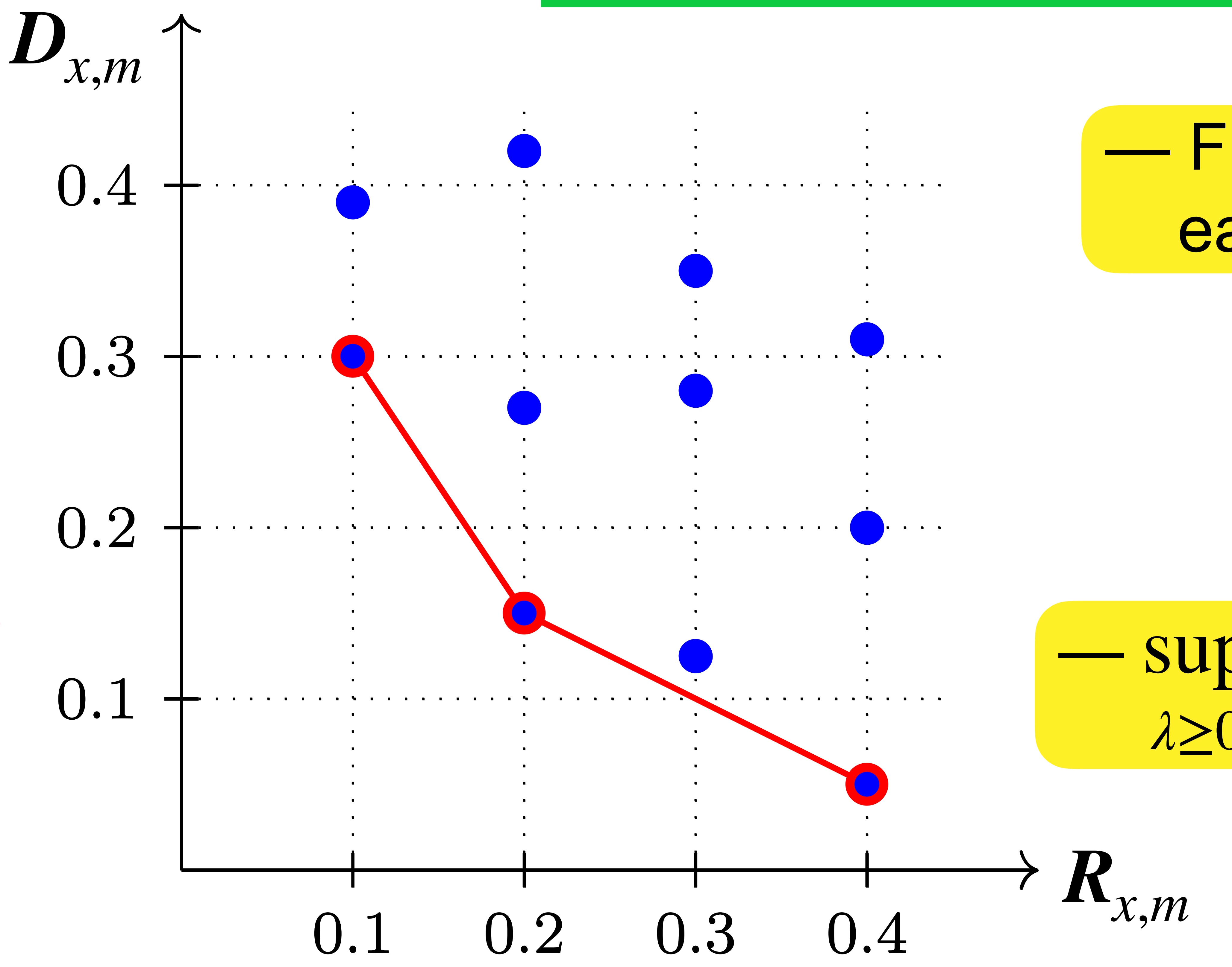
$$D^*(R) = \sup_{\lambda\ge 0}\left\{-\lambda R + \sum_{x\in\mathcal{X}} \boxed{\min_{m\in\mathcal{M}_x}\left[\boldsymbol{D}_{x,m}+\lambda \boldsymbol{R}_{x,m}\right]}\right\}$$

— Fix $x \in \mathcal{X}$

$|\mathcal{M}_x| = 11$ here

— Plot $\left\{\left(\boldsymbol{R}_{x,m}, \boldsymbol{D}_{x,m}\right)\right\}_{m\in\mathcal{M}_x}$

— lower-left concave envelope



— Find minimizing m for each $(\lambda, x) \longrightarrow$ easy!

