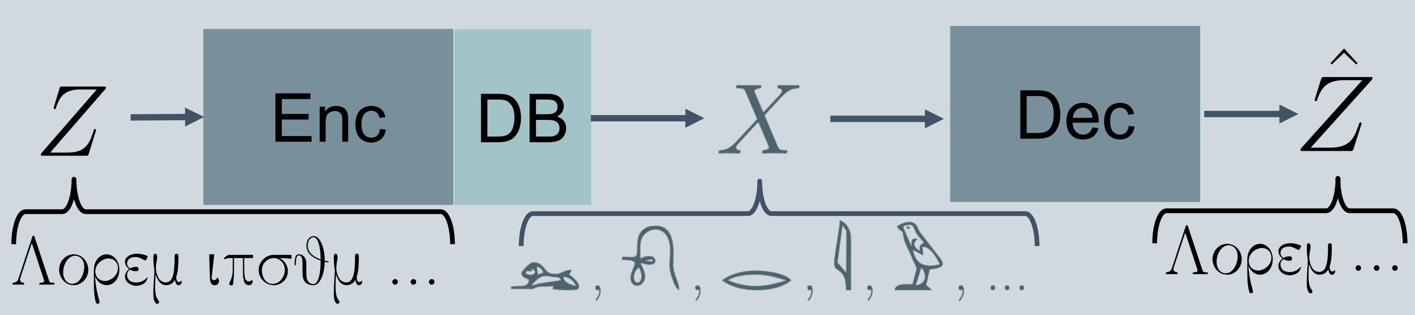
Symbolic Autoencoding for Self-Supervised Sequence Learning

Mohammad Hossein Amani, Nicolas Mario Baldwin*, Amin Mansouri*
Martin Josifoski, Maxime Peyrard, Robert West

Symbolic Auto Encoding (ΣAE)

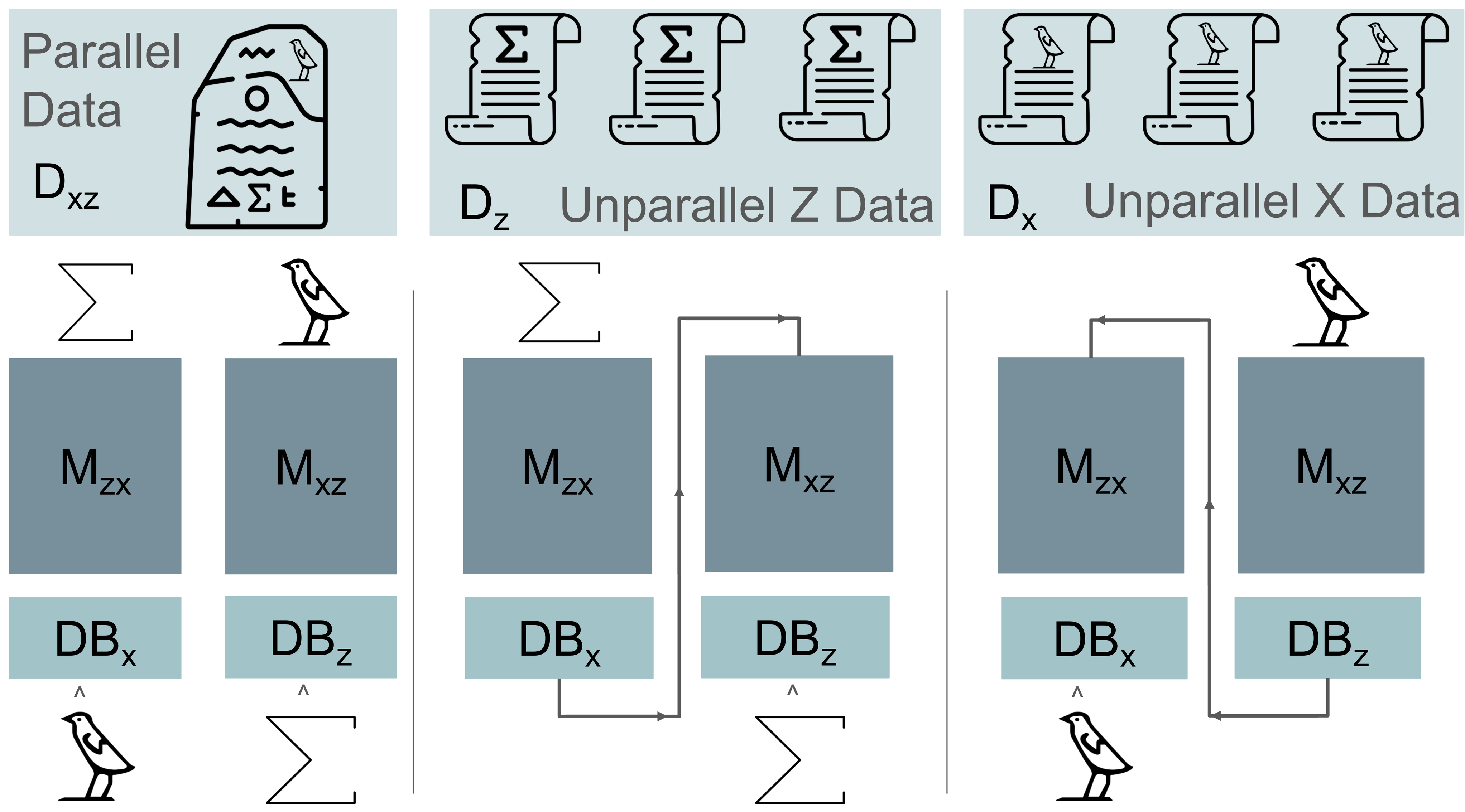
Why ΣAE?

