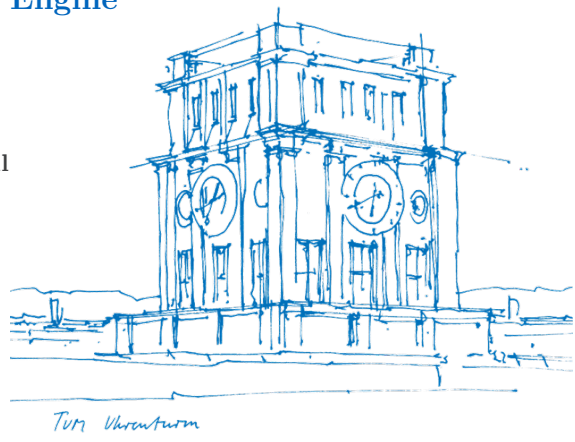
Mímir - A Stateless Private Search Engine

Emma Munisamy*, Gökberk Erdoğan*, Kamil Kluczniak

*Technical University of Munich

WPEC 2026
NIST Workshop on Privacy-Enhancing Cryptography

October 26th



Homomorphic Encryption (HE)

- Computations on encrypted data

$$m_1 + m_2 \;=\; \mathsf{Dec}\big(\mathsf{Enc}(m_1) + \mathsf{Enc}(m_2)\big)$$

Homomorphic Encryption (HE)

- Computations on encrypted data

$$m_1 + m_2 \;=\; \mathsf{Dec}\big(\mathsf{Enc}(m_1) + \mathsf{Enc}(m_2)\big)$$

- Additions, multiplications, etc. $\Longrightarrow$ **in theory any calculation can be done**

Homomorphic Encryption (HE)

- Computations on encrypted data

$$m_1 + m_2 \;=\; \mathsf{Dec}\big(\mathsf{Enc}(m_1) + \mathsf{Enc}(m_2)\big)$$

- Additions, multiplications, etc. $\implies$ **in theory any calculation can be done**
- Based on lattices (LWE, RLWE) $\implies$ **post-quantum secure**

Homomorphic Encryption (HE)

- Computations on encrypted data

$$m_1 + m_2 \;=\; \mathsf{Dec}\big(\mathsf{Enc}(m_1) + \mathsf{Enc}(m_2)\big)$$

- Additions, multiplications, etc. $\implies$ **in theory any calculation can be done**
- Based on lattices (LWE, RLWE) $\implies$ **post-quantum secure**
- Drawback: **large ciphertexts**, **slow computation**

Homomorphic Encryption (HE)

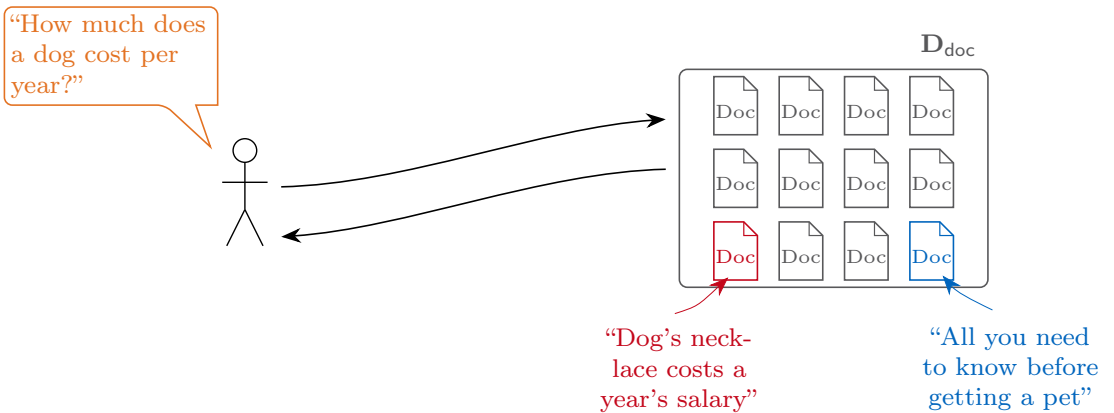
- Computations on encrypted data

$$m_1 + m_2 \;=\; \mathsf{Dec}\big(\mathsf{Enc}(m_1) + \mathsf{Enc}(m_2)\big)$$

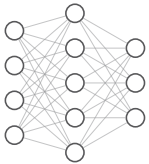
- Additions, multiplications, etc. $\implies$ **in theory any calculation can be done**
- Based on lattices (LWE, RLWE) $\implies$ **post-quantum secure**
- Drawback: **large ciphertexts**, **slow computation**

This work: How practical is HE for real-world applications?

Text Search


"How much does a dog cost per year?"
$\mathbf{D}_{\text{doc}}$
Doc
"Dog's necklace costs a year's salary"
"All you need to know before getting a pet"

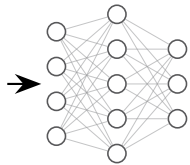
Semantic Embedding Models



Semantic Embedding Models

"All you need to know before getting a pet"

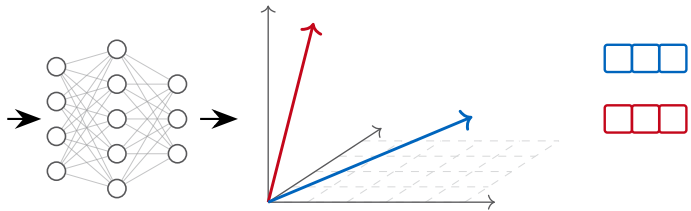
"Dog's necklaces cost a year's salary"



Semantic Embedding Models

"All you need to know before getting a pet"

"Dog's necklaces cost a year's salary"

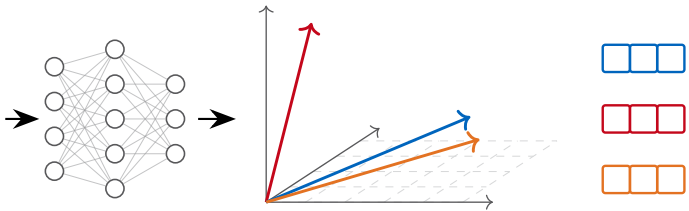


Semantic Embedding Models

"All you need to know before getting a pet"

"Dog's necklaces cost a year's salary"

"How much does a dog cost per year?"

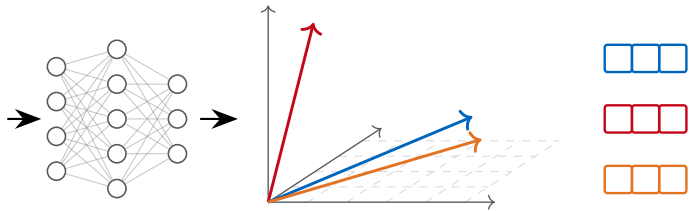


Semantic Embedding Models

"All you need to know before getting a pet"

"Dog's necklaces cost a year's salary"

"How much does a dog cost per year?"



semantic similarity $\Longrightarrow$ "closeness" of embeddings

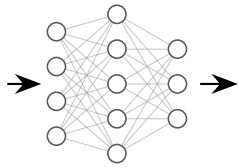
large dot product = related semantic content

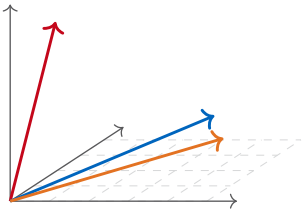
Semantic Embedding Models

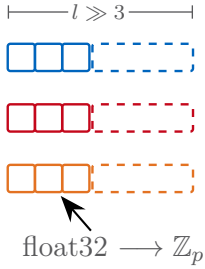
"All you need to know before getting a pet"

"Dog's necklaces cost a year's salary"

"How much does a dog cost per year?"