$\approx 32{,}000^{100}$

$\downarrow$

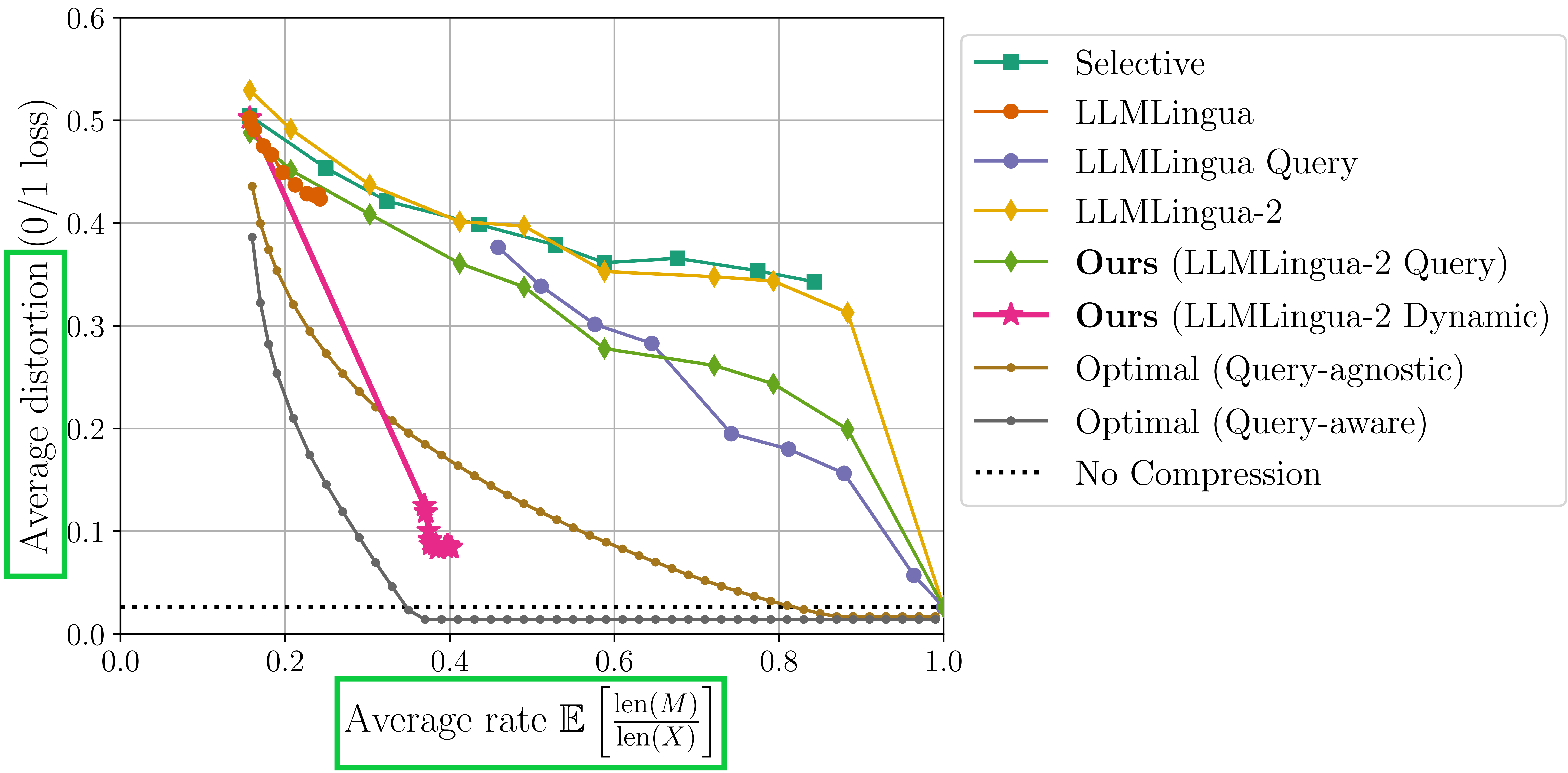
≈ 100

— $\sup_{\lambda\ge 0} \longrightarrow$ max over finite set of slopes

