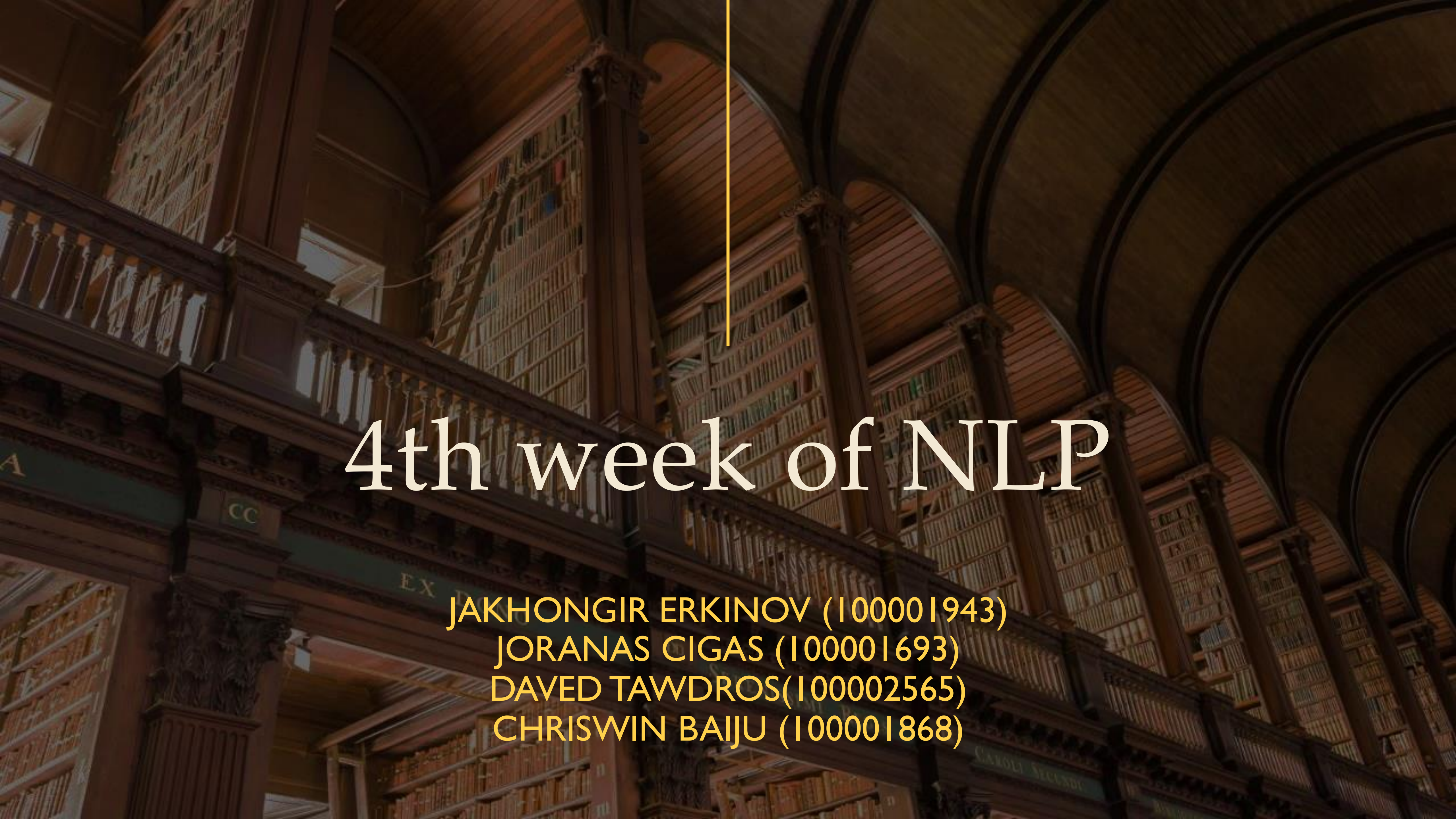