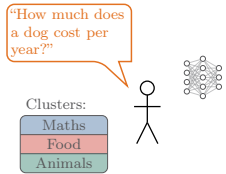


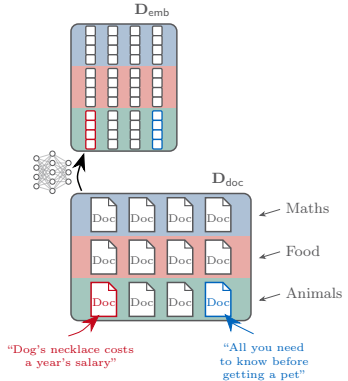


semantic similarity $\Longrightarrow$ "closeness" of embeddings

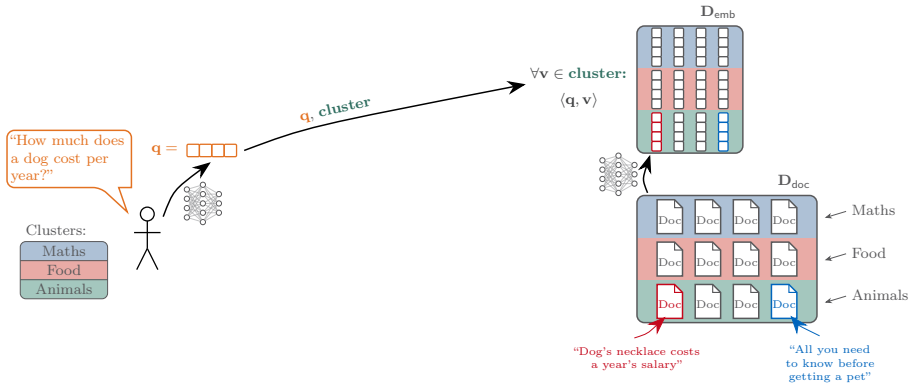
large dot product = related semantic content

Semantic Text Search with Clustering

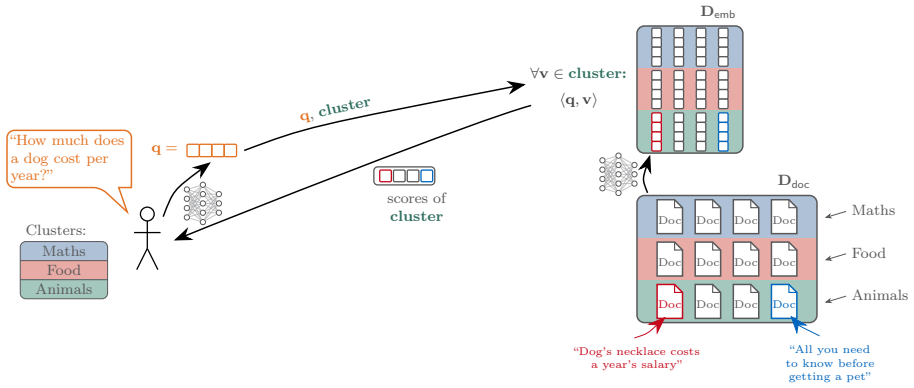




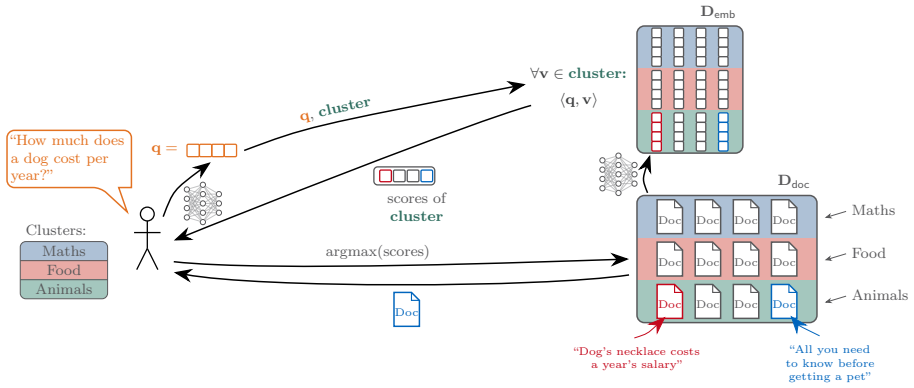
Semantic Text Search with Clustering



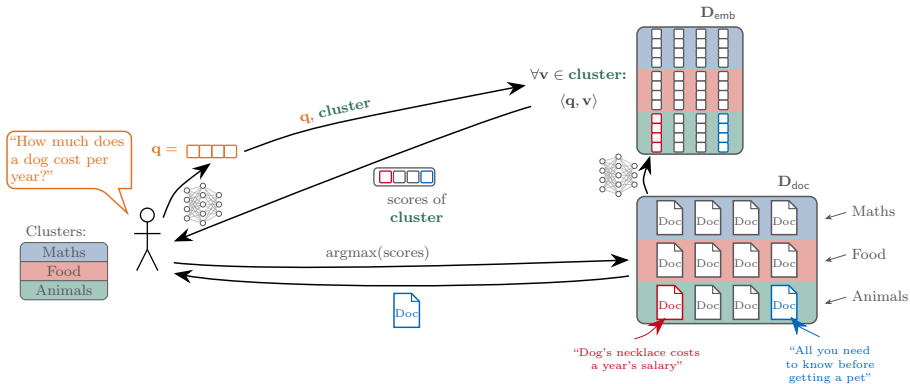
Semantic Text Search with Clustering



Semantic Text Search with Clustering



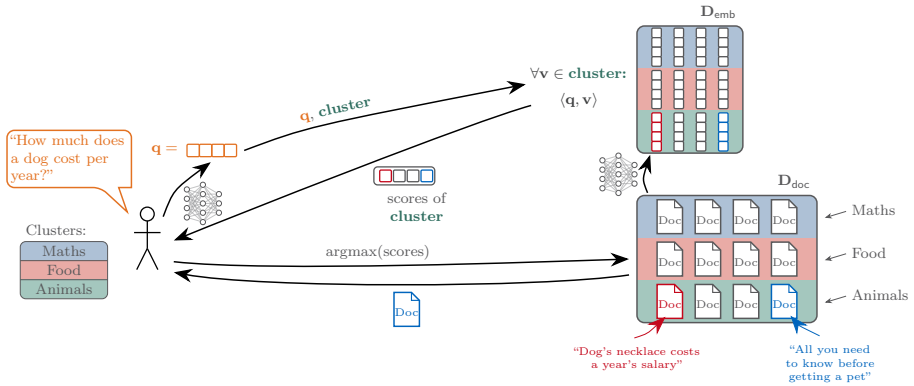
Semantic Text Search with Clustering



Private Semantic Search

An honest-but-curious server cannot distinguish any two given text queries from any public protocol state with more than negligible advantage.