Summary

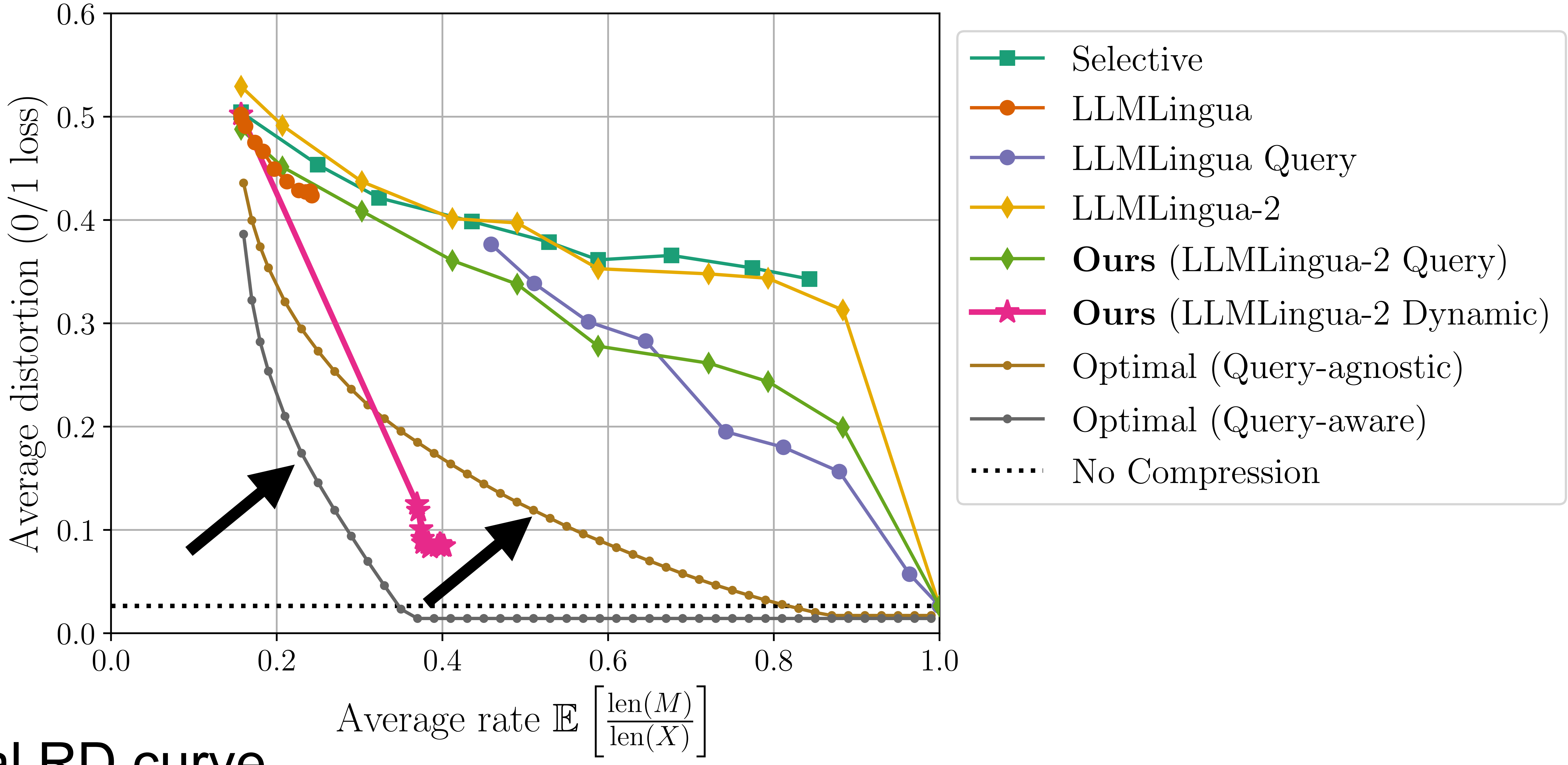
Summary

1. RD framework



Summary

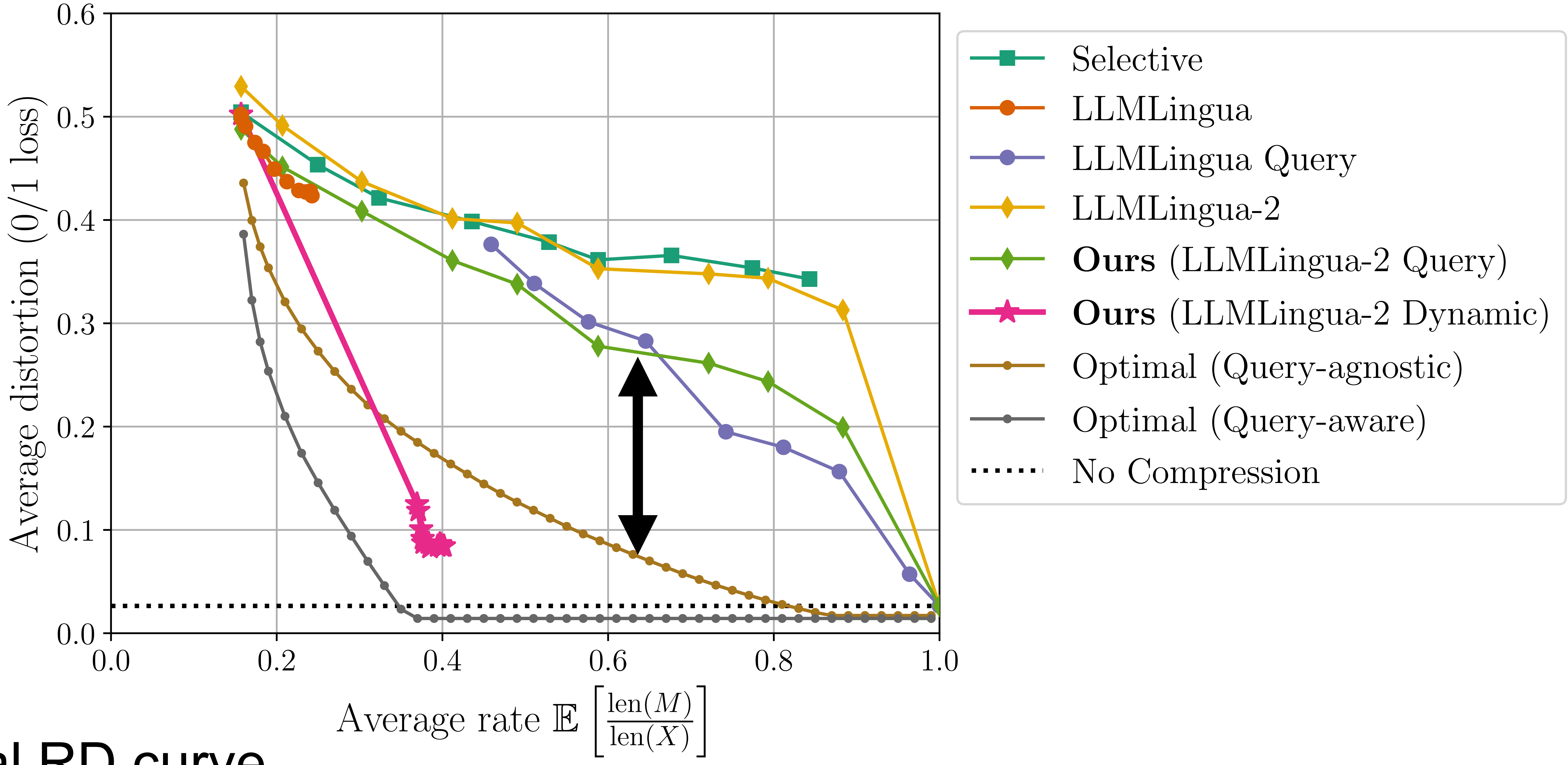
