SCAN ME

# BBScore: A Brownian Bridge Based Metric for Assessing Text Coherence

Zhecheng Sheng, Tianhao Zhang, Chen Jiang, Dongyeop Kang

University of Minnesota
Driven to Discover®

## Motivation

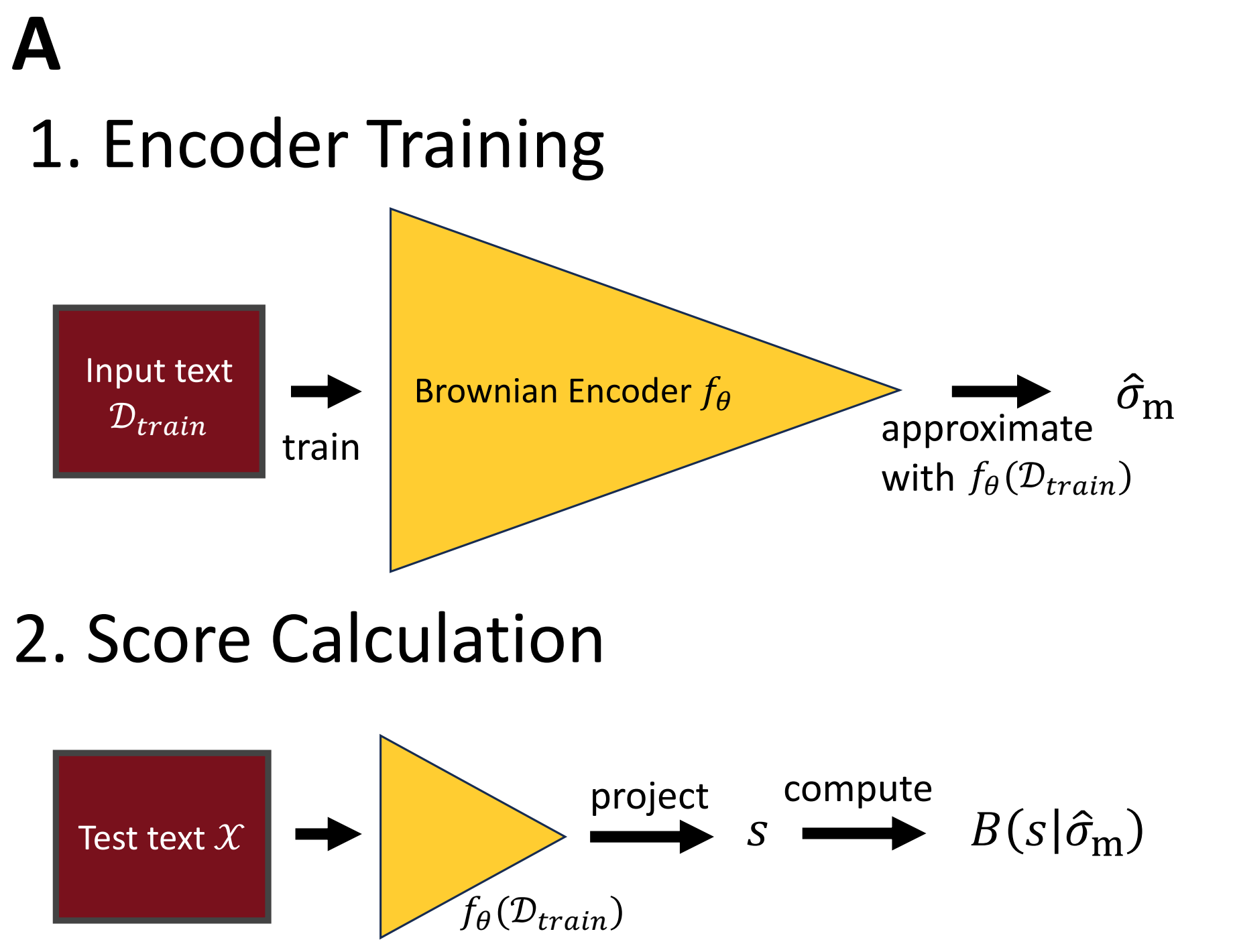
We hypothesis there exists a **sequential** and **cohesive** relationship among sentences in a latent space that represents main goal or opinion of the text. And a coherent text has three properties:

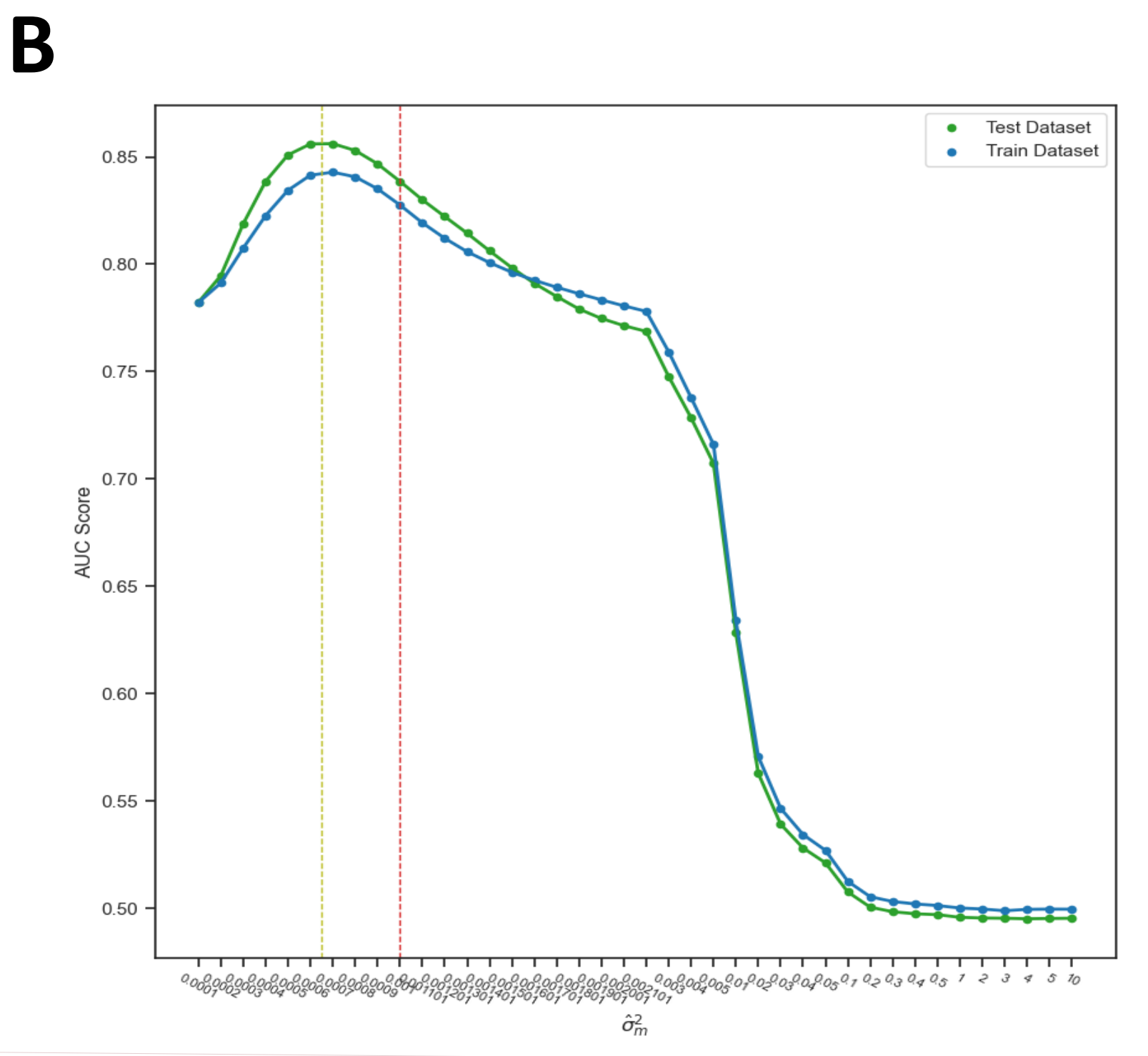
- ✓ Main idea should be kept over the course
- ✓ Sentences at the beginning and end should emphasis the main idea
- ✓ Sentences in the middle can depart from the main idea a little bit but still under control