Semantic Text Search with Clustering

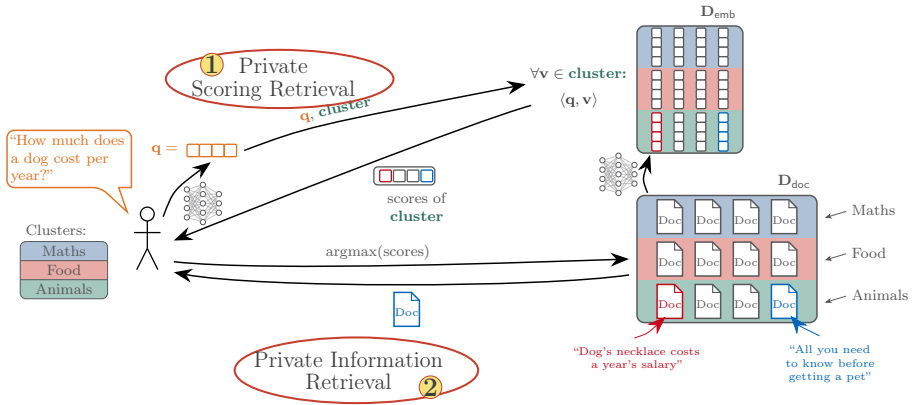


Private Semantic Search

An honest-but-curious server cannot distinguish any two given text queries from any public protocol state with more than negligible advantage.

$\Longrightarrow$ We use **Homomorphic Encryption**

Semantic Text Search with Clustering

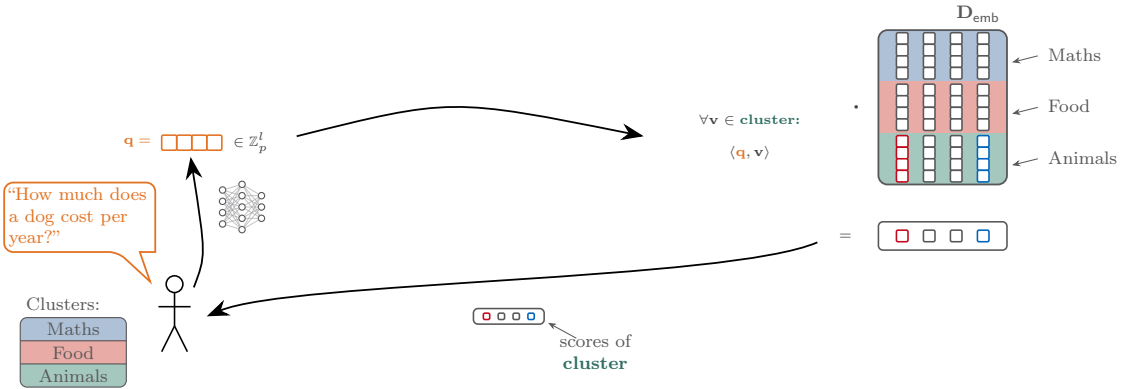


Private Semantic Search

An honest-but-curious server cannot distinguish any two given text queries from any public protocol state with more than negligible advantage.

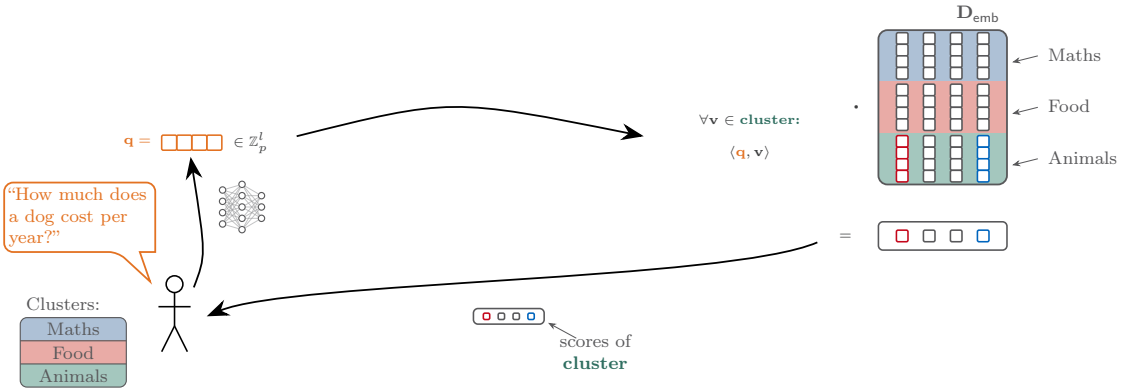
$\Longrightarrow$ We use **Homomorphic Encryption**

Basic Private Scoring Retrieval



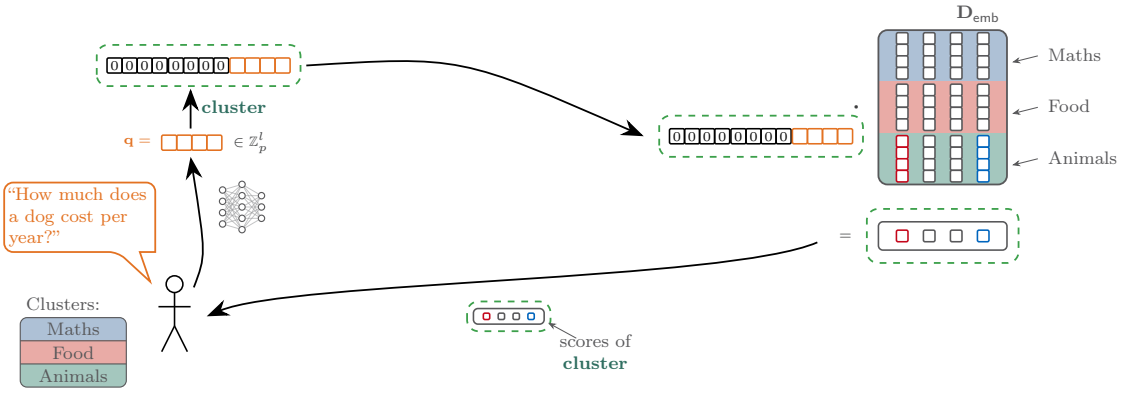
Basic Private Scoring Retrieval

[HDCGZ23] Henzinger et al. "Private web search with Tiptoe." *Proceedings of the 29th symposium on operating systems principles.* 2023



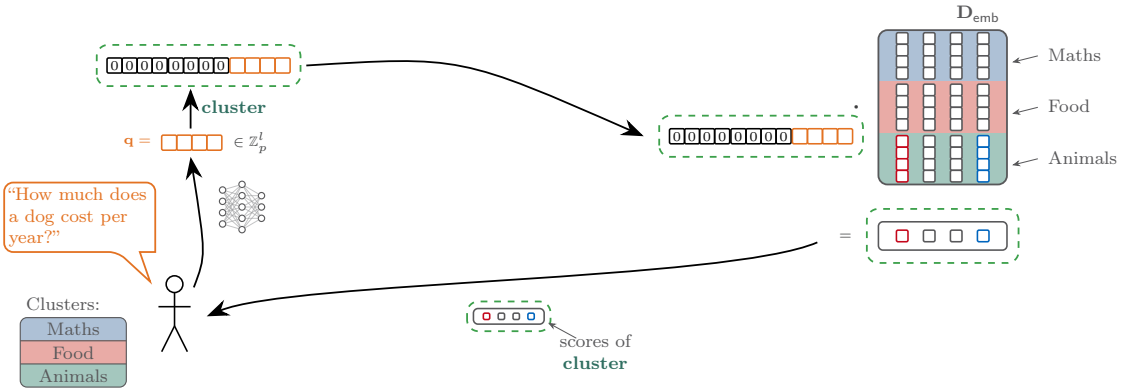
Basic Private Scoring Retrieval

[HDCGZ23] Henzinger et al. "Private web search with Tiptoe." *Proceedings of the 29th symposium on operating systems principles.* 2023