1. RD framework



2. Compute optimal RD curve

Summary

1. RD framework

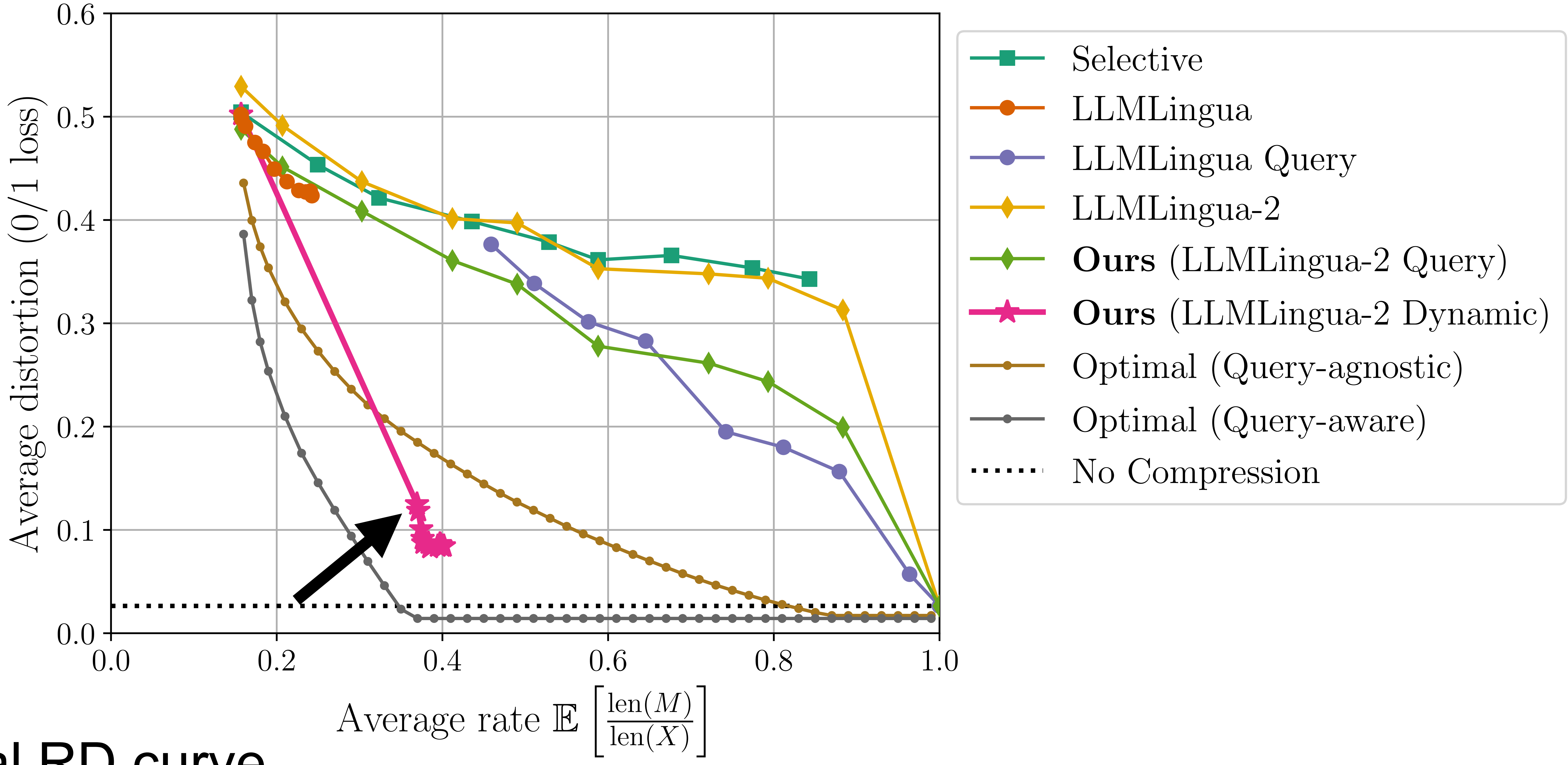


2. Compute optimal RD curve

