CS 957, System-2 AI
Large Concept Models

Mohammad Hossein Rohban | June. 2025

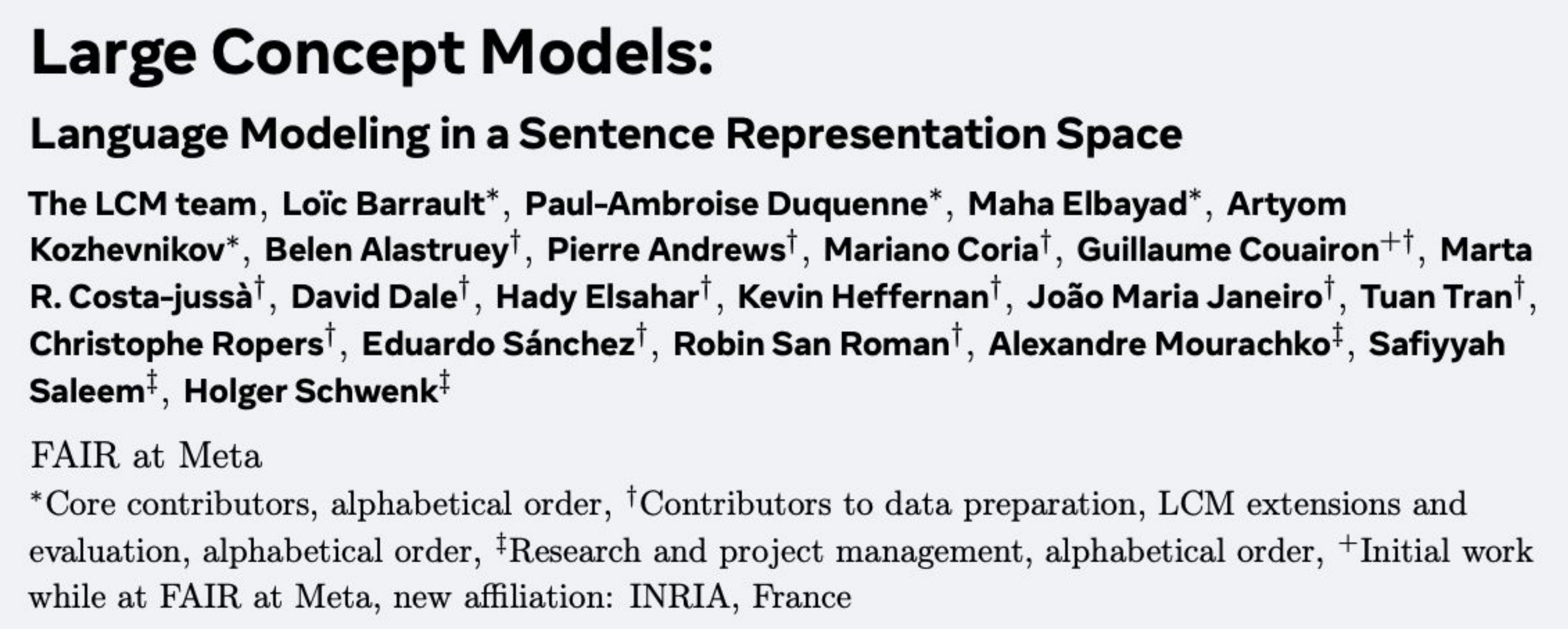
Sharif University of Technology

Motivation

- Explicit reasoning and planning at **multiple levels of abstraction**
- Researchers do not usually prepare detailed speeches by **writing out every single word** they will pronounce
- If probability of correct token generation is p, for a sequence of length L, the probability of generating a correct sentence would be roughly p^L

Let’s brainstorm!

How to go about this?



Large Concept Models:

Language Modeling in a Sentence Representation Space

The LCM team, Loïc Barrault[*], Paul-Ambroise Duquenne[*], Maha Elbayad[*], Artyom Kozhevnikov[*], Belen Alastruey[†], Pierre Andrews[†], Mariano Coria[†], Guillaume Couairon[+†], Marta R. Costa-jussà[†], David Dale[†], Hady Elsahar[†], Kevin Heffernan[†], João Maria Janeiro[†], Tuan Tran[†], Christophe Ropers[†], Eduardo Sánchez[†], Robin San Roman[†], Alexandre Mourachko[‡], Safiyyah Saleem[‡], Holger Schwenk[‡]

FAIR at Meta

[*] Core contributors, alphabetical order, [†] Contributors to data preparation, LCM extensions and evaluation, alphabetical order, [‡] Research and project management, alphabetical order, [+] Initial work while at FAIR at Meta, new affiliation: INRIA, France

Overview

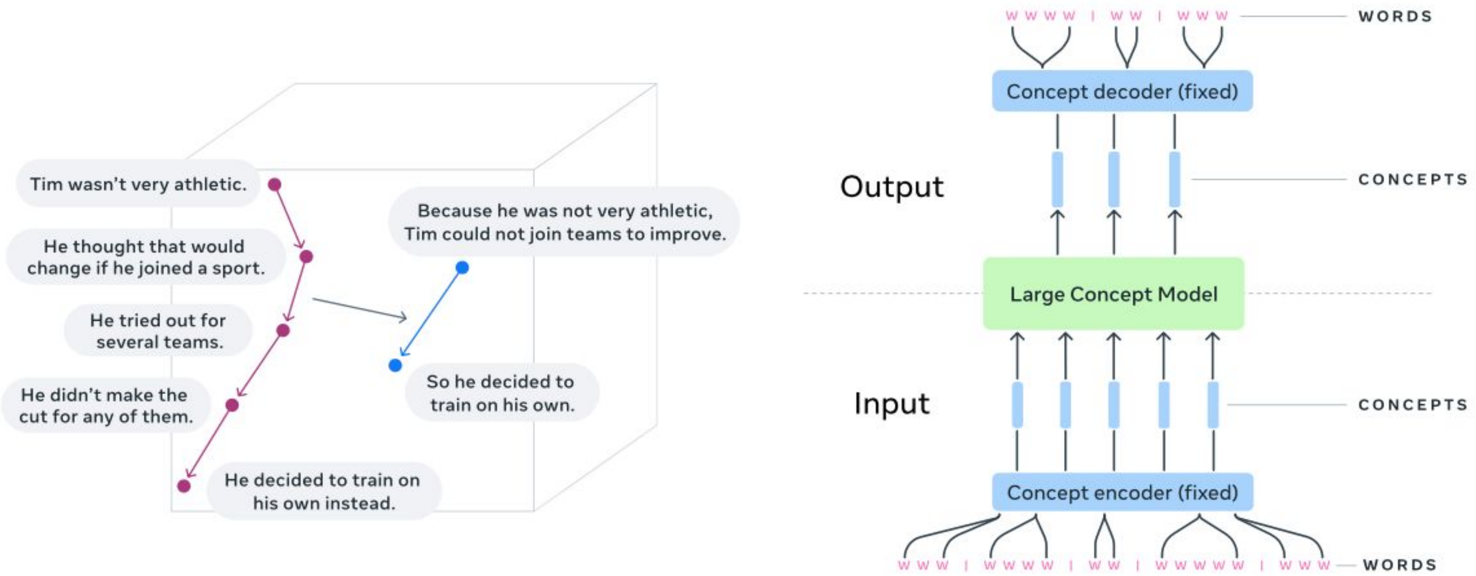
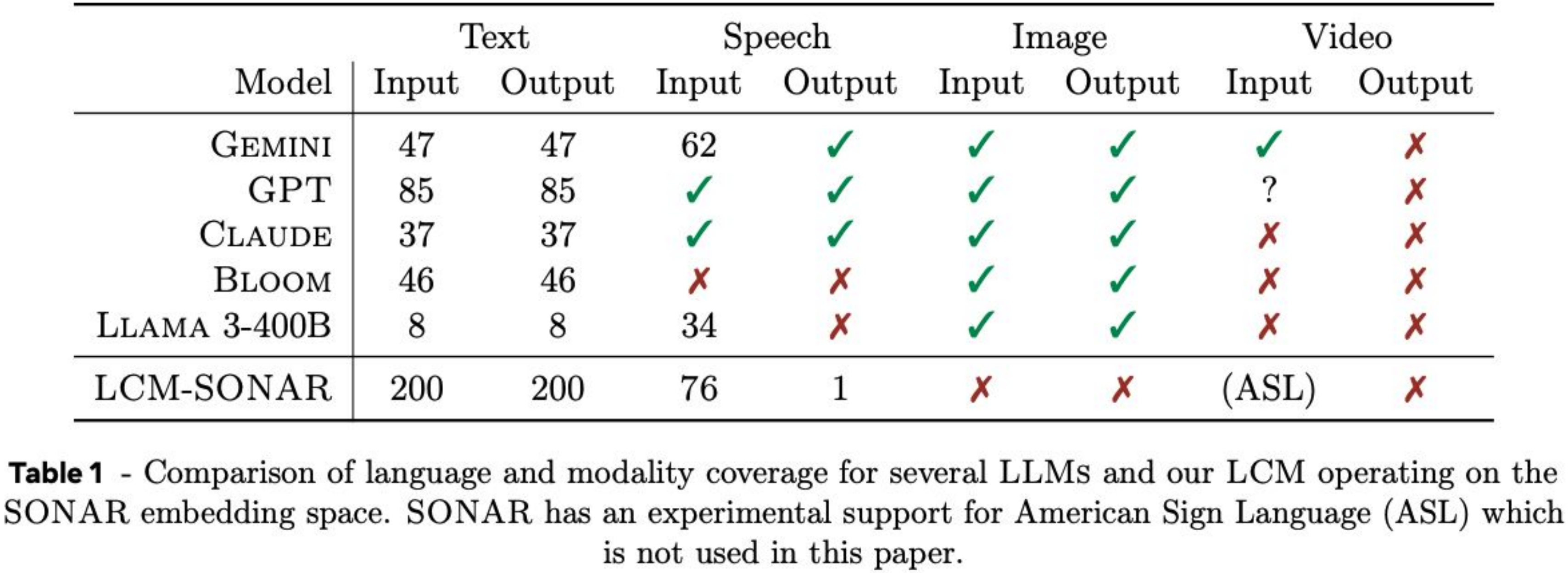