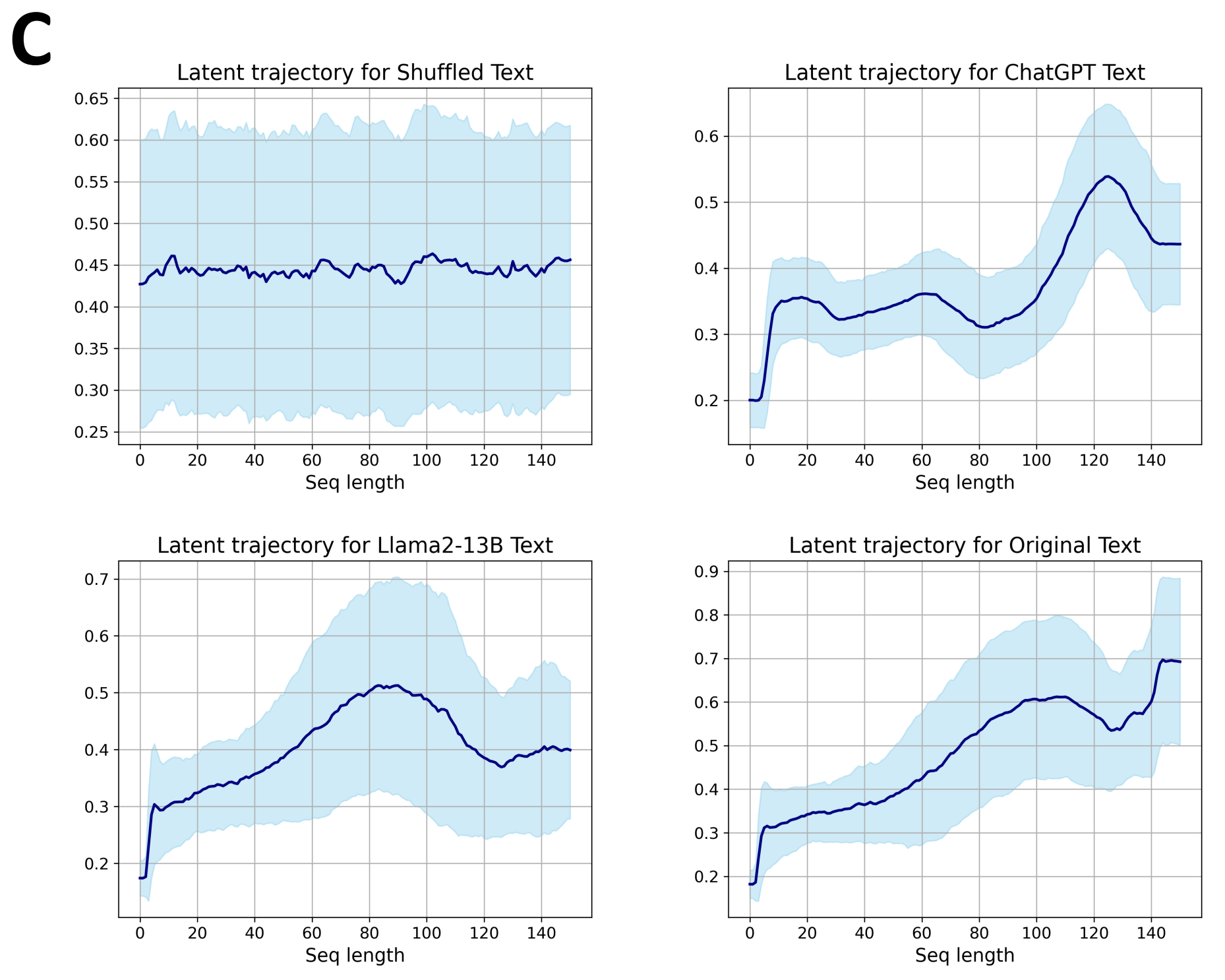
## Method

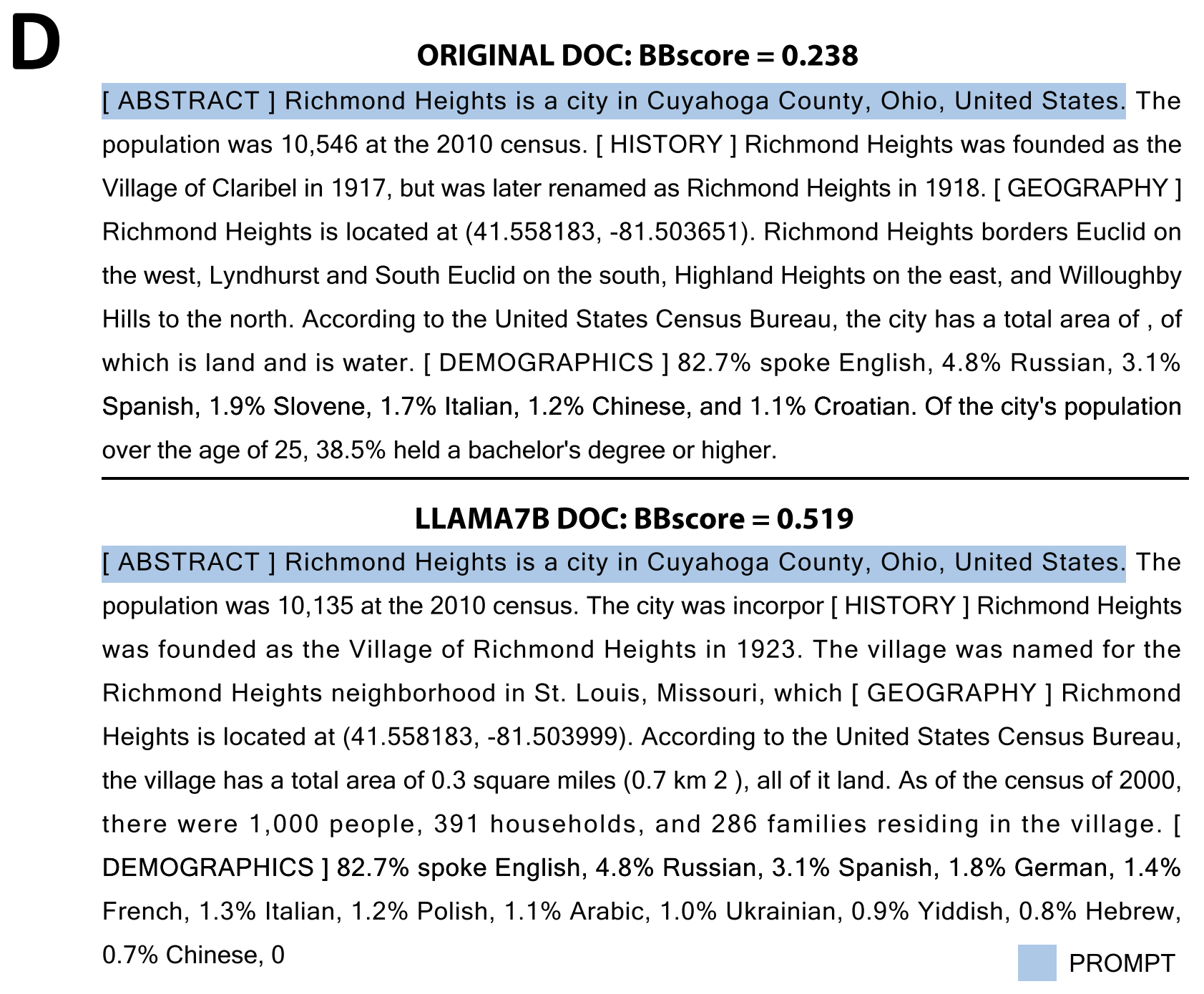
**Brownian Bridge** Consider a sequence $S(t) = s_1, s_2, \dots, s_T$, and $S(t) \sim \mathcal{N}(\mu(t), \sigma^2(t)\mathbf{I}), t \in [0, T]$. We then have $\mu(t) = a + \frac{t}{T}(b-a), \sigma^2 = \frac{t(T-t)}{T}\sigma_m$. This stochastic process models the trajectory in the latent space as both the start and end point are anchored. In our hypothesis, the diffusion coefficient $\sigma_m$ encodes the domain-specific abstract relationship of the input text.

**BBScore** We train a transformer encoder $f_\theta$ with a contrastive Brownian Bridge loss and apply a maximum likelihood estimator for $\hat{\sigma}_m$. BBScore is defined as $B(s|\hat{\sigma}_\mathrm{m}) = \frac{|ln(\prod_{i=2}^{T(s)-1} \mathcal{L}_i)|}{T(s)-2}$, note that $\mathcal{L}_i$ is the likelihood function for sentence $i$ and contains $\hat{\sigma}_\mathrm{m}$.


A
1. Encoder Training
Input text $\mathcal{D}_{train}$
train
Brownian Encoder $f_\theta$
approximate with $f_\theta(\mathcal{D}_{train})$
$\hat{\sigma}_\mathrm{m}$
2. Score Calculation
Test text $\mathcal{X}$
$f_\theta(\mathcal{D}_{train})$
project
$s$
compute
$B(s|\hat{\sigma}_\mathrm{m})$






D

**ORIGINAL DOC: BBscore = 0.238**

[ ABSTRACT ] Richmond Heights is a city in Cuyahoga County, Ohio, United States. The population was 10,546 at the 2010 census. [ HISTORY ] Richmond Heights was founded as the Village of Claribel in 1917, but was later renamed as Richmond Heights in 1918. [ GEOGRAPHY ] Richmond Heights is located at (41.558183, -81.503651). Richmond Heights borders Euclid on the west, Lyndhurst and South Euclid on the south, Highland Heights on the east, and Willoughby Hills to the north. According to the United States Census Bureau, the city has a total area of , of which is land and is water. [ DEMOGRAPHICS ] 82.7% spoke English, 4.8% Russian, 3.1% Spanish, 1.9% Slovene, 1.7% Italian, 1.2% Chinese, and 1.1% Croatian. Of the city's population over the age of 25, 38.5% held a bachelor's degree or higher.