3. Observe a large gap

Summary

1. RD framework



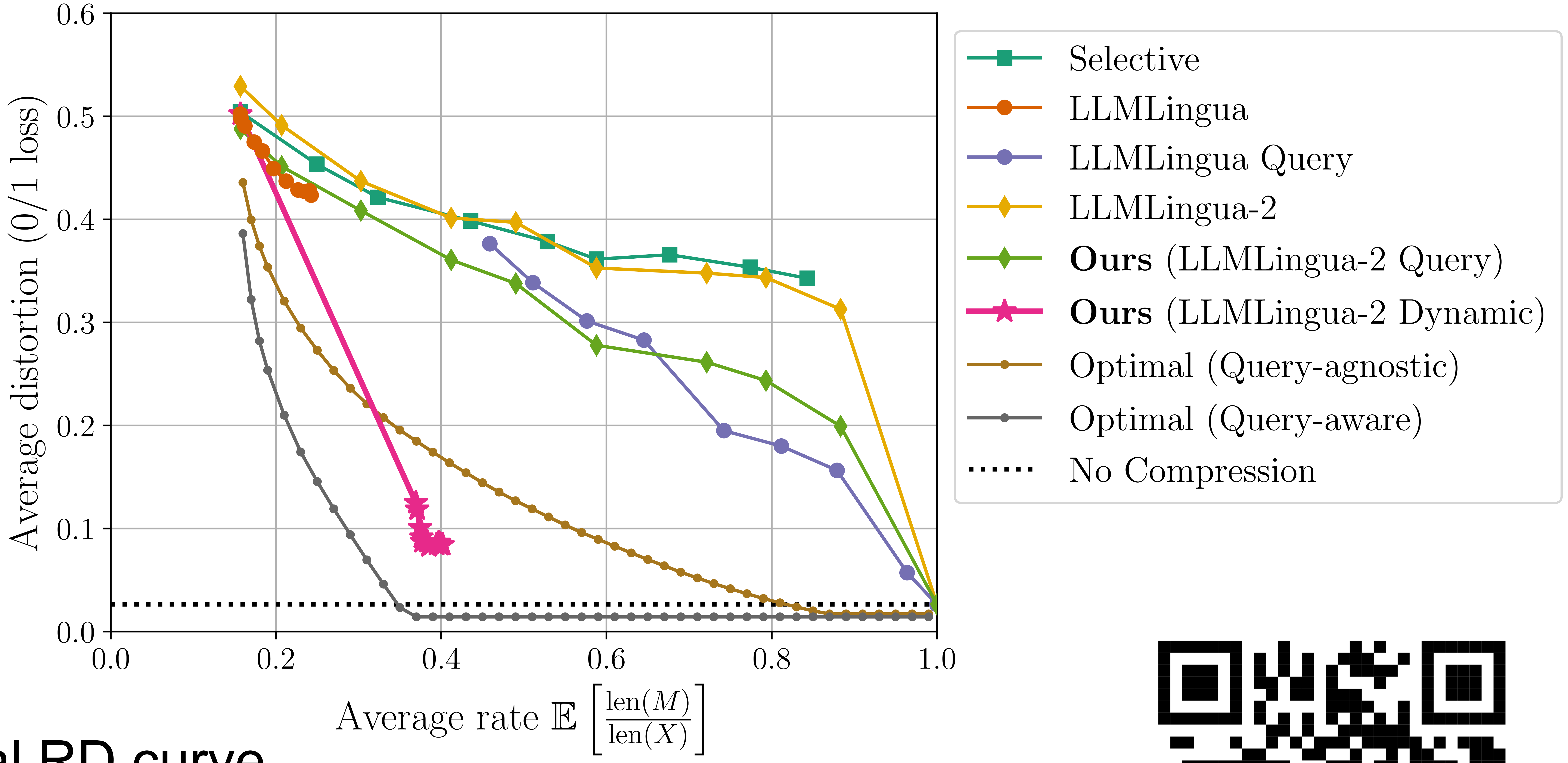
2. Compute optimal RD curve

3. Observe a large gap