Tim wasn't very athletic.
He thought that would change if he joined a sport.
He tried out for several teams.
He didn't make the cut for any of them.
He decided to train on his own instead.
Because he was not very athletic, Tim could not join teams to improve.
So he decided to train on his own.
w w w w | w w | w w w
WORDS
Concept decoder (fixed)
Output
CONCEPTS
Large Concept Model
Input
CONCEPTS
Concept encoder (fixed)
w w w | w w w w | w w | w w w w w | w w w
WORDS

Figure 1 - Left: visualization of reasoning in an embedding space of concepts (task of summarization). Right: fundamental architecture of an LARGE CONCEPT MODEL (LCM). ⋆: concept encoder and decoder are frozen.

Benefits

- Shorter sequence
  - Less computational overhead
- Language agnostic
- Modality Agnostic
  - Avoiding modality competition

Multi-lingual and Multi-Modal



	Text		Speech		Image		Video	
Model	Input	Output	Input	Output	Input	Output	Input	Output
Gemini	47	47	62	✓	✓	✓	✓	✗
GPT	85	85	✓	✓	✓	✓	?	✗
Claude	37	37	✓	✓	✓	✓	✗	✗
Bloom	46	46	✗	✗	✓	✓	✗	✗
Llama 3-400B	8	8	34	✗	✓	✓	✗	✗
LCM-SONAR	200	200	76	1	✗	✗	(ASL)	✗

Table 1 - Comparison of language and modality coverage for several LLMs and our LCM operating on the SONAR embedding space. SONAR has an experimental support for American Sign Language (ASL) which is not used in this paper.

SONAR Encoding

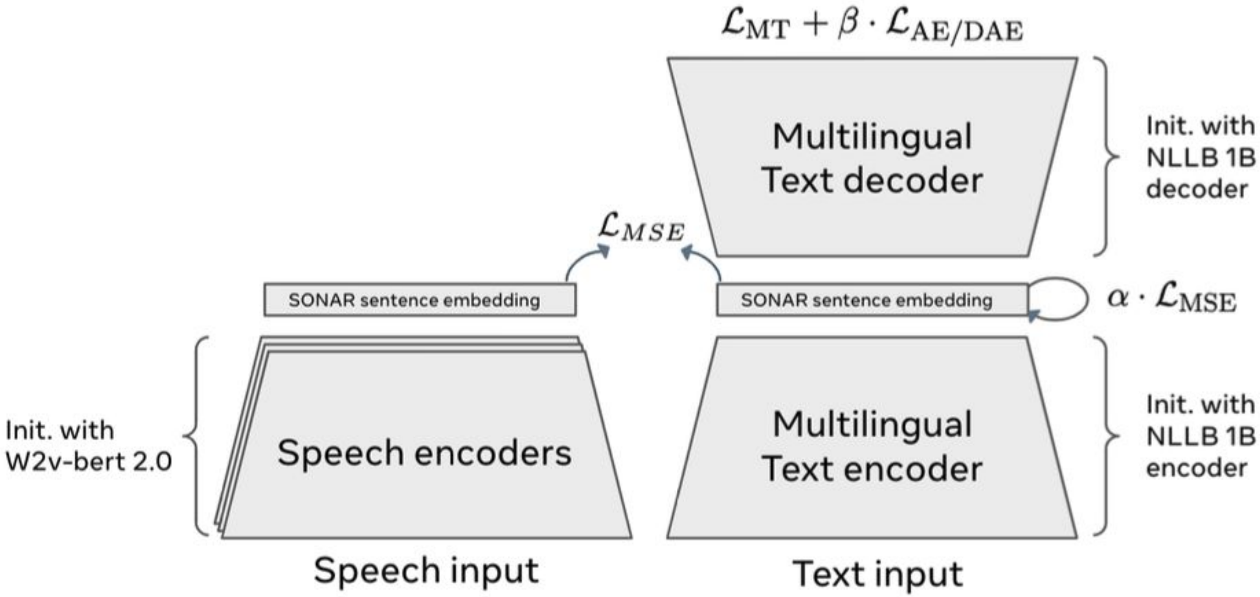

$\mathcal{L}_{\mathrm{MT}} + \beta \cdot \mathcal{L}_{\mathrm{AE/DAE}}$
Multilingual Text decoder
Init. with NLLB 1B decoder
$\mathcal{L}_{MSE}$
SONAR sentence embedding
SONAR sentence embedding
$\alpha \cdot \mathcal{L}_{\mathrm{MSE}}$
Init. with W2v-bert 2.0
Speech encoders
Multilingual Text encoder
Init. with NLLB 1B encoder
Speech input
Text input

Figure 2 - Encoder/decoder bottleneck architecture to train the SONAR text embeddings (right part of figure). Teacher-student approach to extend SONAR to the speech modality (left part).