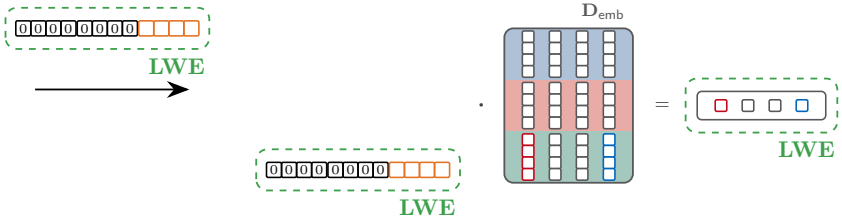
Basic Private Scoring Retrieval

[HDCGZ23] Henzinger et al. "Private web search with Tiptoe." *Proceedings of the 29th symposium on operating systems principles.* 2023

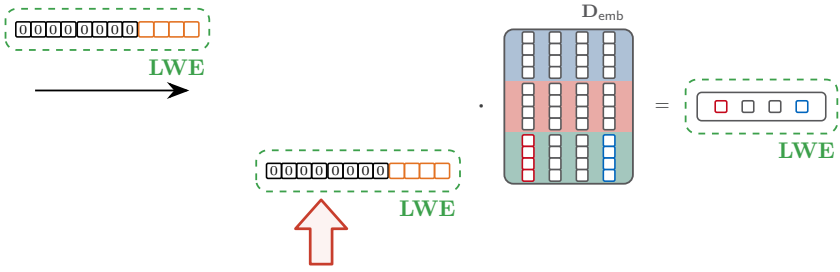


Mímir (this work):
Private Search with 3 improvements for smaller communication cost.

Mímir: 3 Communication Cost Optimizations



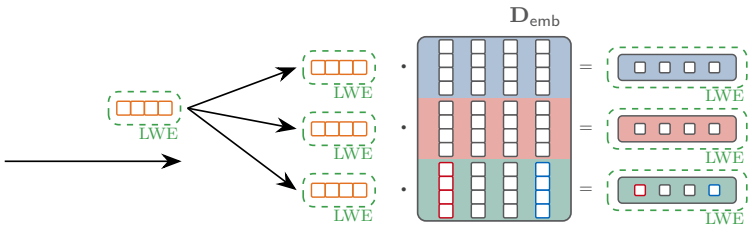
Mímir: 3 Communication Cost Optimizations



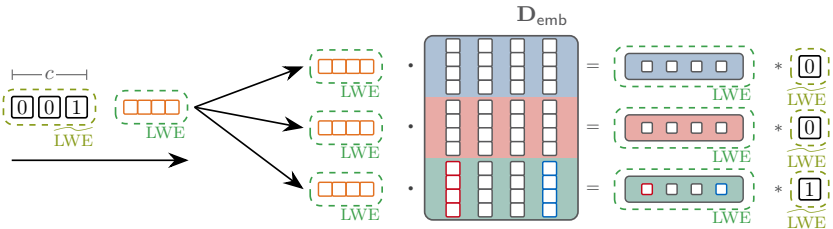
1. Remove redundant 0s

e.g. $c = 100$ clusters $\Longrightarrow$ 99 % are encrypted 0s

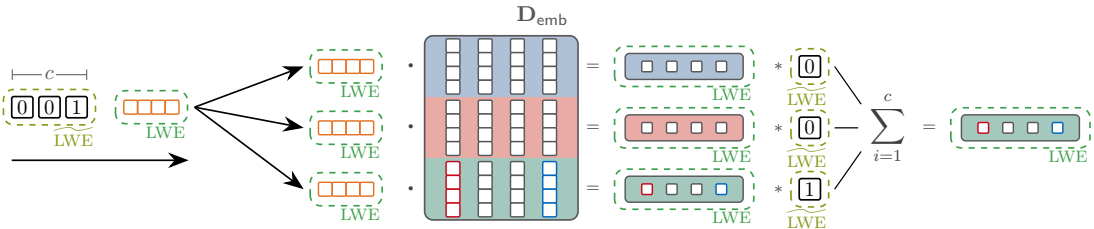
Improvement 1: Remove Redundant 0s



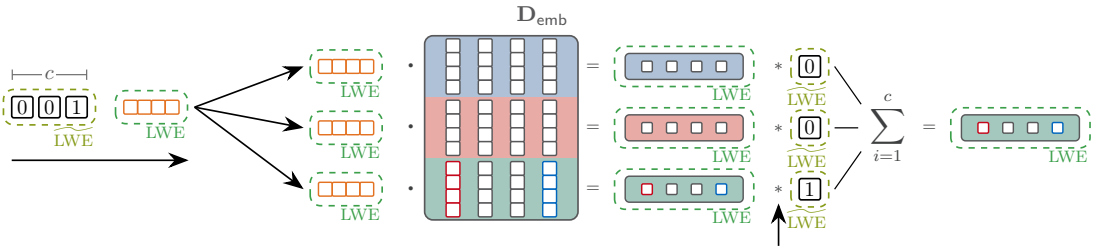
Improvement 1: Remove Redundant 0s



Improvement 1: Remove Redundant 0s

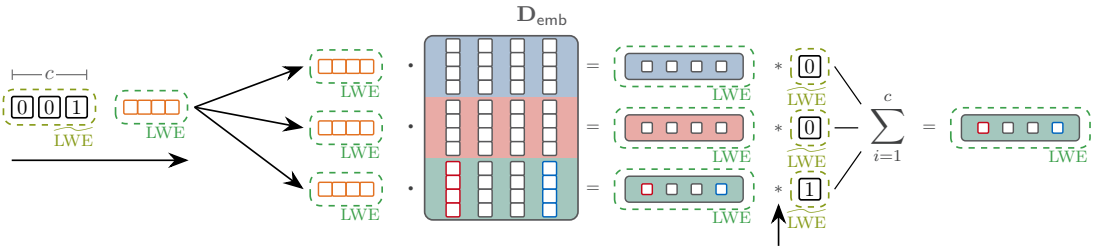


Improvement 1: Remove Redundant 0s



Noise-friendly homomorphic multiplication via $\widetilde{\text{LWE}}$ ciphertexts

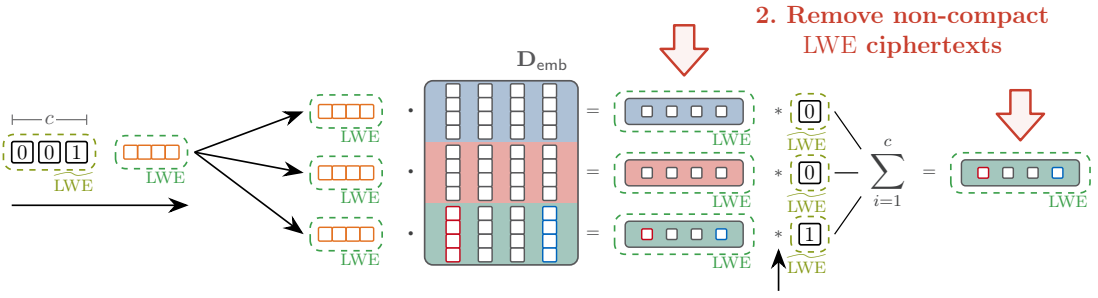
Improvement 1: Remove Redundant 0s



Noise-friendly homomorphic multiplication via $\widetilde{\text{LWE}}$ ciphertexts