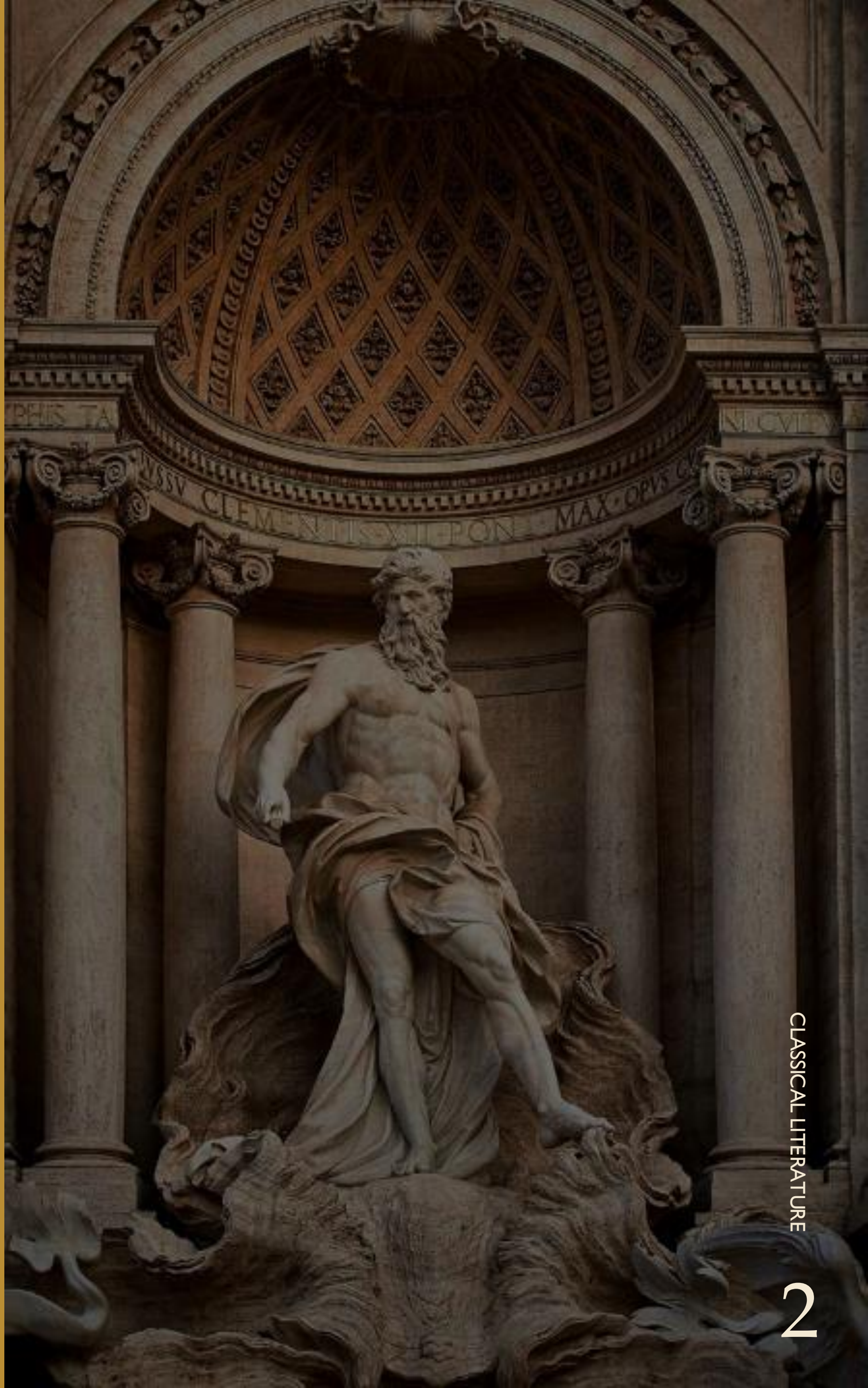