- Humans **think**, **plan**, and **reason** using symbols.
- Symbolic representations capture efficient and concise information, enhancing model **sample efficiency**.

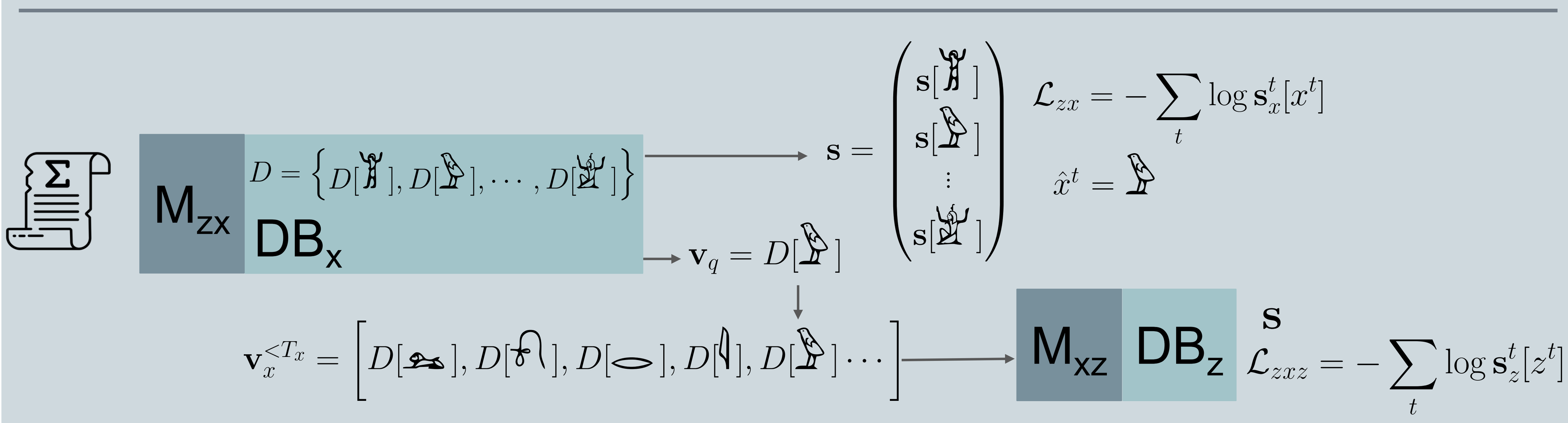


- **X:** Discrete sequential representation of input features
- **Enc, Dec:** The encoder and decoder models. Any sequence model, e.g., transformers trained on next token prediction or diffusion transformers, recurrent neural networks, etc.
- **DB:** Point of quantization in the encoder model, introducing non-differentiability.

Deciphering the Rosetta Stone – a weakly supervised task



Output embedding of a generative model $\mathbf{v} \rightarrow$ DB $\rightarrow$ $\mathbf{s}$ compute the NLL loss when label token exists; $\mathbf{v}_q$ serve as input for subsequent models or layers



$$\mathcal{L}_{zx} = -\sum_t \log \mathbf{s}_x^t[x^t]$$

$$\mathcal{L}_{zxz} = -\sum_t \log \mathbf{s}_z^t[z^t]$$

Straight through Gradients Substitutes

→ Approximate the gradient of the quantized vector with that of it's continuous relaxation: $\frac{\partial \mathbf{v}_q}{\partial \mathbf{v}} \approx \frac{\partial \mathbb{E}_{\mathbf{S}}[D]}{\partial \mathbf{v}}$

DB: $\mathbf{v}_q \leftarrow \mathbf{v}_q + \sum_{i=1}^{|V|} \mathbf{s}[i]D[i] - sg(\sum_{i=1}^{|V|} \mathbf{s}[i]D[i])$

Hidden sequence collapse and EOS Soft-Masking

DB:

$$\mathbf{v}_q^{<T_x} = [D[\cdot], D[\cdot], \cdots < eos >, D[\cdot], \cdots]$$

$$\mathbf{m}^{<T_x} = [\ 1\ ,\ 1\ ,\ \cdots\ ,\ 1\ ,\ 0\ ,\ \cdots]$$

$$\mathbf{v}_q^{<T_x} \leftarrow \mathbf{v}_q^{<T_x} \odot \mathbf{m}^{<T_x} + \left(D[< pad >] \odot (1-\mathbf{m}^{<T_x})\right)$$

→ **Problem.** The model never receives gradient feedback on the discrete decision of when to halt generation.
Solution. Use Gradient Approximation for Halting the Generation. While m is not differentiable, $\mathbb{E}[m]$ is:

$$\mathbb{E}[\mathbf{m}[i]] = \prod_{k=1}^{i-1} (1 - \mathbb{P}(O_k =< eos >))$$

$$\mathbf{m} \leftarrow \mathbf{m} + \mathbb{E}[\mathbf{m}] - sg(\mathbb{E}[\mathbf{m}])$$

— Experimental Results —

Dataset example pairs

Dataset	Sample
SCAN	**X:** look right thrice after run left **Z:** I_TURN_LEFT I_RUN I_TURN_RIGHT I_LOOK I_TURN_RIGHT I_LOOK I_TURN_RIGHT I_LOOK
PCFG SET	**X:** echo append append E18 C13 , L18 M17 , R1 L1 Y1 T18 J18 **Z:** E18 C13 L18 M17 R1 L1 Y1 T18 J18 J18
CFQ	**X:** Who influenced M1 's cinematographer , writer , and editor **Z:** SELECT DISTINCT ?x0 WHERE ?x0 a ns:people.person. ?x0 ns:influence.influence_node.influenced ?x1. ?x1 ns:film.cinematographer.film M1. ?x1 ns:film.editor.film M1. ?x1 ns:film.writer.film M1.
COGS	**X:** Olivia rolled Liam. **Z:** roll . agent (x_1 , Olivia) AND roll . theme (x_1 , Liam)