**LLAMA7B DOC: BBscore = 0.519**

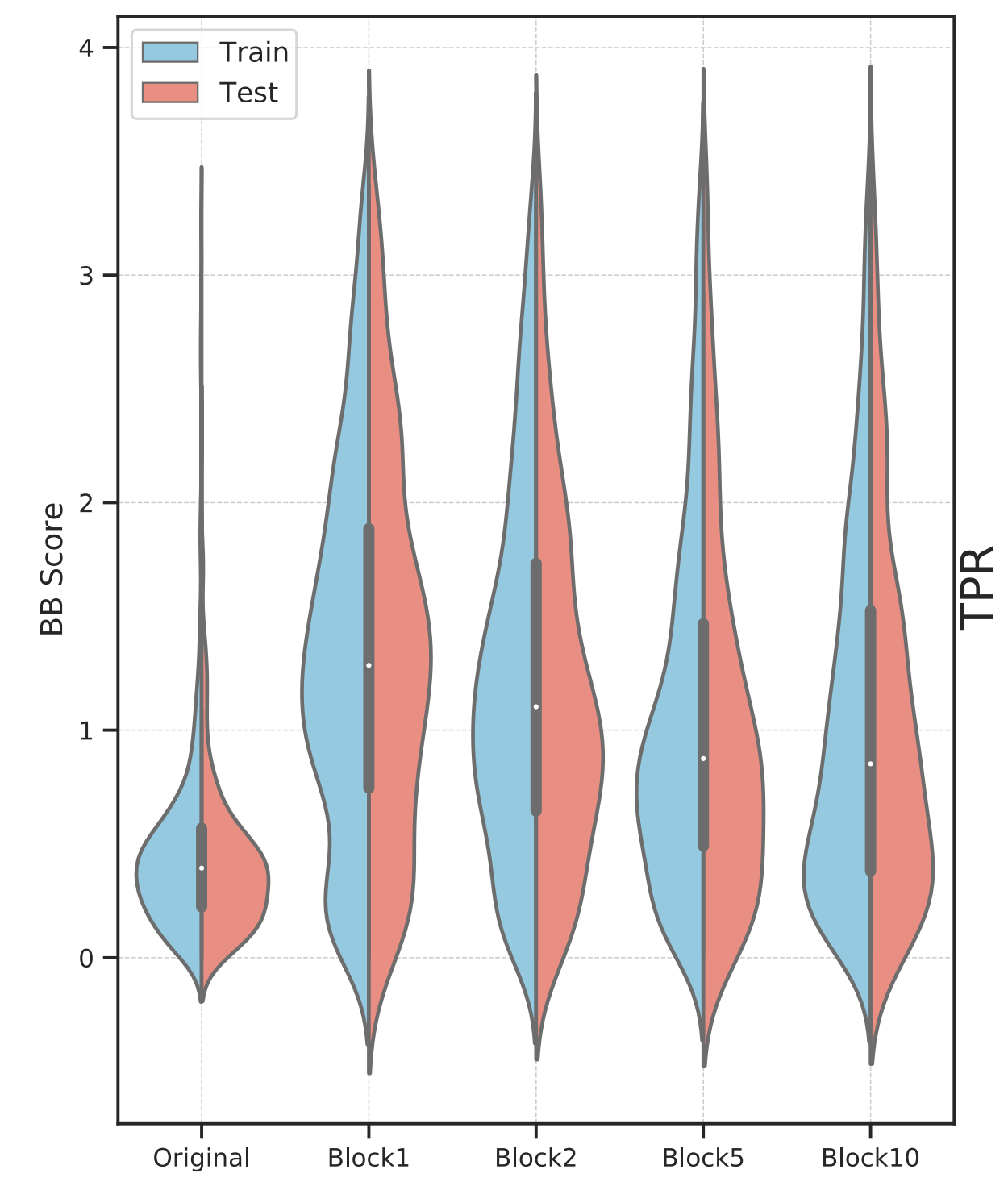
[ ABSTRACT ] Richmond Heights is a city in Cuyahoga County, Ohio, United States. The population was 10,135 at the 2010 census. The city was incorpor [ HISTORY ] Richmond Heights was founded as the Village of Richmond Heights in 1923. The village was named for the Richmond Heights neighborhood in St. Louis, Missouri, which [ GEOGRAPHY ] Richmond Heights is located at (41.558183, -81.503999). According to the United States Census Bureau, the village has a total area of 0.3 square miles (0.7 km 2 ), all of it land. As of the census of 2000, there were 1,000 people, 391 households, and 286 families residing in the village. [ DEMOGRAPHICS ] 82.7% spoke English, 4.8% Russian, 3.1% Spanish, 1.8% German, 1.4% French, 1.3% Italian, 1.2% Polish, 1.1% Arabic, 1.0% Ukrainian, 0.9% Yiddish, 0.8% Hebrew, 0.7% Chinese, 0