4th week of NLP

JAKHONGIR ERKINOV (100001943)
JORANAS CIGAS (100001693)
DAVED TAWDROS(100002565)
CHRISWIN BAIJU (100001868)

AGENDA

- Compressibility
- Narrative Pace
- Topic Drift
- Paragraph Semantic Coherence



CLASSICAL LITERATURE

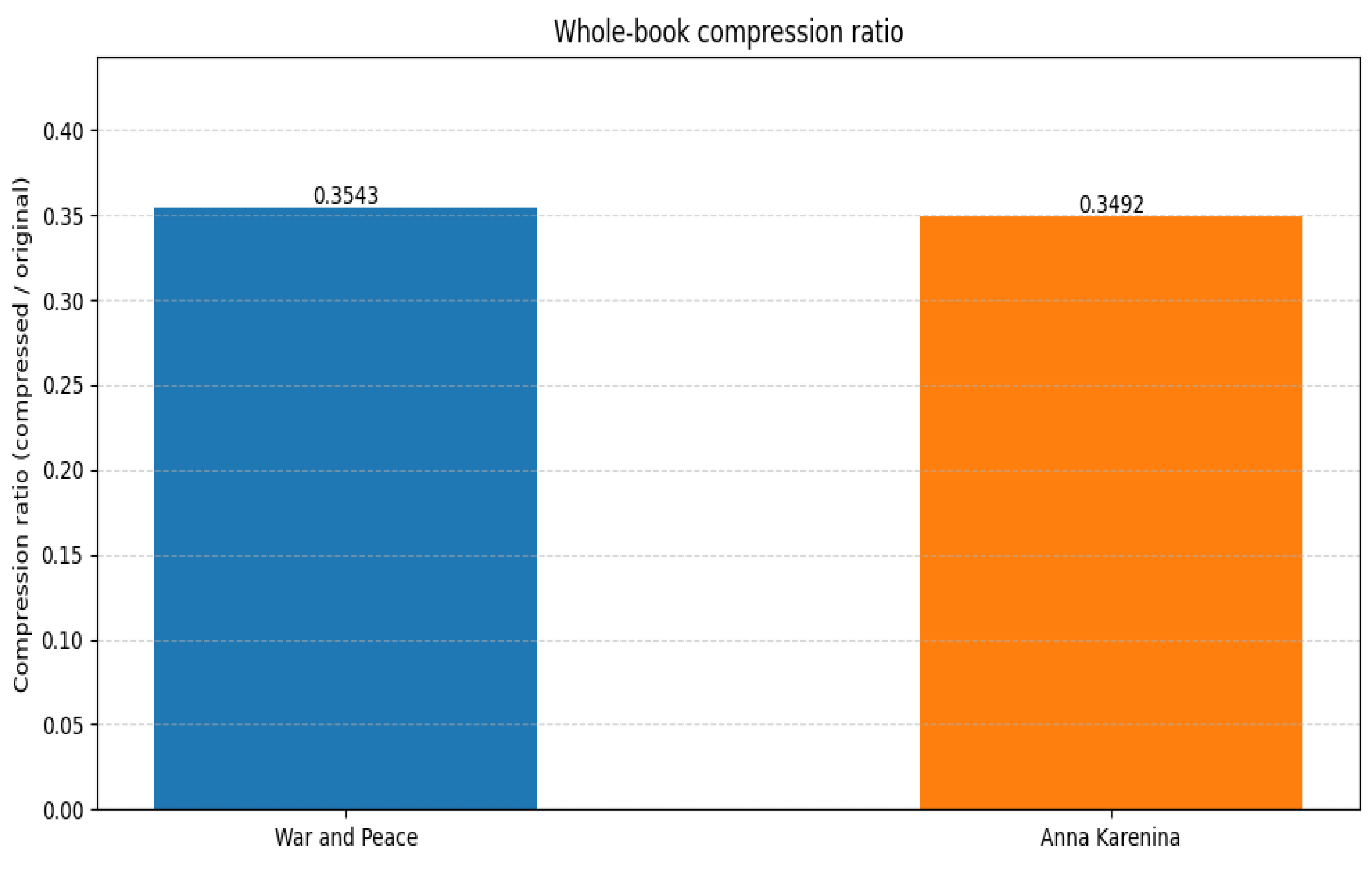
Compressibility

COMPRESSIBILITY

- **The purpose**: Measure how well long novels compress and compare compressibility between books (e.g., *War & Peace*, *Anna Karenina*).

- **Tools:** gzip, bz2, zlib (libraries) — compress raw text and compute ratio = compressed_size / original_size.
  Visualization: pandas, matplotlib.

WHOLE BOOK COMPRESSION RATIO



- Metric: **compression ratio = compressed size / original size** (lower = more compressible).
- Results: *War and Peace* = **0.3543**, *Anna Karenina* = **0.3492**.
- Interpretation: **Anna is slightly more compressible** → marginally more repetition/predictability.
- Caveat: difference is **very small** (~0.005), so treat as *suggestive*, not decisive.

