Explicit Bayesian Inference to Uncover the Latent Themes of Large Language Models

Raymond Li[1] , Chuyuan Li[1*] , Gabriel Murray[2], Giuseppe Carenini[1]

[1] University of British Columbia, Vancouver, BC, Canada

[2] University of Fraser Valley, Abbotsford, BC, Canada

1. Contributions

We introduce a novel approach to interpret LLM generation process through the lens of an explicit Bayesian framework by inferring latent topic variables via variational inference.

***Intrinsic evaluation** of the topic quality shows that the inferred latent topic outperformed SOTA topic models.*

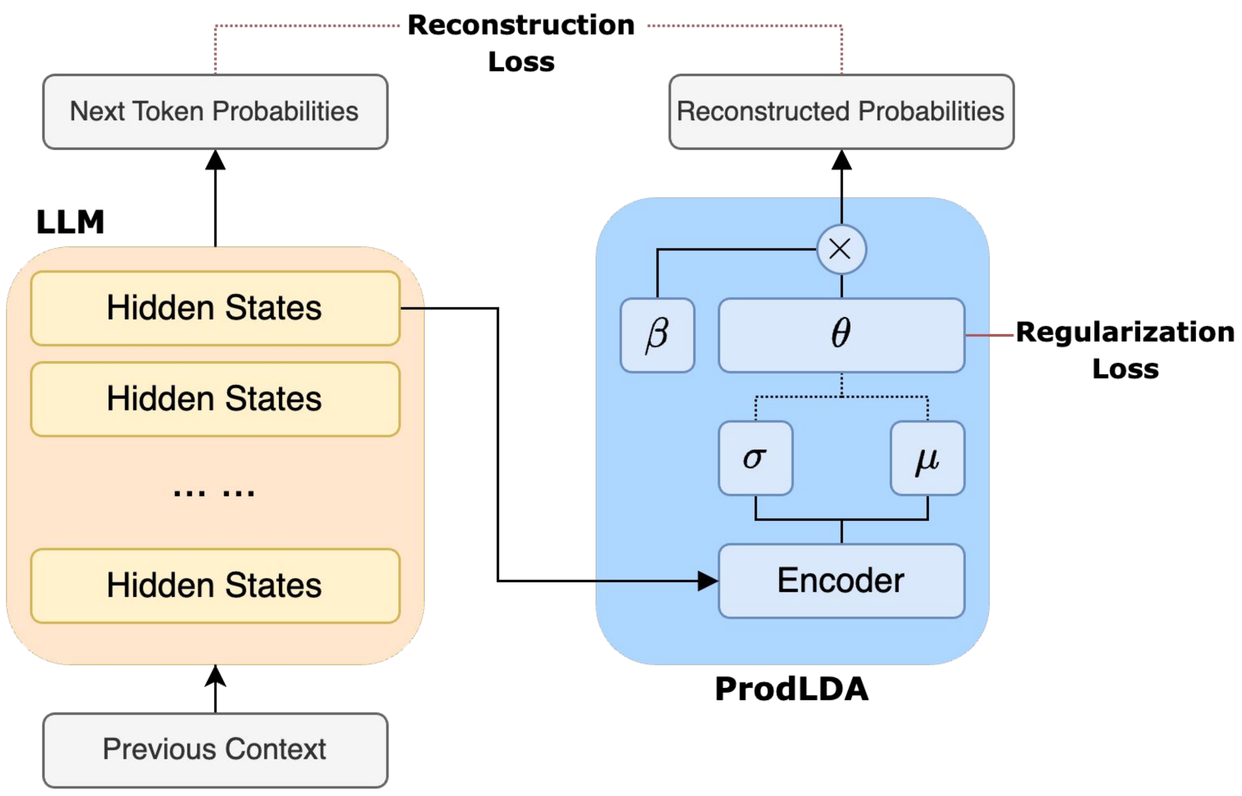
***Extrinsic evaluation** demonstrates downstream utility of the latent topics for dynamically retrieving in-context demonstration examples.*