4. New methods

Summary

1. RD framework



2. Compute optimal RD curve

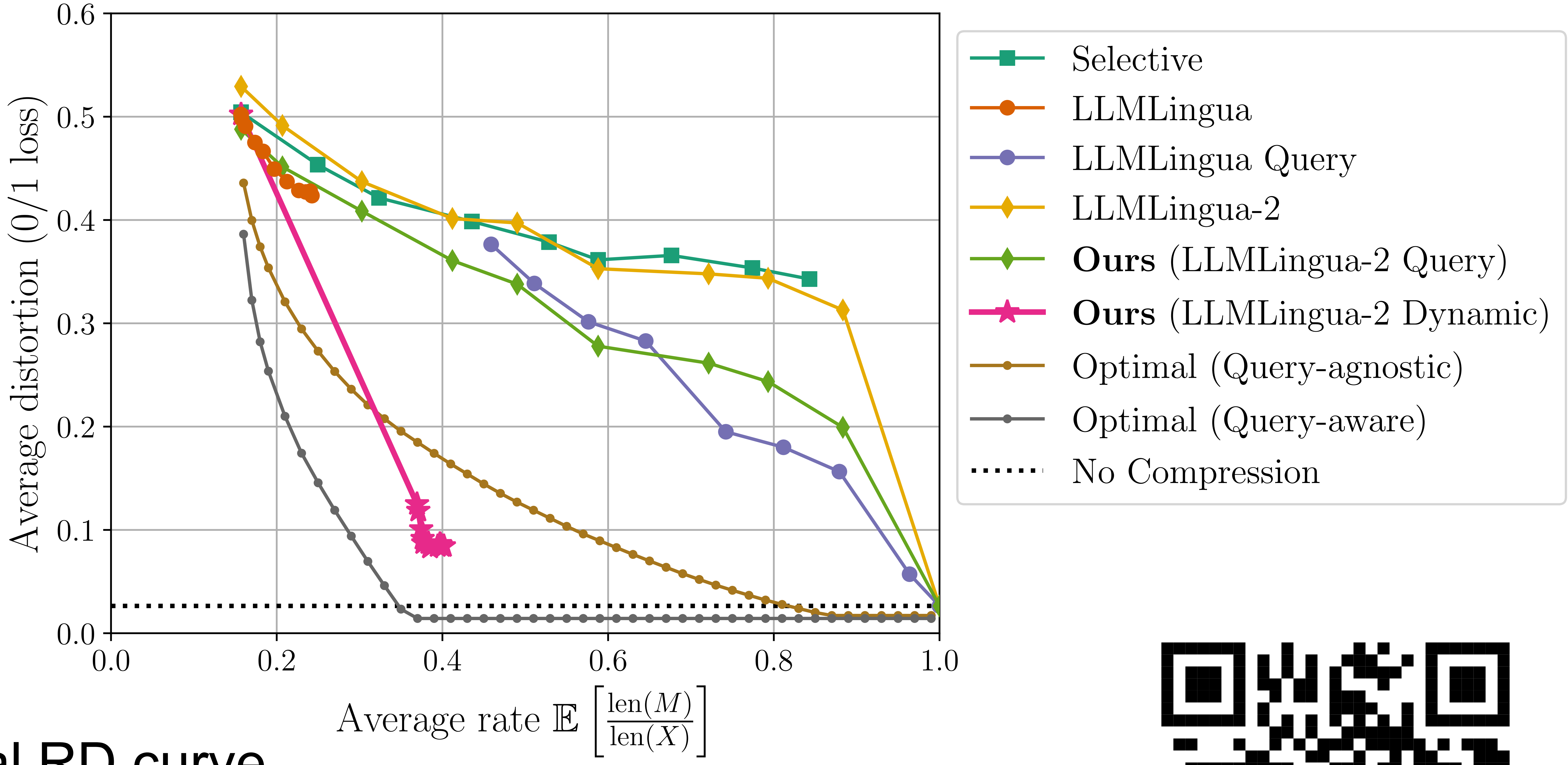
3. Observe a large gap

4. New methods

arXiv:2407.15504

Summary

1. RD framework



2. Compute optimal RD curve

3. Observe a large gap

4. New methods

Thank you!

arXiv:2407.15504