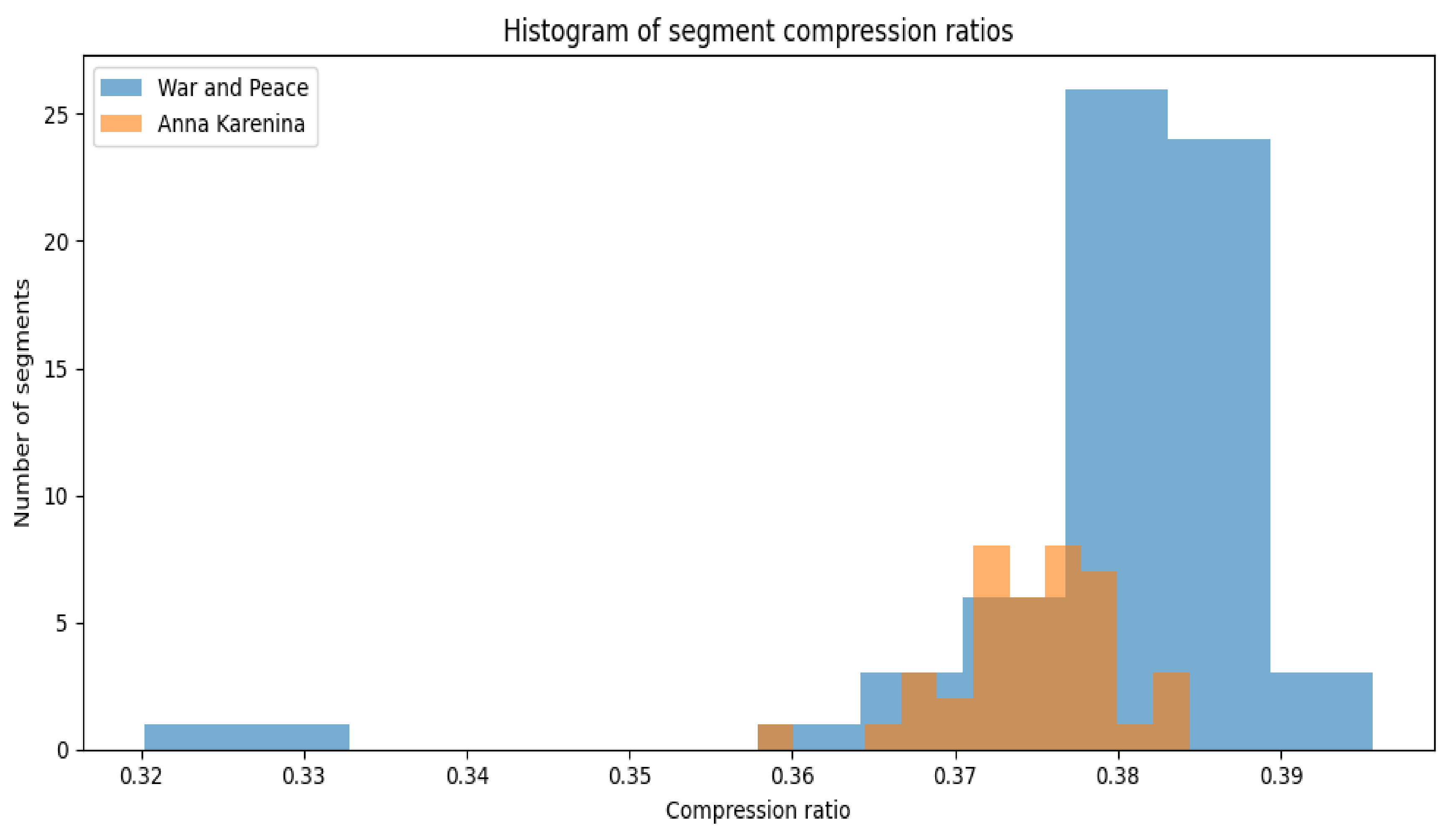
HISTOGRAM OF SEGMENT COMPRESSION RATIOS

Segments used

- War and Peace — **65 segments**
- Anna Karenina — **40 segments**

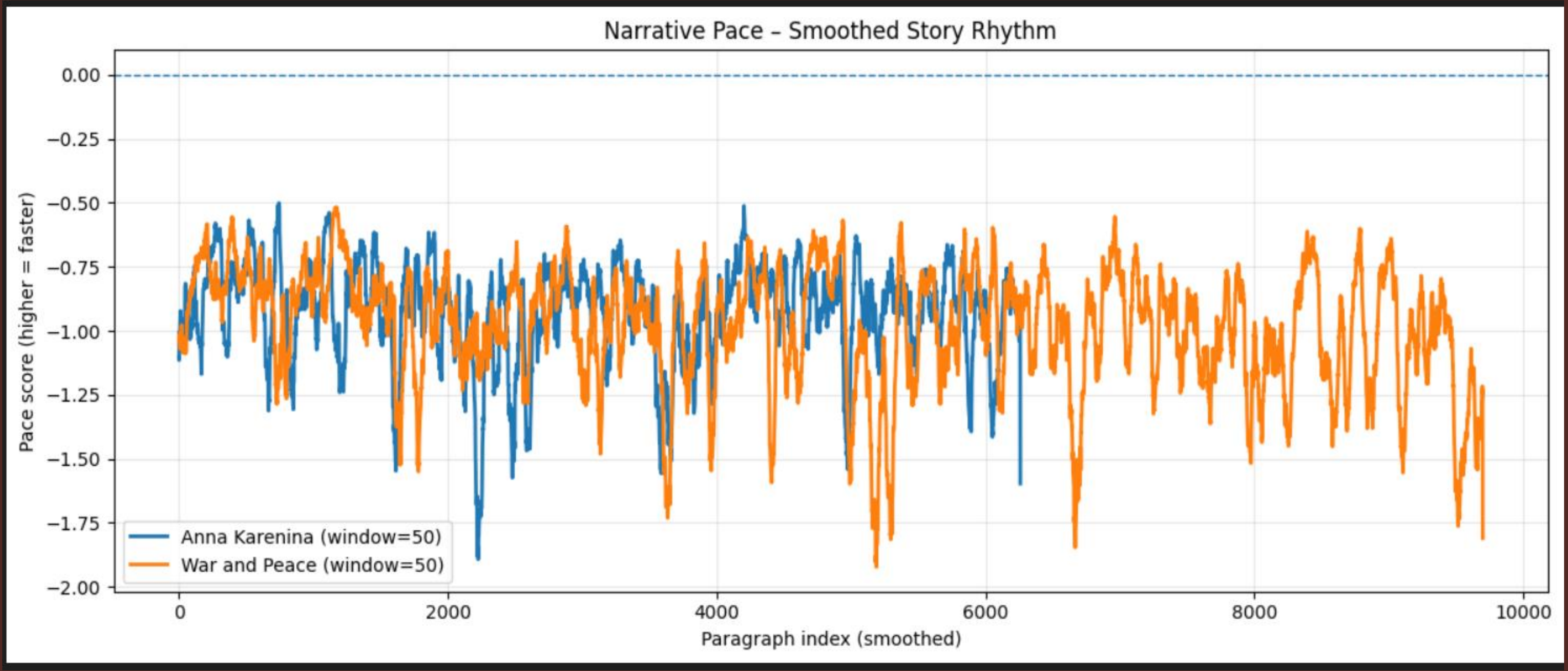
- **Blue bars are taller** mainly because *War and Peace* has more segments (65 vs 40). Histograms count items, so more segments → taller bars even when distributions overlap.

- Interpretation: taller bars = more frequent segments in that ratio bin; not necessarily "more compressible overall" — see slide note for whole-book ratio.