2. Method

***Generative Process**: LLM predicts next token by marginalizing over latent topic variable θ.*

$$P(w_t|w_{1:t-1}) = \int_\theta p(w_t|\theta)p(\theta|w_{1:t-1})d\theta$$

***Variational Inference**: Approximate the posterior using a variational autoencoder (VAE) based on the* ProdLDA *topic model architecture.*



Optimized with a modified ELBO loss, where the decoder learns to reconstruct the next-token probabilities of the LLM.

$$\mathcal{L} = - \mathbb{E}_{w \sim p_{\text{LLM}}(w_t|w_{1:t-1})} [\log p_\psi(w_t \mid \theta)] + \lambda D_{\text{KL}}\Big(q_\phi(\theta \mid w_{1:t-1}) \Big\| p(\theta)\Big)$$

3. Intrinsic Topic Quality

	AGNews			DBPedia			GovReport		
Metrics	LLM	WE	I-RBO	LLM	WE	I-RBO	LLM	WE	I-RBO
$K = 50$									
LDA	1.86	.108	.983	2.10	.123	.952	2.62	.150	.762
ProdLDA	2.30	.147	.984	2.46	.171	**.991**	2.32	.141	.986
ZeroshotTM	2.44	.162	.903	2.66	.204	.916	2.62	.151	.967
GenerativeTM (Ours)	**2.74**	**.269**	**.991**	**2.78**	**.297**	.989	**2.80**	**.254**	**.993**
$K = 100$									
LDA	1.86	.105	.987	1.99	.120	.959	2.40	.148	.807
ProdLDA	2.24	.136	.982	2.49	.170	.989	2.29	.135	.988
ZeroshotTM	2.42	.172	.917	**2.78**	.270	.897	2.60	.148	.984
GenerativeTM (Ours)	**2.69**	**.254**	**.989**	2.76	**.301**	**.990**	**2.77**	**.250**	**.989**

Outperforms all baselines in automatic coherence and diversity metrics.

4. Extrinsic Evaluation for ICL Retrieval

*Using a special prompt (e.g., "**summarize the article in one word**"), take the topic proportion of the next token prediction for retrieval.*

	DBPedia-14	XSum		
	Accuray	ROUGE-1	ROUGE-2	ROUGE-LSum
Zeroshot	52.64	16.93	3.96	14.50
Hidden State	53.14	17.47	4.01	14.73
Probabilities	66.07	17.78	**4.24**	**15.09**
Topic (K = 50)	73.21	17.69	4.08	14.90
Topic (K = 100)	**74.00**	**18.00**	4.12	14.85

Demonstrates significant potential in retrieving examples for classification.

5. Qualitative Examples

Label	ProdLDA	ZeroshotTM	GenerativeTM
Company	operator, base, network, company, operations, distribution, internet, content, subsidiary, computer	business, firm, investment, bank, countries, offices, funds, banks, companies, businesses	retailer, brand, company, distributor, label, shop, manufacturer, firm, supplier, clothing
Film	comedy, role, director, action, films, movie, roles, cinema, film, feature	life, play, drama, film, screenplay, woman, man, adaptation, plot, title	crime, drama, thriller, romance, comedy, suspense, horror, noir, mystery, fantasy
Natural Place	point, level, island, range, mountain, views, elevation, border, peak, hill	point, range, mountain, peak, border, views, end, pass, level, trail	forests, forest, habitat, habitats, woodland, scrub, destruction, vegetation, soils, grass
Written Work	volume, issue, authors, magazine, science, anthology, trade, edition, aspects, review	field, aspects, research, journal, editor, behalf, chief, access, review, peer	book, novel, novels, chapter, books, tale, trilogy, memoir, poem, manga

6. Conclusion & Future Works

- *Propose a novel framework for inferring the latent topics of LLMs by training a VAE to reconstruct the next token probabilities.*
- *For future work, we will adapt this approach to other tasks such as Uncertainty Quantification and Hallucination Detection.*

ACL 2025 VIENNA