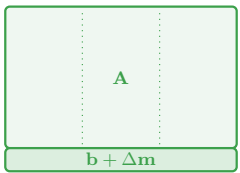
$\Longrightarrow$ Reduces the query length from $l \cdot c$ LWE ciphertexts to l LWE + c $\widetilde{\text{LWE}}$ ciphertexts

Improvement 1: Remove Redundant 0s

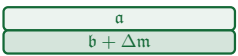


$\Longrightarrow$ Reduces the query length from $l \cdot c$ LWE ciphertexts
to l LWE + c $\widetilde{\text{LWE}}$ ciphertexts

Improvement 2: Use compact RLWE ciphertexts



$\in \mathbb{Z}_q^{(n+1)\times n}$



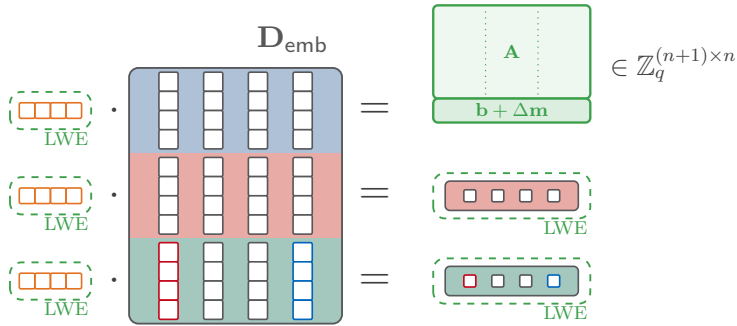
$\in \mathcal{R}_q^2$

with $\mathcal{R}_q = \mathbb{Z}_q[X]/(X^n + 1)$

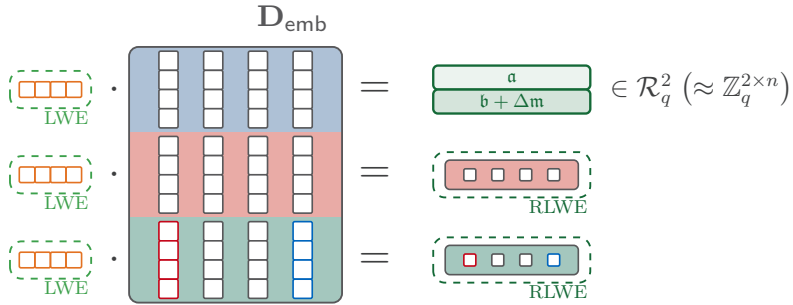
LWE ciphertext of message $\mathbf{m} \in \mathbb{Z}_p^n$

RLWE ciphertext of message $\mathfrak{m} \in \mathcal{R}_p \simeq \mathbb{Z}_p^n$

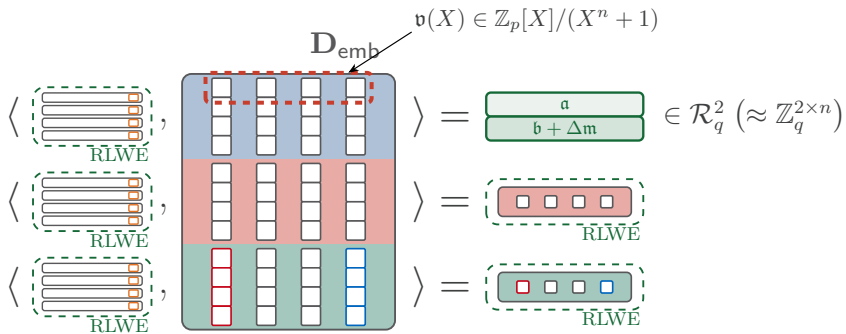
Improvement 2: Use compact RLWE ciphertexts



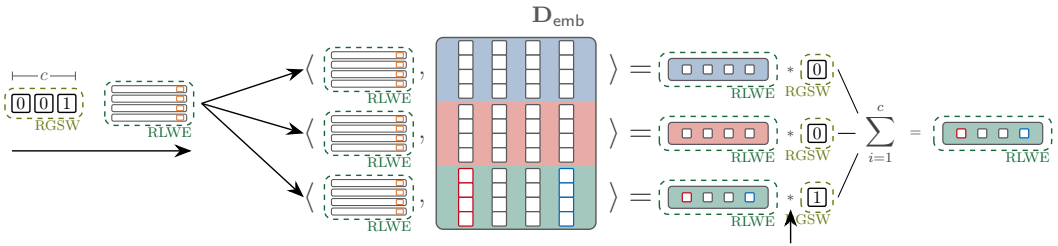
Improvement 2: Use compact RLWE ciphertexts



Improvement 2: Use compact RLWE ciphertexts

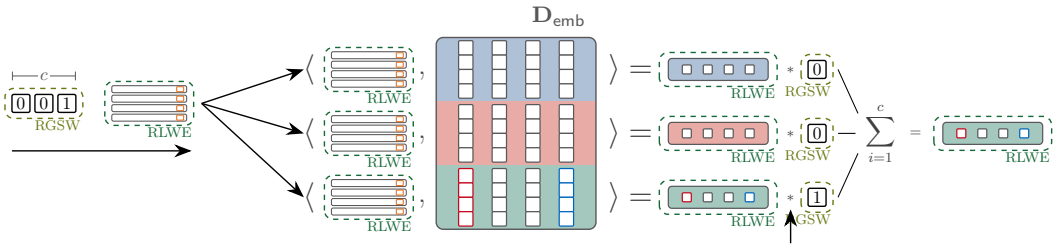


Improvement 2: Use compact RLWE ciphertexts



[DKMS23] De Micheli et al. "Faster Amortized FHEW Bootstrapping using Ring Automorphisms." *IACR Cryptology ePrint Archive, 2023/112.* 2023

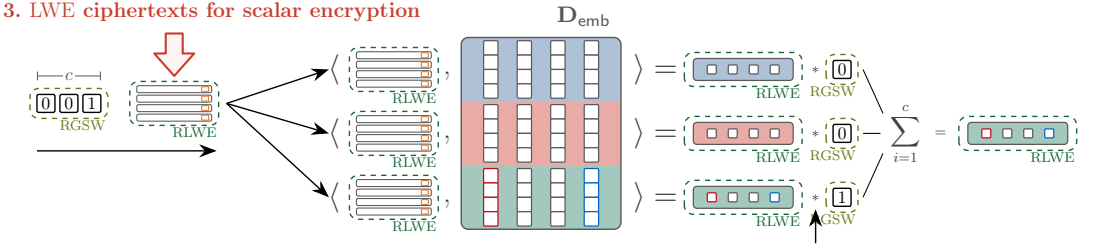
Improvement 2: Use compact RLWE ciphertexts



$\Longrightarrow$ RLWE ciphertexts reduce the download cost by a factor of n

[DKMS23] De Micheli et al. "Faster Amortized FHEW Bootstrapping using Ring Automorphisms." *IACR Cryptology ePrint Archive, 2023/112.* 2023