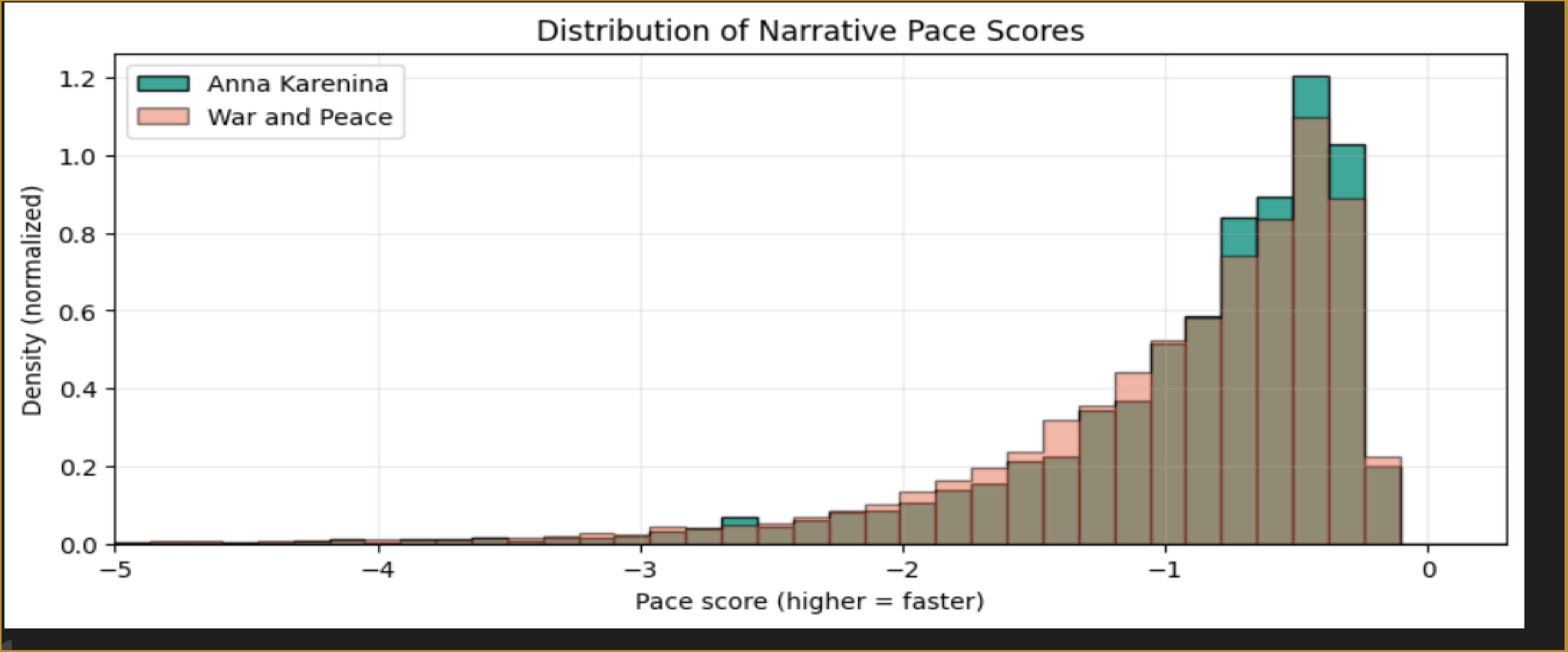


Comparing Narrative Pace in Two Books Using AI

SMOOTHED STORY RHYTHM



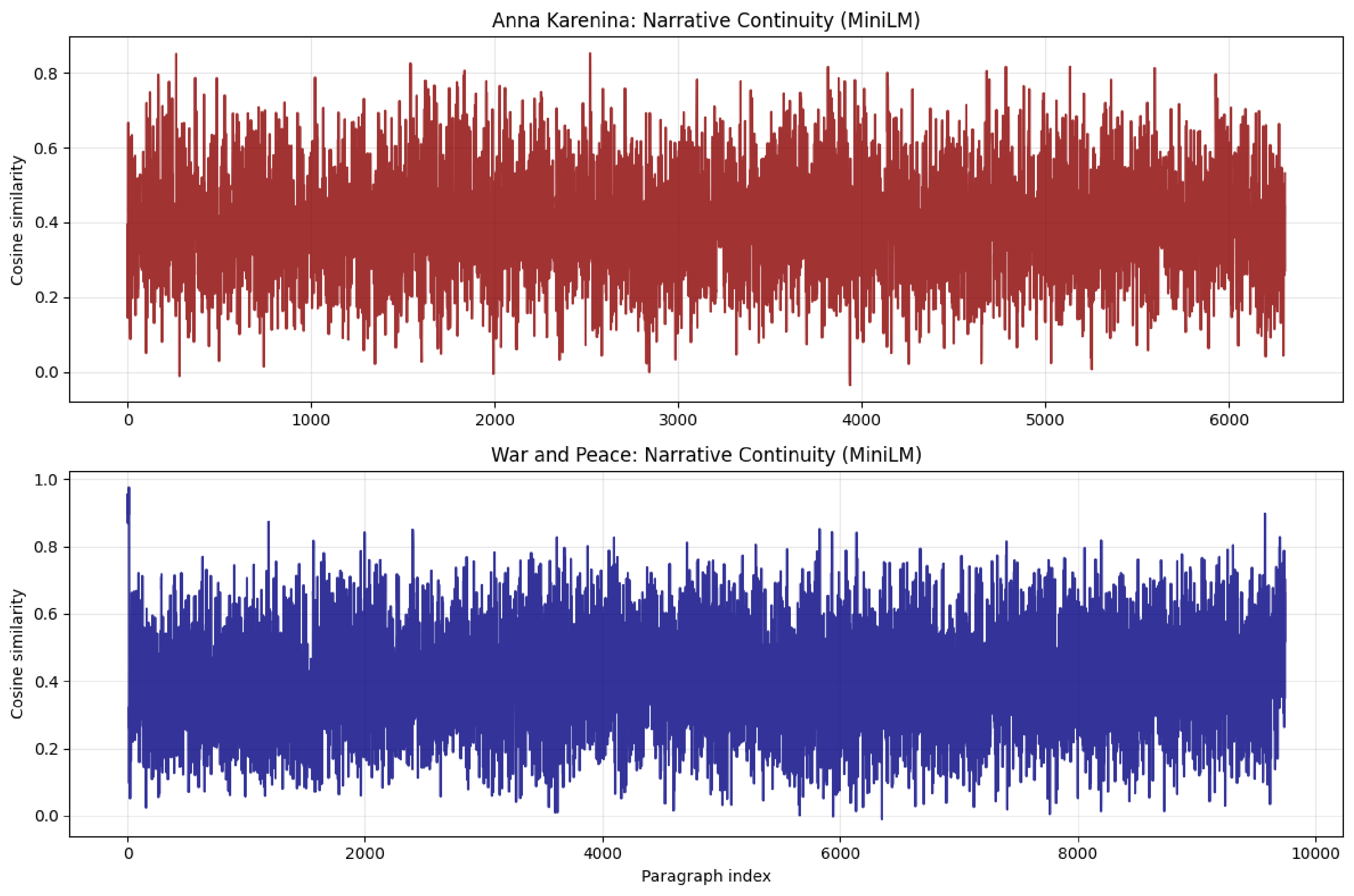
WHICH BOOK IS FASTER?