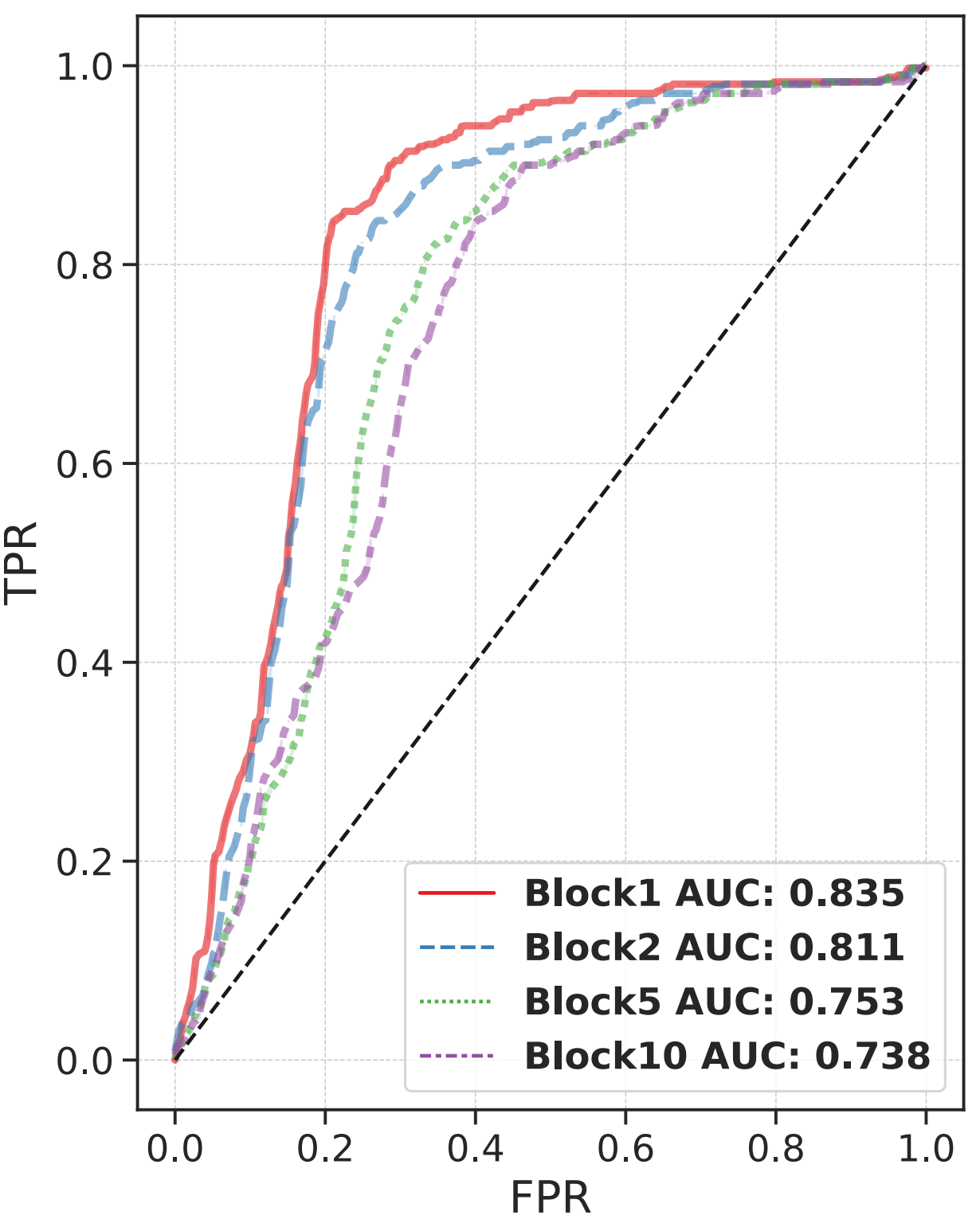
▨ PROMPT

Panel **A** is the procedure of how to generate BBScore from scratch.
Panel **B** demonstrate why a good estimation of $\sigma_m$ is important and validates there is room for a better estimation.
Panel **C** shows the projected latent trajectories for Wikisection text generated by each of the Large Language models plus a random permutation one.
Panel **D** demonstrates two sample text generated by human and LlaMA 7B model and their corresponding BBScore.