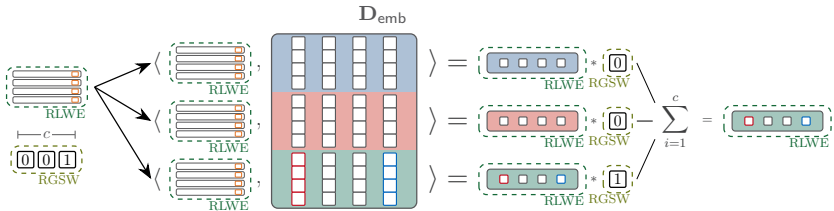
Improvement 2: Use compact RLWE ciphertexts



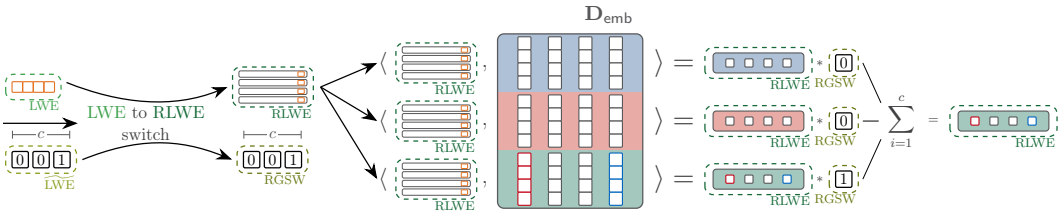
$\Longrightarrow$ RLWE ciphertexts reduce the download cost by a factor of n

[DKMS23] De Micheli et al. "Faster Amortized FHEW Bootstrapping using Ring Automorphisms." *IACR Cryptology ePrint Archive, 2023/112.* 2023

Improvement 3: LWE to RLWE switches for Upload

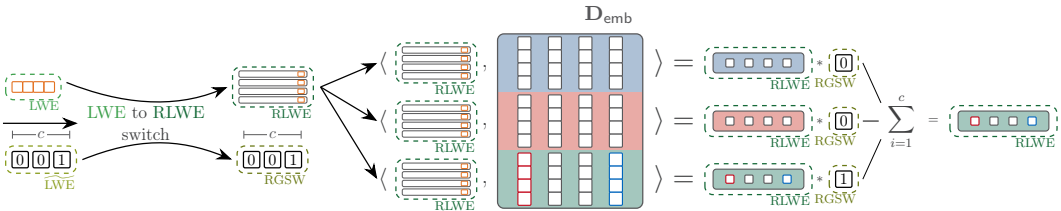


Improvement 3: LWE to RLWE switches for Upload



[CDKS20] Chen et al. "Efficient Homomorphic Conversion Between (Ring) LWE Ciphertexts." *IACR Cryptology ePrint Archive, 2020/015.* 2020

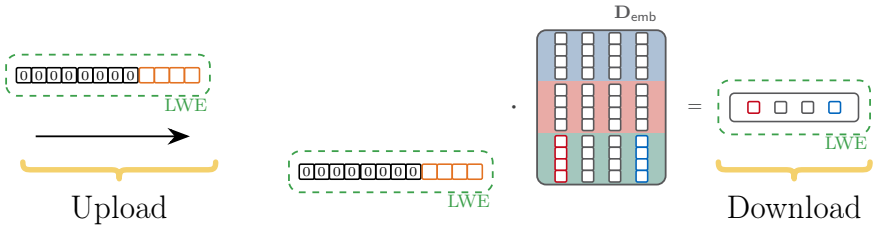
Improvement 3: LWE to RLWE switches for Upload



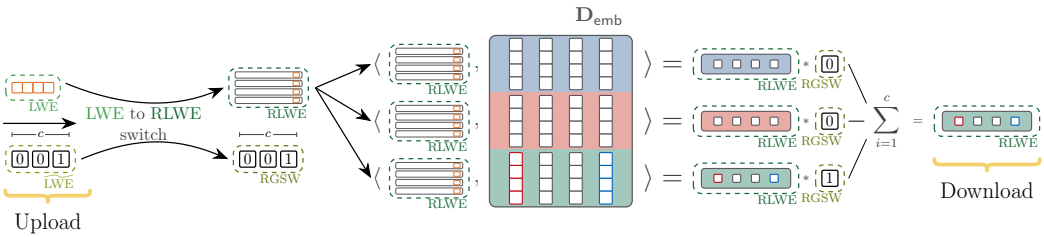
$\Longrightarrow$ LWE ciphertexts reduce the upload cost by a factor of n

[CDKS20] Chen et al. "Efficient Homomorphic Conversion Between (Ring) LWE Ciphertexts." *IACR Cryptology ePrint Archive, 2020/015.* 2020

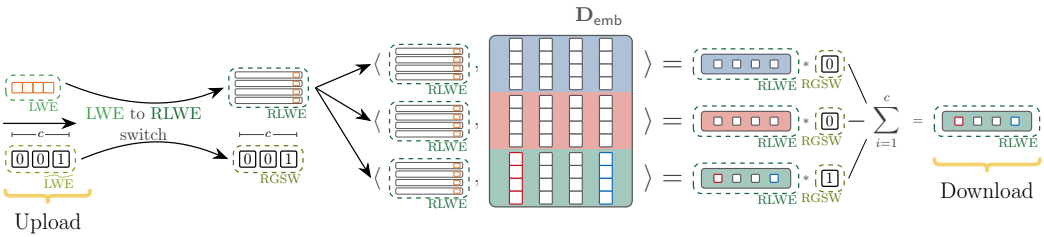
Recap: Trivial Solution



Mímir: Complete Protocol



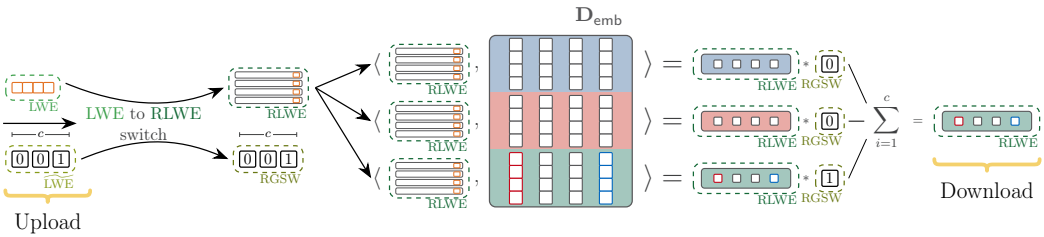
Mímir: Complete Protocol



Summary

Small upload cost $\longrightarrow$ LWE ciphertexts

Mímir: Complete Protocol

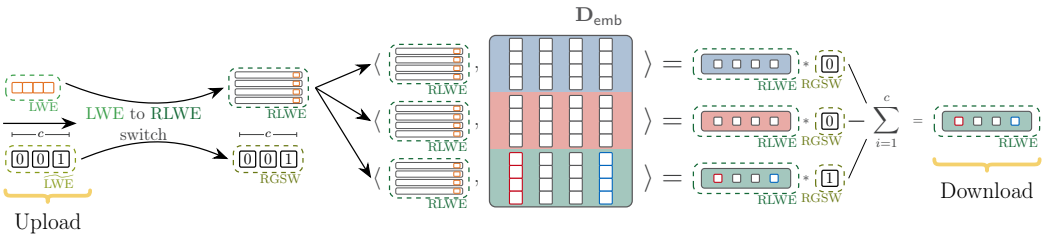


Summary

Small upload cost ⟶ LWE ciphertexts

Noise-friendly multiplications ⟶ RGSW ciphertexts

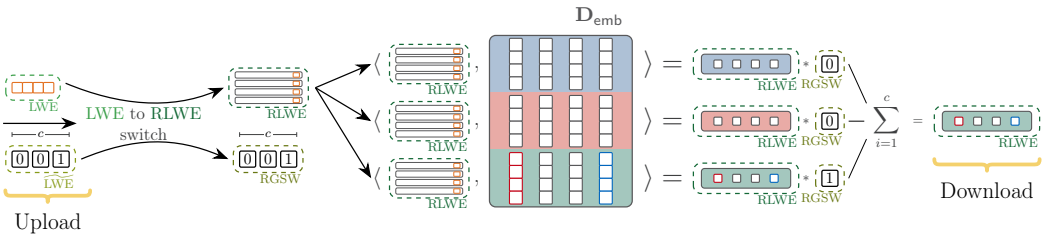
Mímir: Complete Protocol



Summary

Small upload cost ⟶ LWE ciphertexts
Noise-friendly multiplications ⟶ RGSW ciphertexts
Small download cost ⟶ RLWE ciphertexts