Topic Drift Between Paragraphs

Topic Drift & Narrative Scale

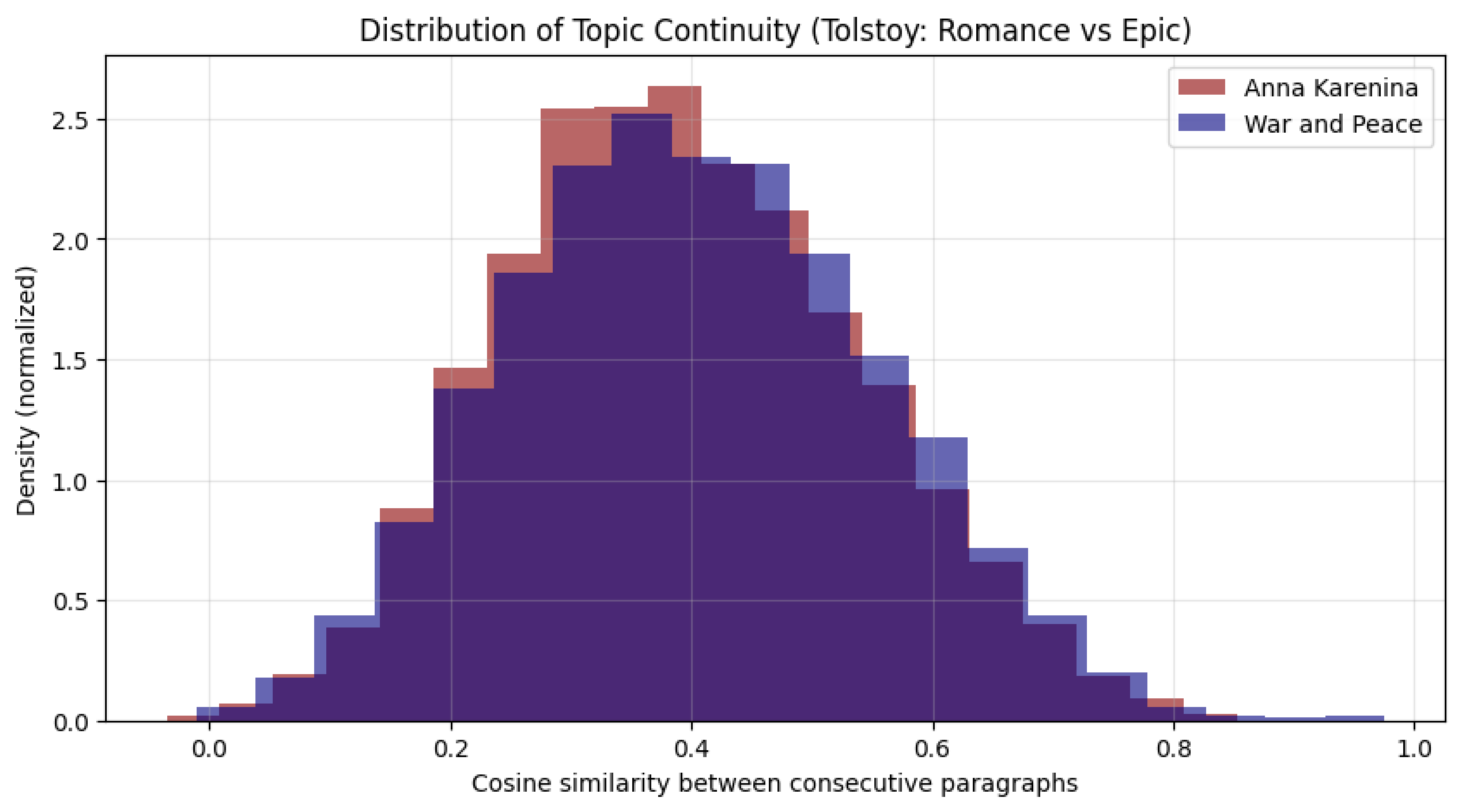
- **Epic Scale:** The blue line (*War and Peace*) extends much further along the X-axis, visually demonstrating its massive length compared to *Anna Karenina*.
- **Dynamic Range:** Both novels fluctuate between 0.0 and 0.8 similarity, showing a dynamic mix of action, dialogue, and description.