## Results

**RQ1**: While designed for global coherence, does BBScore capture both global and local coherence effectively in a synthetic setting?





We use **Wikisection** corpus as our experiment dataset.

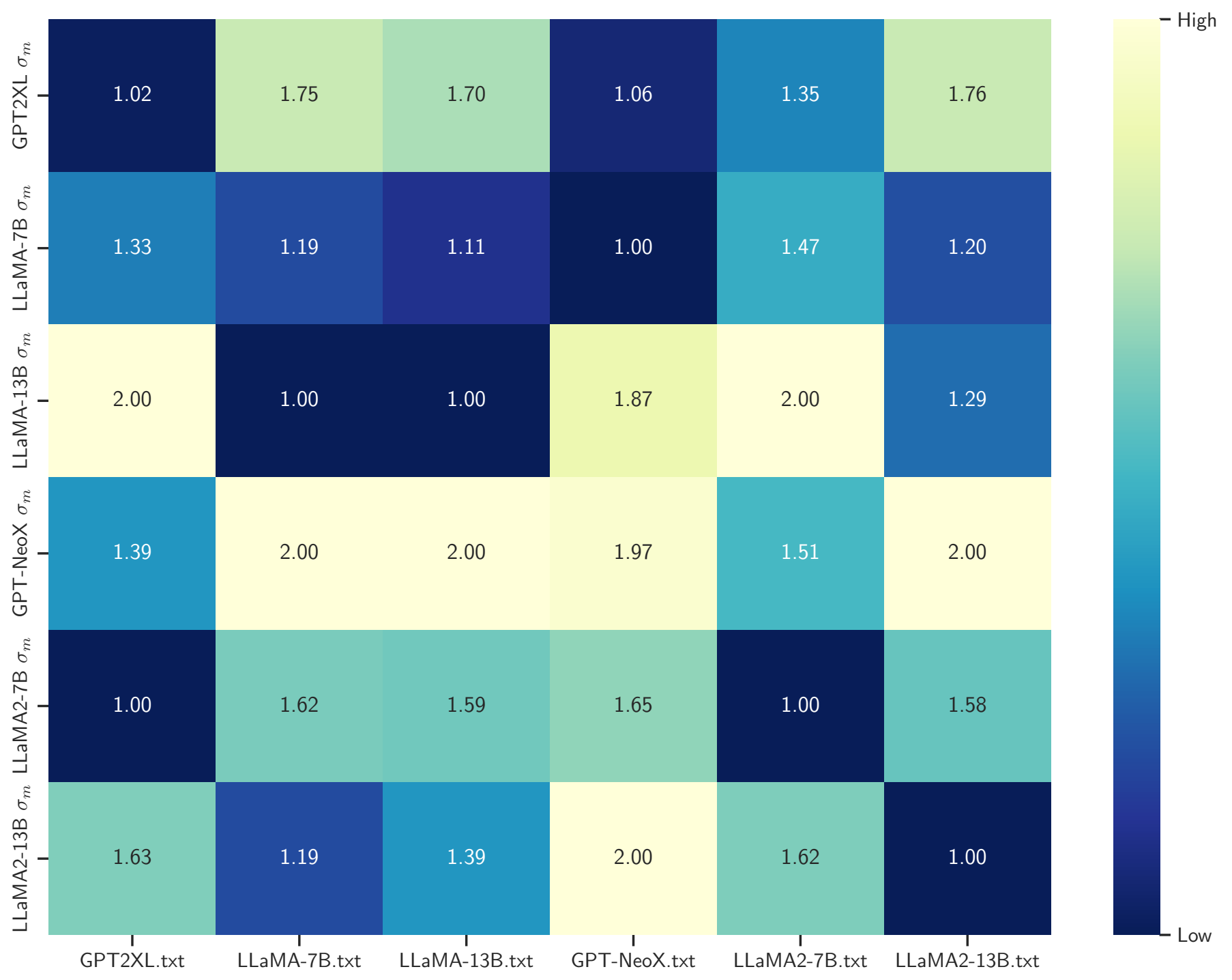
- The violin plot shows the BBScore for the original text and each of the permutated copies. Different number of consecutive sentences are used as block size. From the result we can observe human written articles poses lowest BBScore, which means the corresponding latents means are closer to the underlying Brownian process. The right ROC curves show quantitative results in a paired (original vs shuffled) classification task without any supervision. We can conclude BBScore can inherently distinguish incoherent text from coherent ones.
- In the table below , we compare our methods with previous several baselines from previous work on the global discrimination task. We add a 3-layer MLP for training in the BBScore + CLF method.