Mímir: Complete Protocol



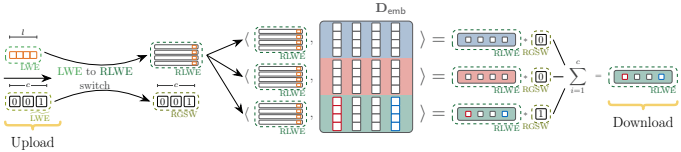
Summary

Small upload cost $\longrightarrow$ LWE ciphertexts
Noise-friendly multiplications $\longrightarrow$ RGSW ciphertexts
Small download cost $\longrightarrow$ RLWE ciphertexts

$\Longrightarrow$ Improved communication cost for **complex server work**

Evaluation: performance metrics

Search quality ↑ **Communication** ↓ **Latency** ↓



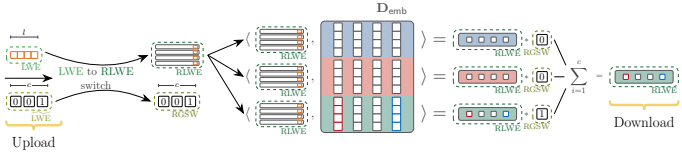
l: embedding length | c: # clusters

Evaluation: performance metrics

Search quality ↑ **Communication** ↓ **Latency** ↓

l ↑

c ↓



l: embedding length | c: # clusters

Evaluation: performance metrics

Search quality ↑

l ↑

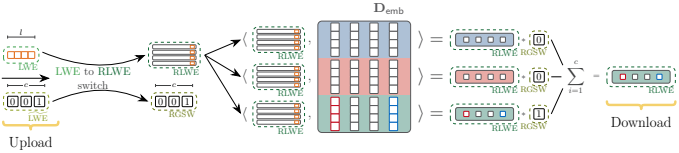
c ↓

Communication ↓

l ↓

c ↑

Latency ↓



l: embedding length | c: # clusters

Evaluation: performance metrics

Search quality ↑

l ↑

c ↓

Communication ↓

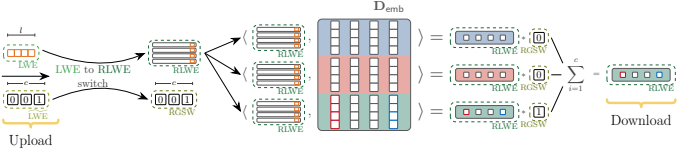
l ↓

c ↑

Latency ↓

l ↓

c ↓



l: embedding length | c: # clusters

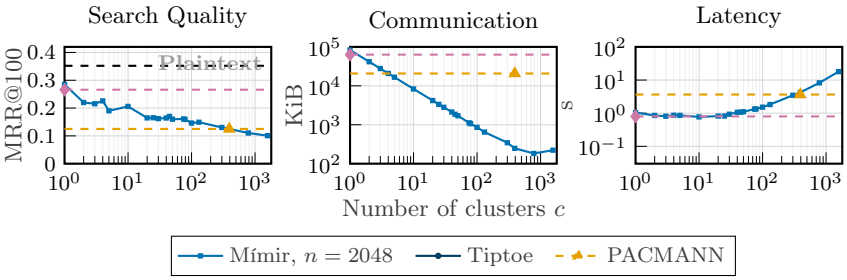
Evaluation: Clustering is the performance knob

Dataset: MS MARCO with ≈ 3 million websites

l: embedding length | c: # clusters

Evaluation: Clustering is the performance knob

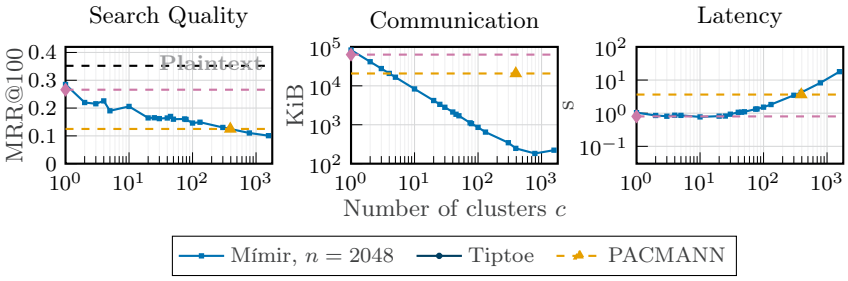
Dataset: MS MARCO with ≈ 3 million websites



l: embedding length | c: # clusters

Evaluation: Clustering is the performance knob

Dataset: MS MARCO with ≈ 3 million websites

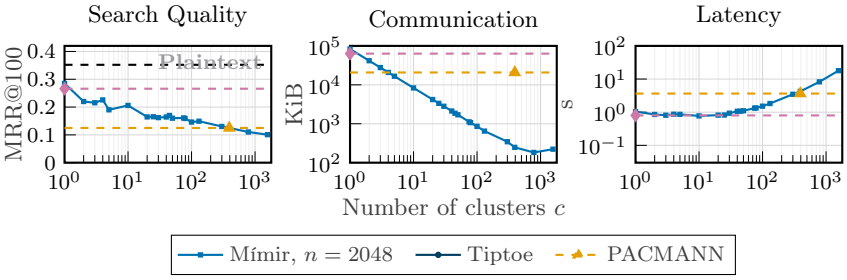


- **Tiptoe** [HDCGZ23]: optimized for **low communication cost**

l: embedding length | c: # clusters

Evaluation: Clustering is the performance knob

Dataset: MS MARCO with ≈ 3 million websites

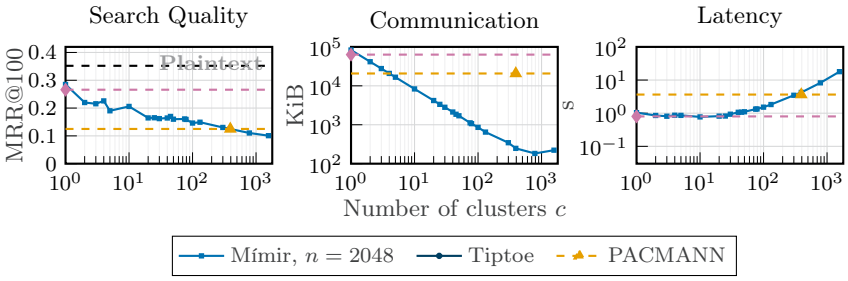


- **Tiptoe** [HDCGZ23]: optimized for **low communication cost**
- **PACMANN** [ZSF25]: optimized for **high accuracy**

l: embedding length | c: # clusters

Evaluation: Clustering is the performance knob

Dataset: MS MARCO with ≈ 3 million websites



- **Tiptoe** [HDCGZ23]: optimized for **low communication cost**
- **PACMANN** [ZSF25]: optimized for **high accuracy**