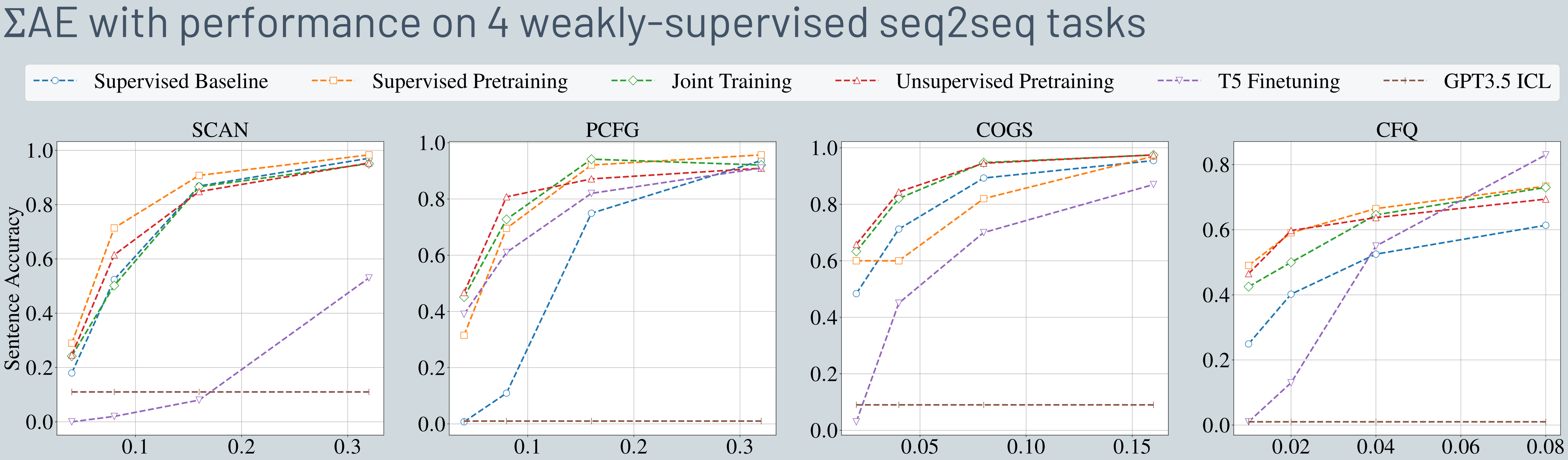
- Datasets chosen for their compositional complexity, controlled environments, and precise accuracy measures.
- We measure **token-level** and **sentence-level accuracy** on the Z space (the longer sequence).
- In **Unsupervised Compression Experiments** we demonstrate the feasibility of symbolic autoencoding with straight-through gradient updates.
- In **Weakly Supervised Experiments** we study the efficiency of ΣAE in utilizing small amounts of parallel data and a large unparallel corpus in a Rosetta Stone setting.

ΣAE sentence accuracy in unsupervised compression task

	SCAN	PCFG	COGS	CFQ
Softmax DB	1.00 0.96	0.74 0.31	0.98 0.55	0.99 0.69
Gumbel DB	0.98 0.74	0.75 0.36	0.98 0.51	0.99 0.43
VQ DB	1.00 0.93	0.44 0.00	0.94 0.03	0.90 0.00

- Trains the encoder to compress the longer sequence to a short code and the decoder to reconstruct the sequence.
- The shorter code has same meta features (vocabulary size and maximum length) as the ground truth shorter sequence.

ΣAE with performance on 4 weakly-supervised seq2seq tasks



- Results for **Softmax DB**: In the forward pass, we select the most likely token, and in the backward pass, we differentiate through a Softmax average of dictionary embeddings.
- Three Baselines: **Finetuning T5-Large, in-context learning with GPT-3.5**, and **supervised training from scratch** using only the available parallel data.

Paper?

Code & Data?

EPFL
ICML International Conference On Machine Learning
Differentiable Almost Everything Workshop