Sentence Segmentation

1. SpaCy segmenter (SPACY): Rule-based; relies on punctuation and capitalization; suitable for high-resource languages
2. Segment any Text (SAT): predict sentence boundaries at the token level; less reliant on punctuation and capitalization
3. Force a length cap on the sentences to make them short.

How to evaluate these segmentors?

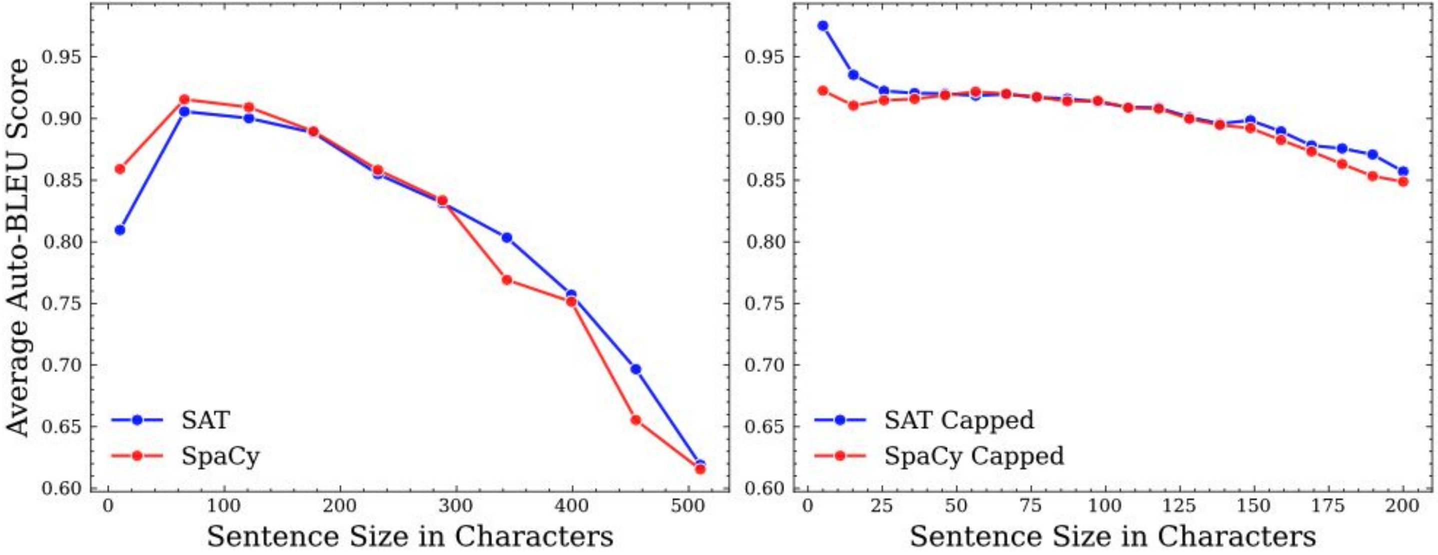


Figure 3 - **Segmenters quality.** Average Auto-BLEU scores for different sentence segmentation methods depending on sentence length, for both out of the box (left) and capped implementations (right).

Base LCM

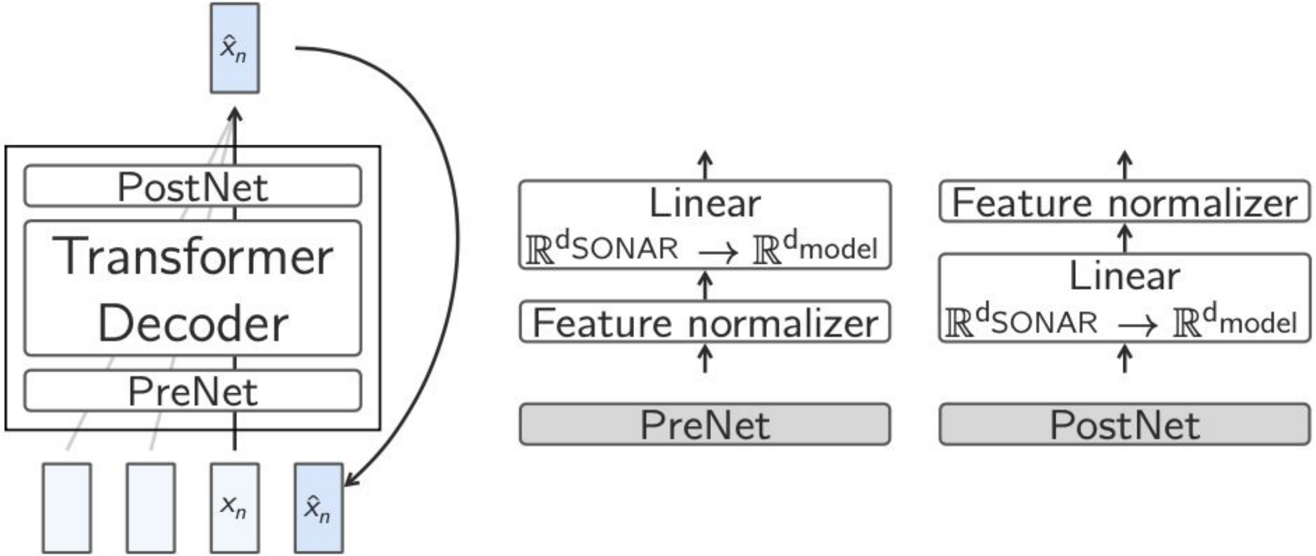
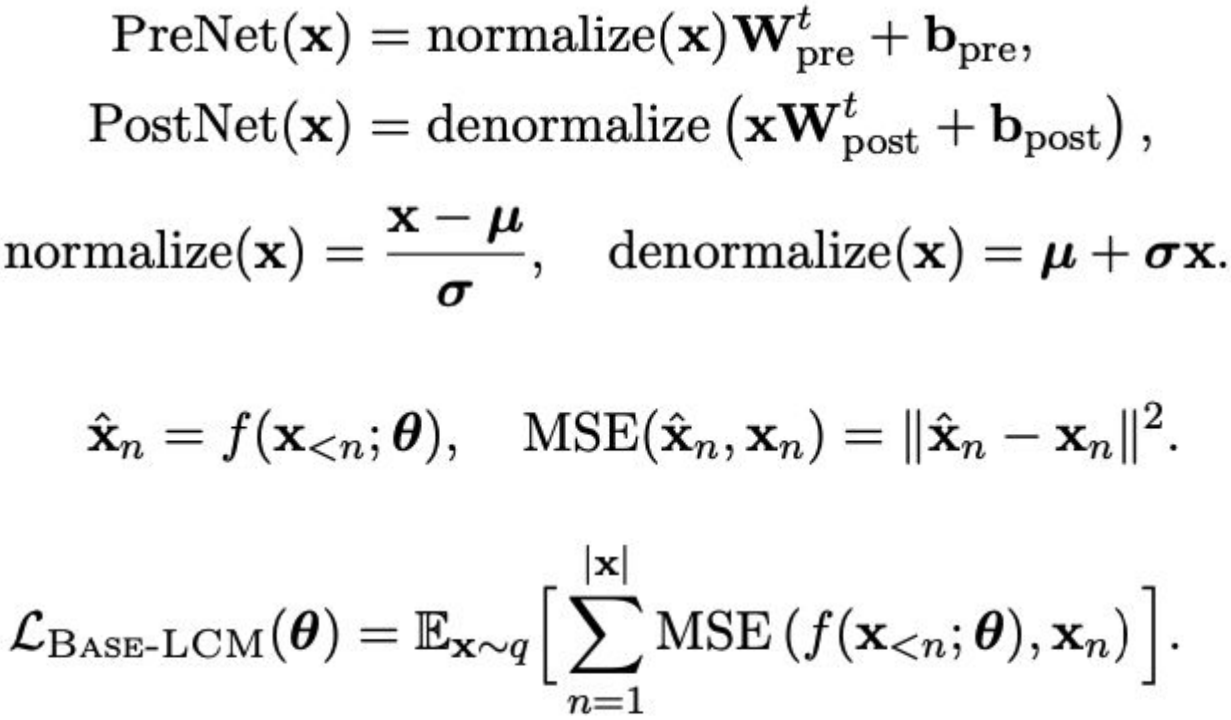


Figure 4 - **TheBase-LCM.** Illustration of the BASE-LCM. At its core is a standard decoder-only Transformer surrounded with a PreNet and a PostNet.