$\Longrightarrow$ both are designed for the **stateful** setting

l: embedding length | c: # clusters

Results - A Dynamic Database Setting

	c	Accuracy % of non-private	Communication per query	Latency total
Tiptoe [HDCGZ23]	393	35%	20.26 MiB	3.65 s

l: embedding length | c: # clusters

Results - A Dynamic Database Setting

	c	Accuracy % of non-private	Communication per query	Latency total
Tiptoe [HDCGZ23]	393	35%	20.26 MiB	3.65 s
Mímir I – lowest comm.	393	35% ≈	0.24 MiB ↓ ×84	4.23 s ↑ ×1.2

l: embedding length | c: # clusters

Results - A Dynamic Database Setting

	c	Accuracy % of non-private	Communication per query	Latency total
Tiptoe [HDCGZ23]	393	35%	20.26 MiB	3.65 s
Mímir I – lowest comm.	393	35% ≈	0.24 MiB ↓ ×84	4.23 s ↑ ×1.2
Mímir II – Tiptoe's comm.	4	63% ↑ ×1.8	20.34 MiB ≈	0.87 s ↓ ×4.2

l: embedding length | c: # clusters

Results - A Dynamic Database Setting

	c	Accuracy % of non-private	Communication per query	Latency total
Tiptoe [HDCGZ23]	393	35%	20.26 MiB	3.65 s
Mímir I – lowest comm.	393	35% ≈	0.24 MiB ↓ ×84	4.23 s ↑ ×1.2
Mímir II – Tiptoe's comm.	4	63% ↑ ×1.8	20.34 MiB ≈	0.87 s ↓ ×4.2
PACMANN [ZSF25]	1	76%	61.85 MiB	0.80 s

l: embedding length | c: # clusters

Results - A Dynamic Database Setting

	c	Accuracy % of non-private	Communication per query	Latency total
Tiptoe [HDCGZ23]	393	35%	20.26 MiB	3.65 s
Mímir I – lowest comm.	393	35% ≈	0.24 MiB ↓ ×84	4.23 s ↑ ×1.2
Mímir II – Tiptoe's comm.	4	63% ↑ ×1.8	20.34 MiB ≈	0.87 s ↓ ×4.2
PACMANN [ZSF25]	1	76%	61.85 MiB	0.80 s
Mímir III – highest accuracy	1	81% ↑ ×1.1	44.46 MiB ↓ ×1.4	1.66 s ↑ ×2.1

l: embedding length | c: # clusters

Results - A Dynamic Database Setting

	c	Accuracy % of non-private	Communication per query	Latency total
Tiptoe [HDCGZ23]	393	35%	20.26 MiB	3.65 s
Mímir I – lowest comm.	393	35% ≈	0.24 MiB ↓ ×84	4.23 s ↑ ×1.2
Mímir II – Tiptoe's comm.	4	63% ↑ ×1.8	20.34 MiB ≈	0.87 s ↓ ×4.2
PACMANN [ZSF25]	1	76%	61.85 MiB	0.80 s
Mímir III – highest accuracy	1	81% ↑ ×1.1	44.46 MiB ↓ ×1.4	1.66 s ↑ ×2.1
Mímir IV – fastest	10	60% ↑ ×1.7	8.13 MiB ↓ ×2.5	0.77 s ↓ ×4.7

l: embedding length | c: # clusters

Results - A Dynamic Database Setting

	c	Accuracy % of non-private	Communication per query	Latency total
Tiptoe [HDCGZ23]	393	35%	20.26 MiB	3.65 s
Mímir I – lowest comm.	393	35% ≈	0.24 MiB ↓ ×84	4.23 s ↑ ×1.2
Mímir II – Tiptoe's comm.	4	63% ↑ ×1.8	20.34 MiB ≈	0.87 s ↓ ×4.2
PACMANN [ZSF25]	1	76%	61.85 MiB	0.80 s
Mímir III – highest accuracy	1	81% ↑ ×1.1	44.46 MiB ↓ ×1.4	1.66 s ↑ ×2.1
Mímir IV – fastest	10	60% ↑ ×1.7	8.13 MiB ↓ ×2.5	0.77 s ↓ ×4.7

$\Longrightarrow$ Mímir trades communication cost for computation cost

Thank you!

l: embedding length | c: # clusters

Backup: Results in the Stateful Setting

Online latency assumes a static database, that allows a client to do most computations *offline* in a preprocessing step.
The *total* latency assumes the database changing frequently, forcing to do offline and online computations per query.

	c	MRR@100	Communication per query	Latency offline	Latency online	Latency total
Plaintext	1	0.352	12.26 MiB	–	–	–
Tiptoe [HDCGZ23]	393	0.125	20.26 MiB	3.57 s	0.08 s	3.65 s
Mímir I – lowest comm.	393	0.125 ≈	0.24 MiB ↓ ×84	–	4.23 s ↑ ×53	4.23 s ↑ ×1.2
Mímir II – Tiptoe's comm.	4	0.222 ↑ ×1.8	20.34 MiB ≈	–	0.87 s ↑ ×11	0.87 s ↓ ×4.2
PACMANN [ZSF25]	1	0.266	61.85 MiB	0.74 s	0.06 s	0.80 s
Mímir III – highest accuracy	1	0.286 ↑ ×1.1	44.46 MiB ↓ ×1.4	–	1.66 s ↑ ×28	1.66 s ↑ ×2.1
Mímir IV – fastest	10	0.211 ↑ ×1.7	8.13 MiB ↓ ×2.5	–	0.77 s ↑ ×9.7	0.77 s ↓ ×4.7

l: embedding length | c: # clusters

Backup: Parameters of Mímir I–IV

Parameter	Mímir I	Mímir II	Mímir III	Mímir IV
Security parameter, $n = N$	2048	2048	4096	2048
Standard deviation of the error, σ	3.2	3.2	3.2	10.4
Ciphertext modulus q	53 bits	53 bits	58 bits	55 bits
Security level, λ	128.8	128.8	248.5	128.1
Embedding vector length, l	192	192	192	192
Embedding precision ρ	5 bits	5 bits	5 bits	5 bits
Plaintext modulus, p	98305	98305	98305	98305
Number of CRT component rings, r	2	2	1	2
Component ring moduli, p_R	$\{314, 315\}$	$\{314, 315\}$	98305	$\{314, 315\}$
Number of clusters c	393	4	1	10
Decomposition base B_1	3	3	5	5
LWEtoRLWE decomposition degree d_1	34	34	25	24
Decomposition base B_2	17	179	3	116
LWE′toRGSW decomposition degree d_2	13	7	37	8
Decryption failure rate	$9.1 \cdot 10^{-13}$	$3.4 \cdot 10^{-15}$	$9.1 \cdot 10^{-13}$	$2.3 \cdot 10^{-13}$

All configurations: 3 213 835 documents (MS MARCO), $\lambda \geq 128$, decryption failure rate $\leq 2^{-40}$.