| Methods | Train | | | | Test | | | |
|---|---|---|---|---|---|---|---|---|
| | $\mathcal{D}_{b=1}$ | $\mathcal{D}_{b=2}$ | $\mathcal{D}_{b=5}$ | $\mathcal{D}_{b=10}$ | $\mathcal{D}_{b=1}$ | $\mathcal{D}_{b=2}$ | $\mathcal{D}_{b=5}$ | $\mathcal{D}_{b=10}$ |
| EntityGrid (Barzilay and Lapata 2008) | 79.17 | 86.20 | 74.51 | 60.70 | 85.73 | 82.79 | 75.81 | 64.65 |
| UnifiedCoherence (Moon et al. 2019) | **99.75** | 98.60 | 97.10 | 96.21 | **99.73** | 97.86 | 96.90 | 96.09 |
| BBScore | 76.29 | 75.12 | 73.04 | 73.12 | 83.39 | 80.71 | 79.36 | 78.66 |
| BBScore + CLF | 99.12 | **98.95** | **98.68** | **98.54** | 98.92 | **98.46** | **98.25** | **98.50** |

**RQ2**: Can BBScore detects texts where the departure from desired coherence are not manipulated and recognize their differences ?

| Methods | Train | | | | | | Test | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ♣ | ♠ | ♦ | ♥ | ♢ | ♡ | ♣ | ♠ | ♦ | ♥ | ♢ | ♡ |
| EntityGrid | 54.43 | 36.99 | 18.64 | 25.23 | 43.50 | 25.86 | 59.57 | 36.17 | 25.88 | 32.19 | 39.40 | 22.85 |
| BBScore | 76.96 | 85.07 | 82.21 | 82.42 | 73.07 | 75.29 | 77.75 | 83.74 | 81.16 | 81.82 | 75.63 | 75.13 |
| BBScore + CLF | **83.21** | **85.51** | **88.04** | **88.03** | **92.19** | **91.89** | **83.01** | **86.17** | **86.18** | **86.73** | **92.02** | **91.89** |

♣: GPT2XL ♠: GPT-NeoX ♦: LLaMA-7B ♥: LLaMA-13B ♢: LLaMA2-7B ♡: LLaMA2-13B



| | GPT2XL.txt | LLaMA-7B.txt | LLaMA-13B.txt | GPT-NeoX.txt | LLaMA2-7B.txt | LLaMA2-13B.txt |
|---|---|---|---|---|---|---|
| GPT2XL $\sigma_m$ | 1.02 | 1.75 | 1.70 | 1.06 | 1.35 | 1.76 |
| LLaMA-7B $\sigma_m$ | 1.33 | 1.19 | 1.11 | 1.00 | 1.47 | 1.20 |
| LLaMA-13B $\sigma_m$ | 2.00 | 1.00 | 1.00 | 1.87 | 2.00 | 1.29 |
| GPT-NeoX $\sigma_m$ | 1.39 | 2.00 | 2.00 | 1.97 | 1.51 | 2.00 |
| LLaMA2-7B $\sigma_m$ | 1.00 | 1.62 | 1.59 | 1.65 | 1.00 | 1.58 |
| LLaMA2-13B $\sigma_m$ | 1.63 | 1.19 | 1.39 | 2.00 | 1.62 | 1.00 |

- **AI-Human discrimination**

In the AI-Human discrimination task, we pair an AI generated Wikisection article with the original one and ask the model to distinguish between them. The tables shows the results from multiple text generation models and BBScore accomplish satisfying performances in separating the pair across all the conditions

- **LLM detection**

In this task we use different LLMs as the writing agents and examine when trained with a mixture of coherent text with different styles, whether $\hat{\sigma}_\mathrm{m}$ can recognize the differences at test time. The heatmap displays the pairwise normalized distance between $\hat{\sigma}_\mathrm{m}$ estimated from an input text and the one obtained from training data. In the results, we observe that 5 out of the 6 test texts are in close proximity (among the top 2) to their respective source LLMs,

## > Takeaways

- Comparing to previous methods, BBScore presents a more flexible way to model text coherence for downstream tasks and reaches good performance
- Generalization to out of domain text requires further investigations
- Estimation of $\sigma_m$ is important in BBScore calculation