Details



$$\mathrm{PreNet}(\mathbf{x}) = \mathrm{normalize}(\mathbf{x})\mathbf{W}^{t}_{\mathrm{pre}} + \mathbf{b}_{\mathrm{pre}},$$

$$\mathrm{PostNet}(\mathbf{x}) = \mathrm{denormalize}\left(\mathbf{x}\mathbf{W}^{t}_{\mathrm{post}} + \mathbf{b}_{\mathrm{post}}\right),$$

$$\mathrm{normalize}(\mathbf{x}) = \frac{\mathbf{x} - \boldsymbol{\mu}}{\boldsymbol{\sigma}}, \quad \mathrm{denormalize}(\mathbf{x}) = \boldsymbol{\mu} + \boldsymbol{\sigma}\mathbf{x}.$$

$$\hat{\mathbf{x}}_n = f(\mathbf{x}_{<n}; \boldsymbol{\theta}), \quad \mathrm{MSE}(\hat{\mathbf{x}}_n, \mathbf{x}_n) = \|\hat{\mathbf{x}}_n - \mathbf{x}_n\|^2.$$

$$\mathcal{L}_{\textsc{Base-LCM}}(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{x}\sim q}\Big[\sum_{n=1}^{|\mathbf{x}|} \mathrm{MSE}\left(f(\mathbf{x}_{<n}; \boldsymbol{\theta}), \mathbf{x}_n\right)\Big].$$

What are the potential problems with base LCM?

Let’s brainstorm

Diffusion-based LCM

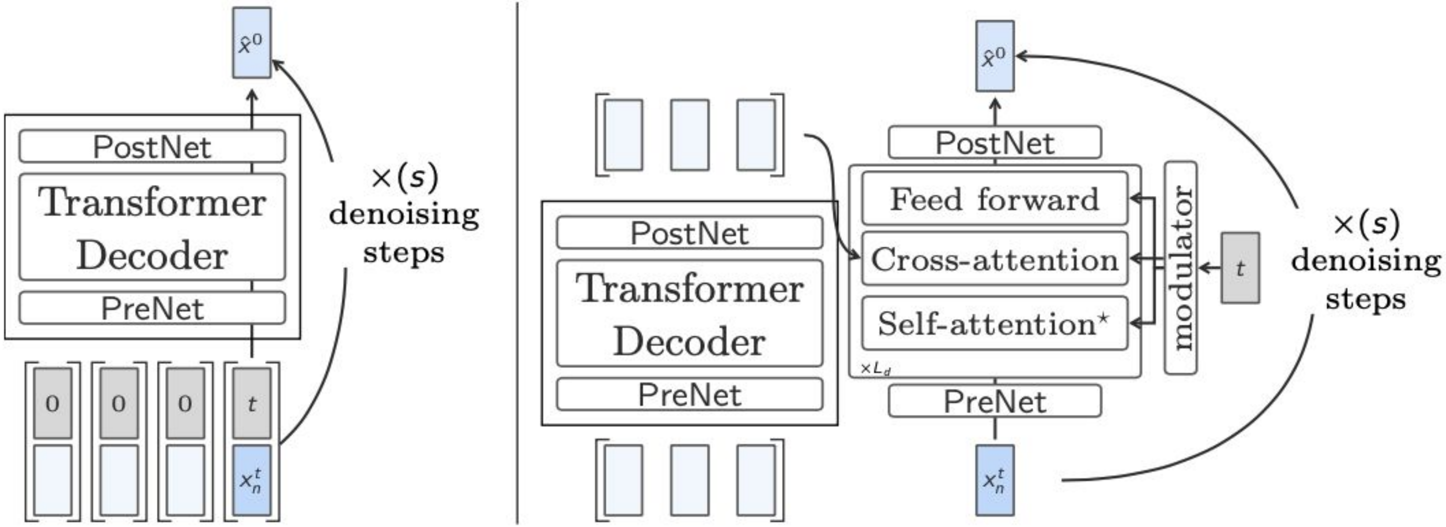


Figure 6 - Inference with diffusion-based LCMs. In the left-hand side, an illustration of the ONE-TOWER LCM and on the right-hand side an illustration of the TWO-TOWER LCM.

Details

$$\mathrm{q}(\mathbf{x}^t|\mathbf{x}^0) := \mathcal{N}(\alpha_t \mathbf{x}^0, \sigma_t^2 \mathbf{I}).$$

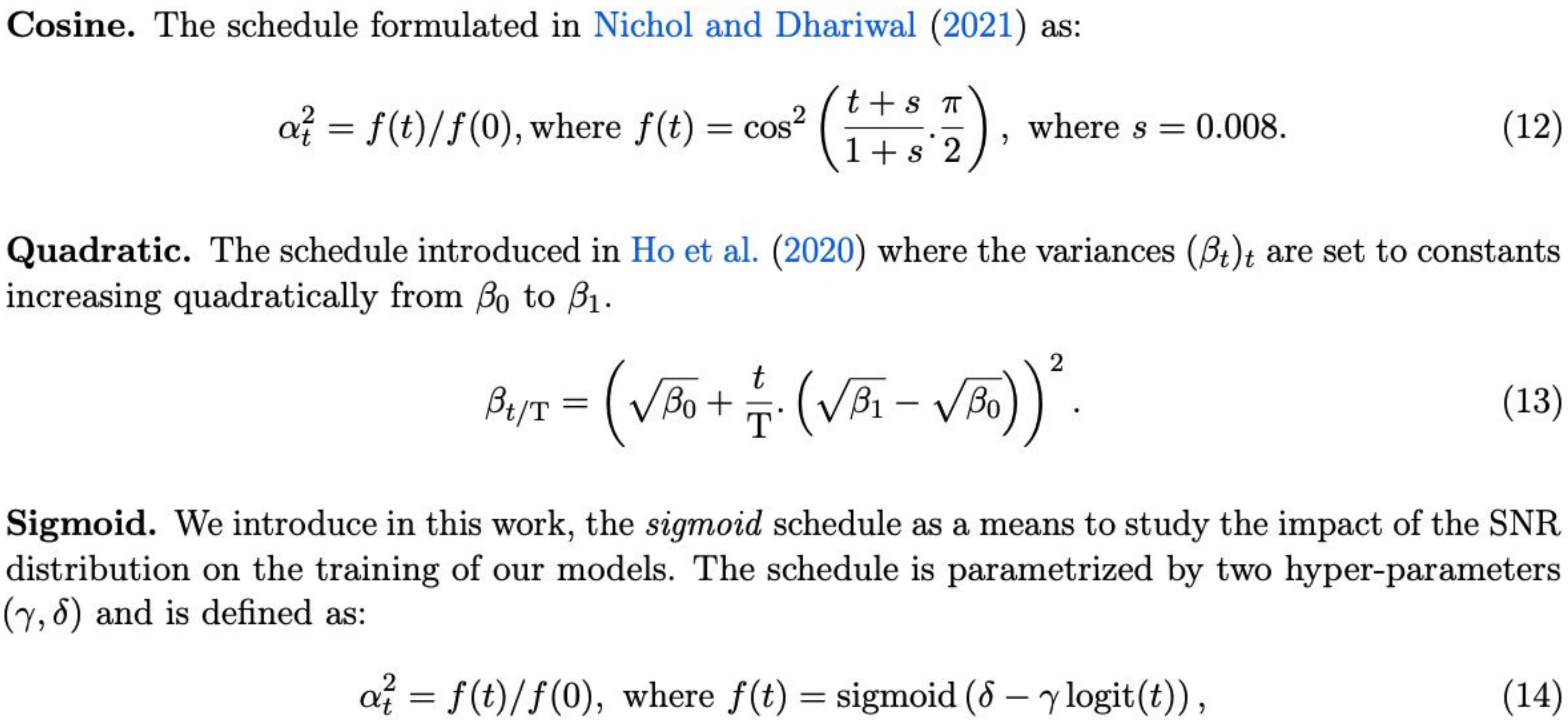
$$\mathbf{x}^t = \alpha_t \mathbf{x}^0 + \sigma_t \boldsymbol{\epsilon} \quad \text{where } \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