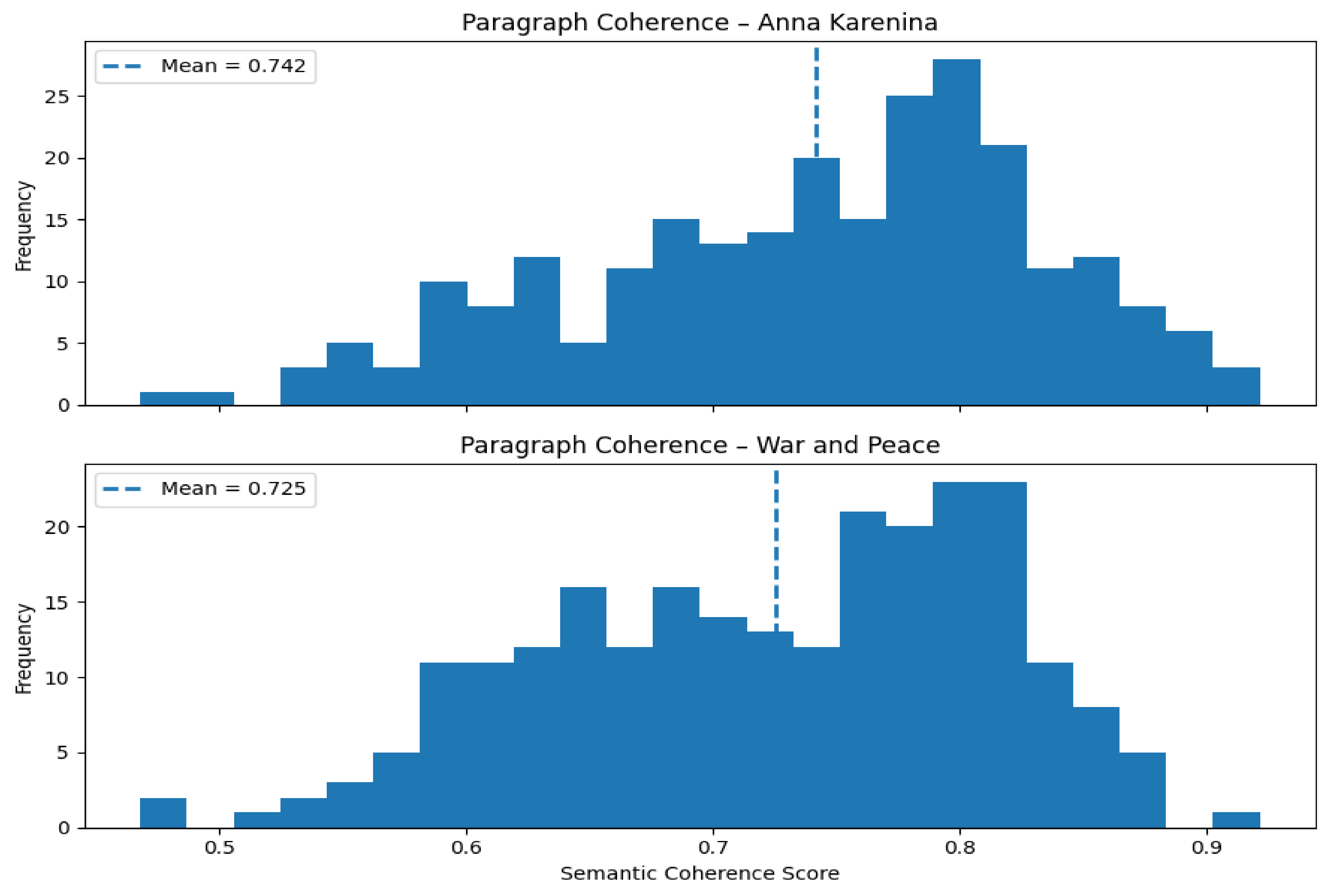
CLASSICAL LITERATURE

DISTRIBUTION OF NARRATIVE CONTINUITY

- **Perfect Overlap:** The two distributions align almost perfectly, indicating that the "rhythm" of the writing is identical in both books.
- **Stylistic Fingerprint:** The peak around 0.4 similarity suggests a specific "Tolstoy Value"—a consistent level of narrative density he maintains regardless of the subject matter.



- Anna Karenina
- **Paragraphs analysed:** 250
- **Mean coherence: 0.742**
- **Std (spread): 0.092**
- **Range: 0.47 – 0.92**
- War and Peace
- **Paragraphs analysed:** 242
- **Mean coherence: 0.725** (slightly lower than *Anna Karenina*)
- **Std: 0.089** → most paragraphs lie roughly in **[0.64, 0.81]**
- **Range: 0.47 – 0.91**



Thank you for attention