$$\alpha_t^2 = \mathrm{sigmoid}(\lambda_t), \qquad \sigma_t^2 = \mathrm{sigmoid}(-\lambda_t) = 1 - \mathrm{sigmoid}(\lambda_t), \qquad \lambda_t = \log\left(\alpha_t^2 / \sigma_t^2\right),$$

$$\mathrm{q}(\mathbf{x}^{1\ldots\mathrm{T}}|\mathbf{x}^0) := \prod_{t=1}^{\mathrm{T}} \mathrm{q}(\mathbf{x}^t|\mathbf{x}^{t-1}), \quad \mathrm{q}(\mathbf{x}^t|\mathbf{x}^{t-1}) := \mathcal{N}(\mathbf{x}^t; \sqrt{1-\beta_t}\mathbf{x}^{t-1}, \beta_t \mathbf{I}),$$

$$\alpha_t^2 = \prod_{s=1}^{t} (1 - \beta_s).$$

Details



Cosine. The schedule formulated in Nichol and Dhariwal (2021) as:

$$\alpha_t^2 = f(t)/f(0), \text{where } f(t) = \cos^2\left(\frac{t+s}{1+s}.\frac{\pi}{2}\right), \quad \text{where } s = 0.008. \tag{12}$$

Quadratic. The schedule introduced in Ho et al. (2020) where the variances $(\beta_t)_t$ are set to constants increasing quadratically from β_0 to β_1.

$$\beta_{t/\mathrm{T}} = \left(\sqrt{\beta_0} + \frac{t}{\mathrm{T}}.\left(\sqrt{\beta_1} - \sqrt{\beta_0}\right)\right)^2. \tag{13}$$

Sigmoid. We introduce in this work, the *sigmoid* schedule as a means to study the impact of the SNR distribution on the training of our models. The schedule is parametrized by two hyper-parameters (γ, δ) and is defined as:

$$\alpha_t^2 = f(t)/f(0), \;\text{where } f(t) = \text{sigmoid}\left(\delta - \gamma\,\text{logit}(t)\right), \tag{14}$$

Details

$$\mathrm{p}_{\boldsymbol{\theta}}(\mathbf{x}^{0:\mathrm{T}}) \coloneqq \mathrm{p}(\mathbf{x}^{\mathrm{T}}) \prod_{t=1}^{\mathrm{T}} \mathrm{p}_{\boldsymbol{\theta}}(\mathbf{x}^{t-1}|\mathbf{x}^{t}), \qquad \mathrm{p}_{\boldsymbol{\theta}}(\mathbf{x}^{t-1}|\mathbf{x}^{t}) \coloneqq \mathcal{N}(\mathbf{x}^{t-1}; \boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{x}^{t}, t), \boldsymbol{\Sigma}_{\boldsymbol{\theta}}(\mathbf{x}^{t}, t)),$$

$$\mathcal{L}(\boldsymbol{\theta}) \coloneqq \mathbb{E}_{t\sim\mathcal{U}(0,1)}\left[\omega(t)\mathcal{L}(t,\boldsymbol{\theta})\right], \quad \mathcal{L}(t,\boldsymbol{\theta}) \coloneqq \mathbb{E}_{\mathbf{x}^0,\boldsymbol{\epsilon}}\left[\left\|\mathbf{x}^0 - \boldsymbol{\mu}_{\boldsymbol{\theta}}(\alpha_t\mathbf{x}^0 + \sigma_t\boldsymbol{\epsilon}, t)\right\|_2^2\right].$$

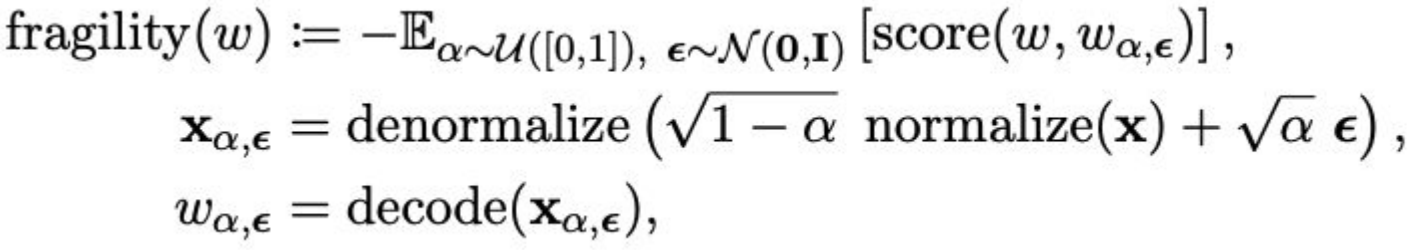
$$\omega(t) = \max(\min(\exp(\lambda_t), \lambda_{\max}), \lambda_{\min}), \; \lambda_t = \log(\alpha_t^2/\sigma_t^2),$$

$$\nabla_x \log_\gamma p(x|y) = (1-\gamma)\nabla_x \log p(x) + \gamma\nabla_x \log p(x|y),$$

$$\mathcal{L}_{\text{fragility}}(\boldsymbol{\theta}) \coloneqq \mathbb{E}_{t\sim\mathcal{U}(0,1),\mathbf{x}^0,\boldsymbol{\epsilon}}\left[\omega(\mathbf{x}^0)\ \left\|\mathbf{x}^0 - \boldsymbol{\mu}_{\boldsymbol{\theta}}(\alpha_t\mathbf{x}^0 + \sigma_t\boldsymbol{\epsilon}, t)\right\|_2^2\right],$$

$$\omega(\mathbf{x}^0) = \text{sigmoid}(a\ \ \text{fragility}(\mathbf{x}^0) + b),$$

Fragility



$$
\begin{aligned}
\text{fragility}(w) &:= -\mathbb{E}_{\alpha \sim \mathcal{U}([0,1]),\ \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0},\mathbf{I})} \left[\text{score}(w, w_{\alpha,\boldsymbol{\epsilon}})\right], \\
\mathbf{x}_{\alpha,\boldsymbol{\epsilon}} &= \text{denormalize}\left(\sqrt{1-\alpha}\ \ \text{normalize}(\mathbf{x}) + \sqrt{\alpha}\ \boldsymbol{\epsilon}\right), \\
w_{\alpha,\boldsymbol{\epsilon}} &= \text{decode}(\mathbf{x}_{\alpha,\boldsymbol{\epsilon}}),
\end{aligned}
$$

Training One-Tower LCM

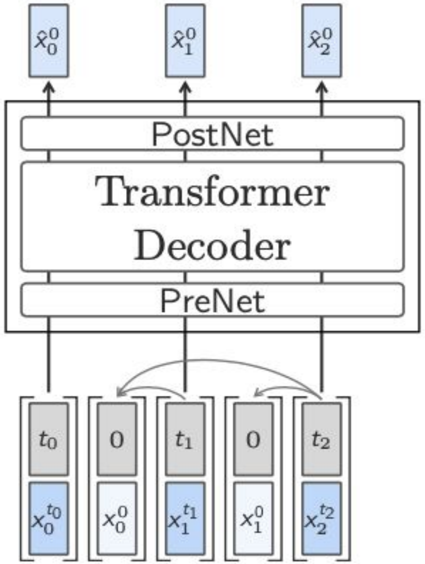
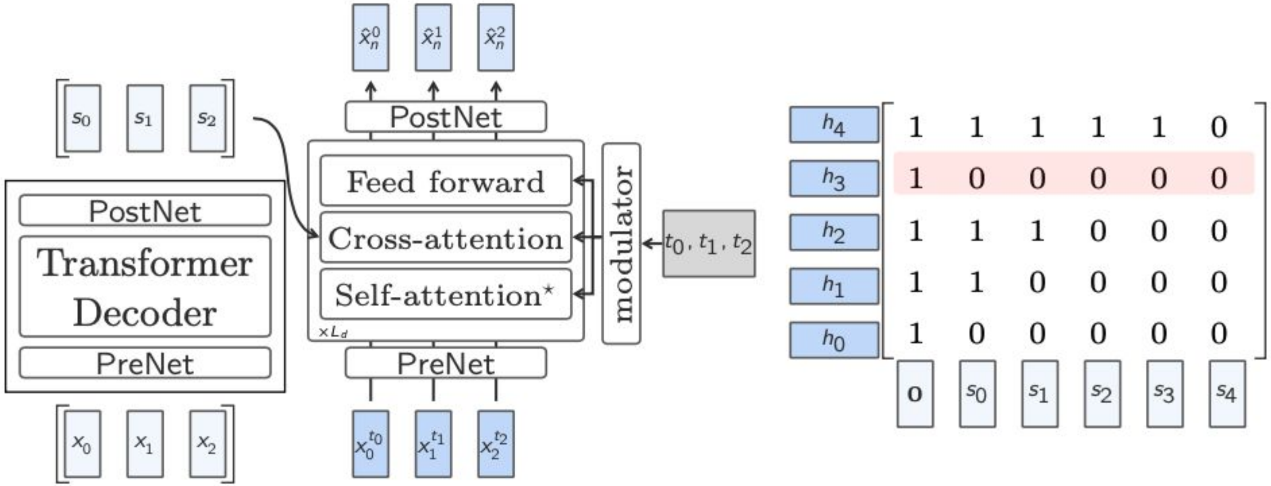


Figure 7 - Training of One-Tower diffusion LCM. Interleaving the clean and noisy embeddings and sampling different diffusion timesteps allows for efficient training.

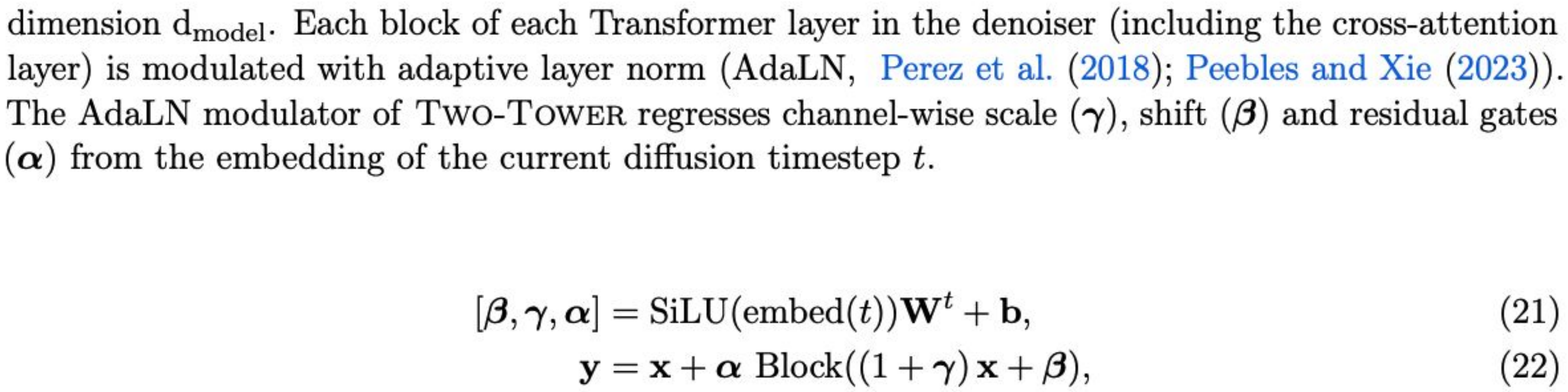
Training Two-Tower LCM



$$
\begin{array}{c|cccccc}
h_4 & 1 & 1 & 1 & 1 & 1 & 0 \\
h_3 & 1 & 0 & 0 & 0 & 0 & 0 \\
h_2 & 1 & 1 & 1 & 0 & 0 & 0 \\
h_1 & 1 & 1 & 0 & 0 & 0 & 0 \\
h_0 & 1 & 0 & 0 & 0 & 0 & 0 \\
\hline
 & \mathbf{0} & s_0 & s_1 & s_2 & s_3 & s_4
\end{array}
$$

Figure 8 - Training Two-Tower diffusion LCM. On the left panel, a TWO-TOWER forward pass in training time in order to denoise multiple embeddings in parallel. On the right side panel a visualization of the denoiser's cross-attention masks with the red highlighted row signaling a sample dropped to train the denoiser unconditionally. $(h_1, \ldots, h_4)$ denotes the sequence of intermediate representations in the denoiser right before the cross-attention layer.

Training Two-Tower LCM



dimension $\mathrm{d_{model}}$. Each block of each Transformer layer in the denoiser (including the cross-attention layer) is modulated with adaptive layer norm (AdaLN, Perez et al. (2018); Peebles and Xie (2023)). The AdaLN modulator of Two-Tower regresses channel-wise scale ($\boldsymbol{\gamma}$), shift ($\boldsymbol{\beta}$) and residual gates ($\boldsymbol{\alpha}$) from the embedding of the current diffusion timestep t.

$$[\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\alpha}] = \mathrm{SiLU}(\mathrm{embed}(t))\mathbf{W}^t + \mathbf{b}, \tag{21}$$

$$\mathbf{y} = \mathbf{x} + \boldsymbol{\alpha}\ \mathrm{Block}((1 + \boldsymbol{\gamma})\,\mathbf{x} + \boldsymbol{\beta}), \tag{22}$$

Efficiency of LCMs

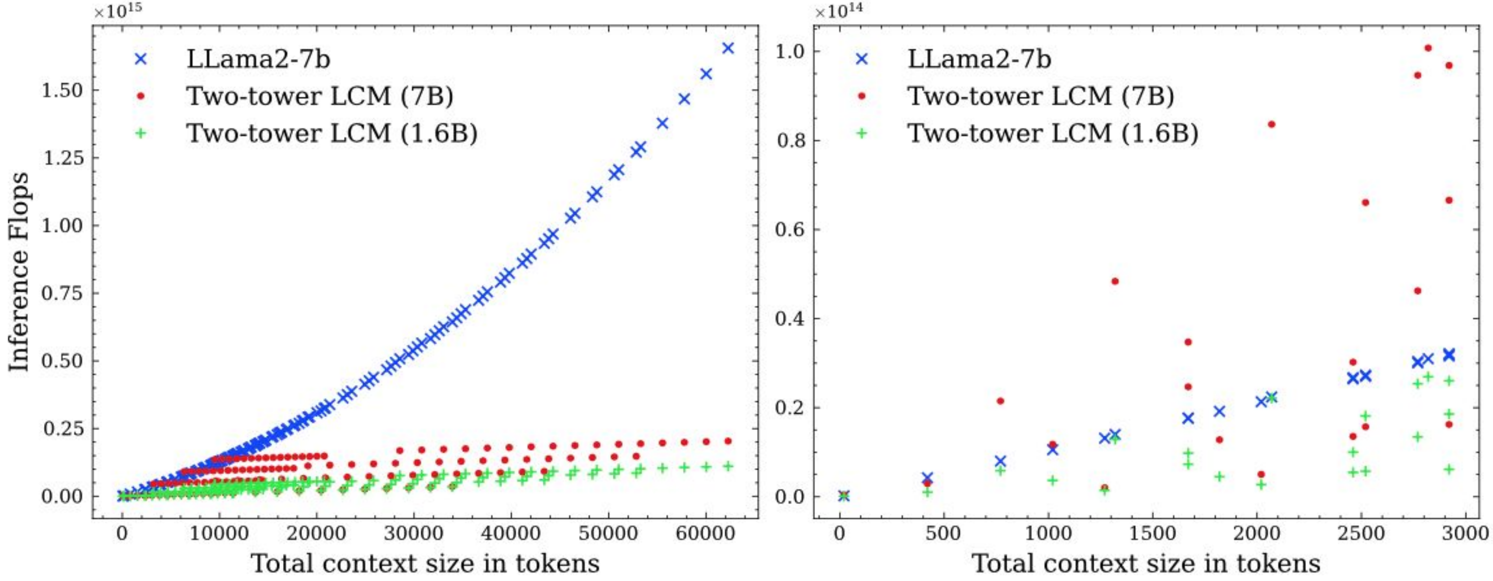
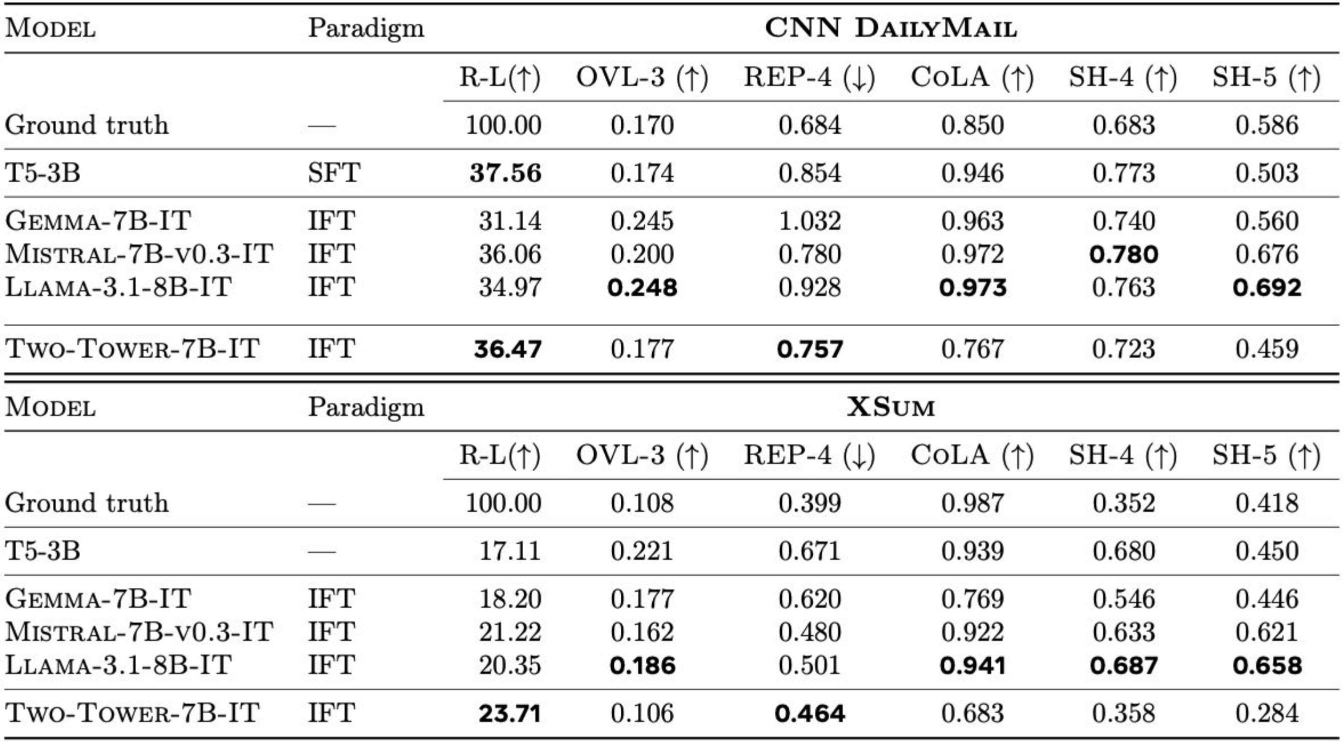


Figure 13 - Theoretical inference Flops of LCMs and LLLms. We evaluate the inference flops for different text lengths (in LLAMA2 tokens) with a variable average sentence length. Only extremely short sentences (≤ 10 tokens) favor LLMs.

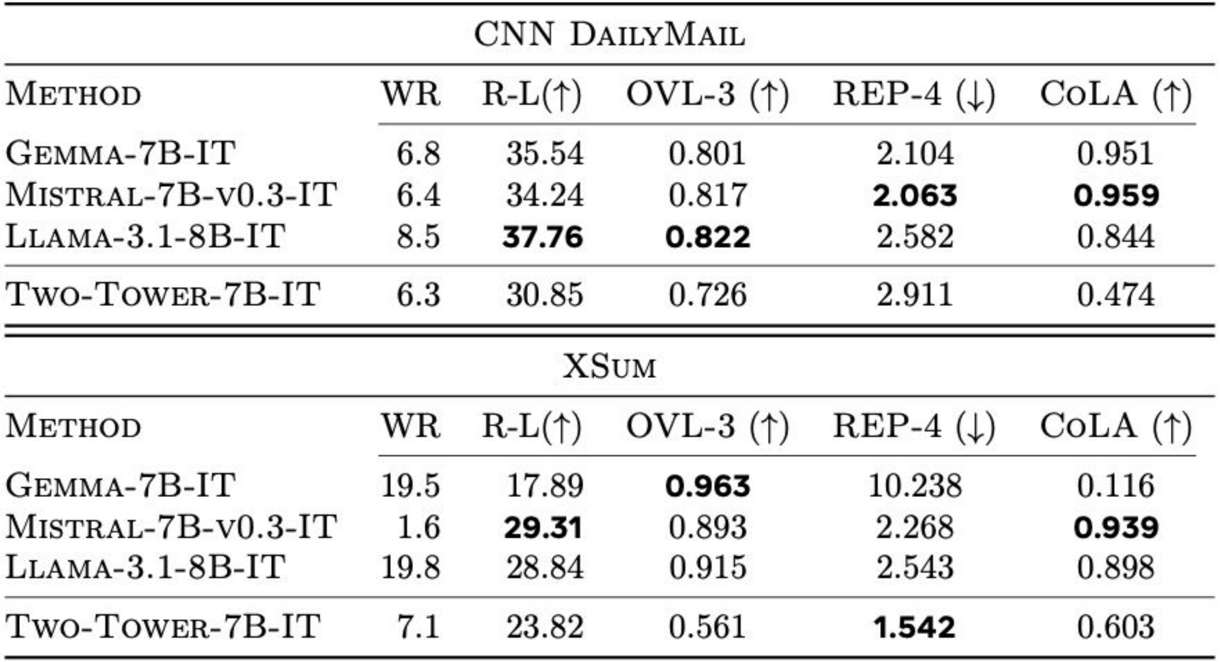
Results



Model	Paradigm	CNN DailyMail					
		R-L(↑)	OVL-3 (↑)	REP-4 (↓)	CoLA (↑)	SH-4 (↑)	SH-5 (↑)
Ground truth	—	100.00	0.170	0.684	0.850	0.683	0.586
T5-3B	SFT	**37.56**	0.174	0.854	0.946	0.773	0.503
Gemma-7B-IT	IFT	31.14	0.245	1.032	0.963	0.740	0.560
Mistral-7B-v0.3-IT	IFT	36.06	0.200	0.780	0.972	**0.780**	0.676
Llama-3.1-8B-IT	IFT	34.97	**0.248**	0.928	**0.973**	0.763	**0.692**
Two-Tower-7B-IT	IFT	**36.47**	0.177	**0.757**	0.767	0.723	0.459
Model	Paradigm	XSum					
		R-L(↑)	OVL-3 (↑)	REP-4 (↓)	CoLA (↑)	SH-4 (↑)	SH-5 (↑)
Ground truth	—	100.00	0.108	0.399	0.987	0.352	0.418
T5-3B	—	17.11	0.221	0.671	0.939	0.680	0.450
Gemma-7B-IT	IFT	18.20	0.177	0.620	0.769	0.546	0.446
Mistral-7B-v0.3-IT	IFT	21.22	0.162	0.480	0.922	0.633	0.621
Llama-3.1-8B-IT	IFT	20.35	**0.186**	0.501	**0.941**	**0.687**	**0.658**
Two-Tower-7B-IT	IFT	**23.71**	0.106	**0.464**	0.683	0.358	0.284

Table 10 - Performance on the CNN DailyMail and XSum summarization tasks.

Results



CNN DailyMail

Method	WR	R-L(↑)	OVL-3 (↑)	REP-4 (↓)	CoLA (↑)
Gemma-7B-IT	6.8	35.54	0.801	2.104	0.951
Mistral-7B-v0.3-IT	6.4	34.24	0.817	**2.063**	**0.959**
Llama-3.1-8B-IT	8.5	**37.76**	**0.822**	2.582	0.844
Two-Tower-7B-IT	6.3	30.85	0.726	2.911	0.474

XSum

Method	WR	R-L(↑)	OVL-3 (↑)	REP-4 (↓)	CoLA (↑)
Gemma-7B-IT	19.5	17.89	**0.963**	10.238	0.116
Mistral-7B-v0.3-IT	1.6	**29.31**	0.893	2.268	**0.939**
Llama-3.1-8B-IT	19.8	28.84	0.915	2.543	0.898
Two-Tower-7B-IT	7.1	23.82	0.561	**1.542**	0.603

Table 12 - Performance on the summary expansion tasks of CNN DailyMail and XSum, evaluated with the metrics described in Table 8. WR is the word count ratio between the hypothesis and the source summary.

Results

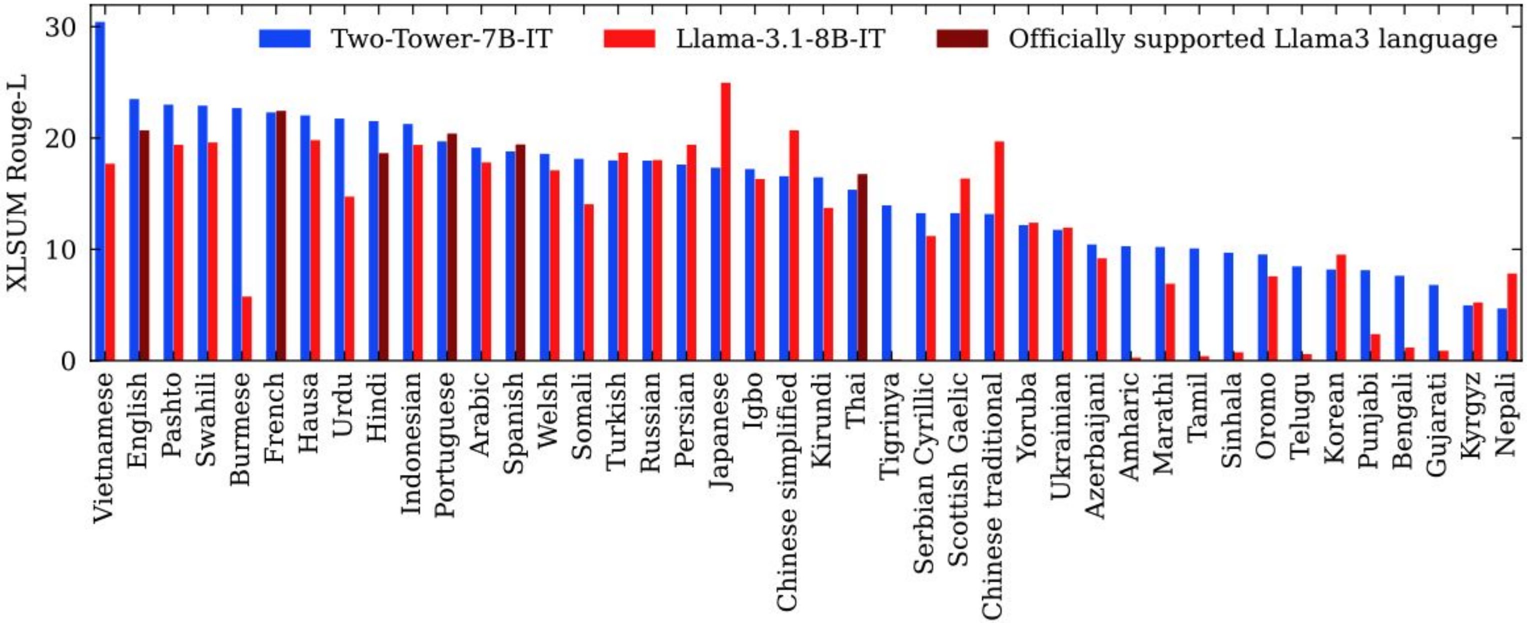


Figure 16 - ROUGE-L scores on XLSUM for LLAMA-3.1-8B-IT and TWO-TOWER-7B-IT.