Embedding Method Comparison Report
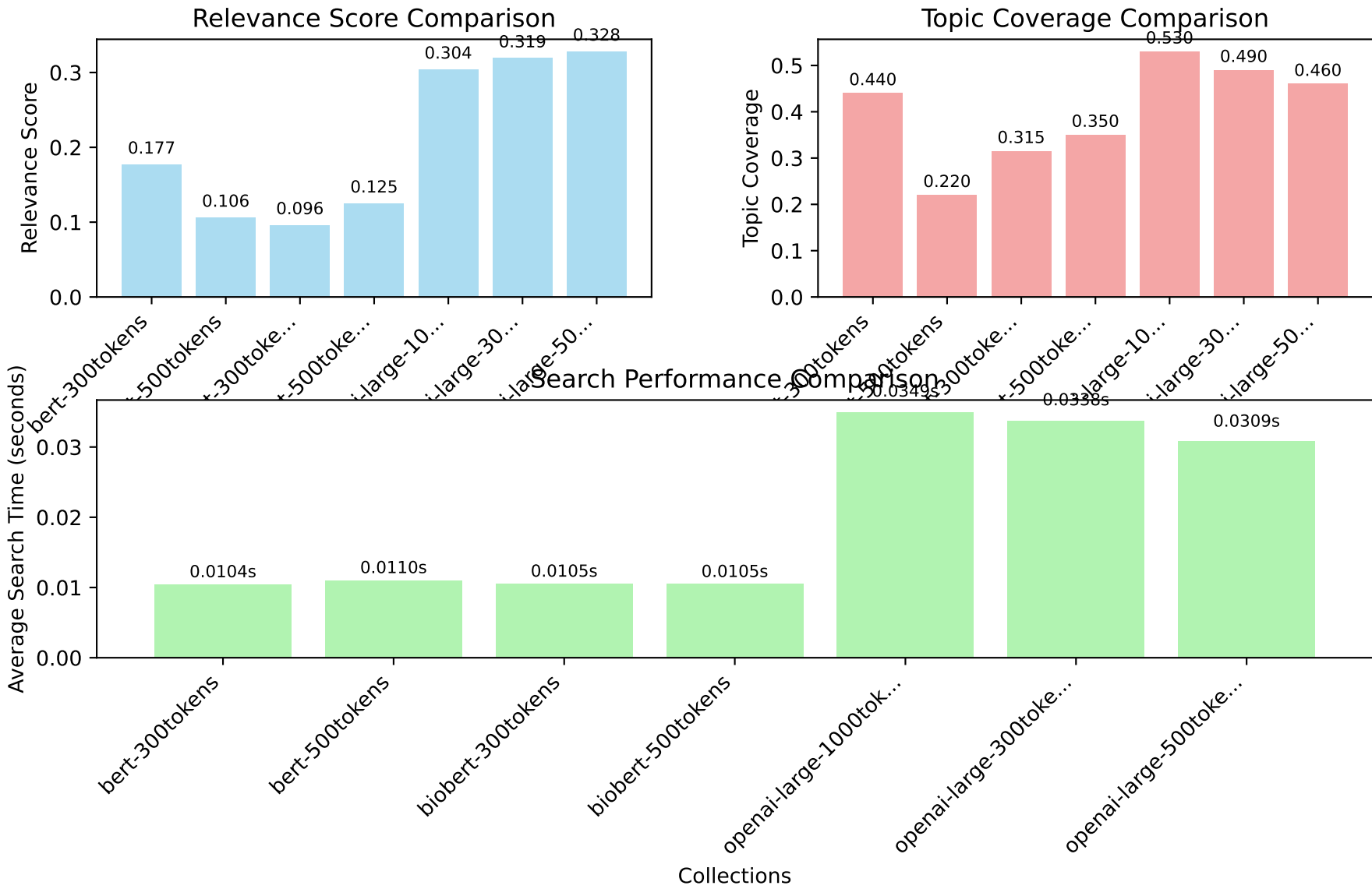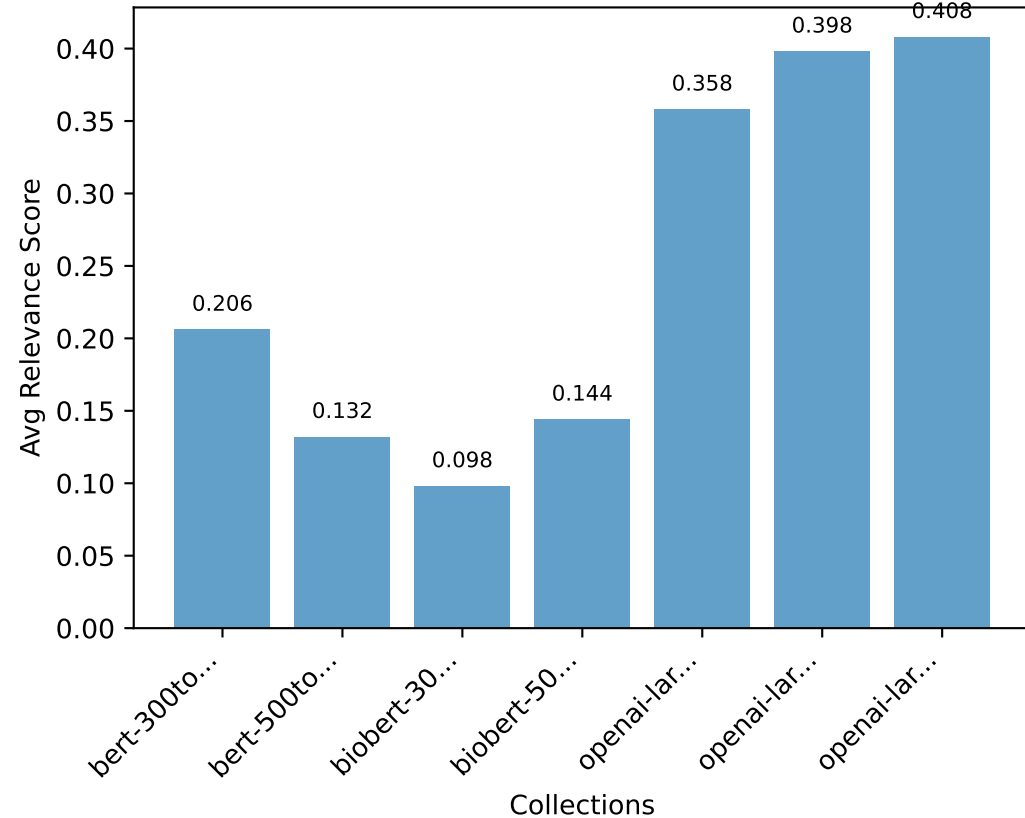
Performance Summary by Collection

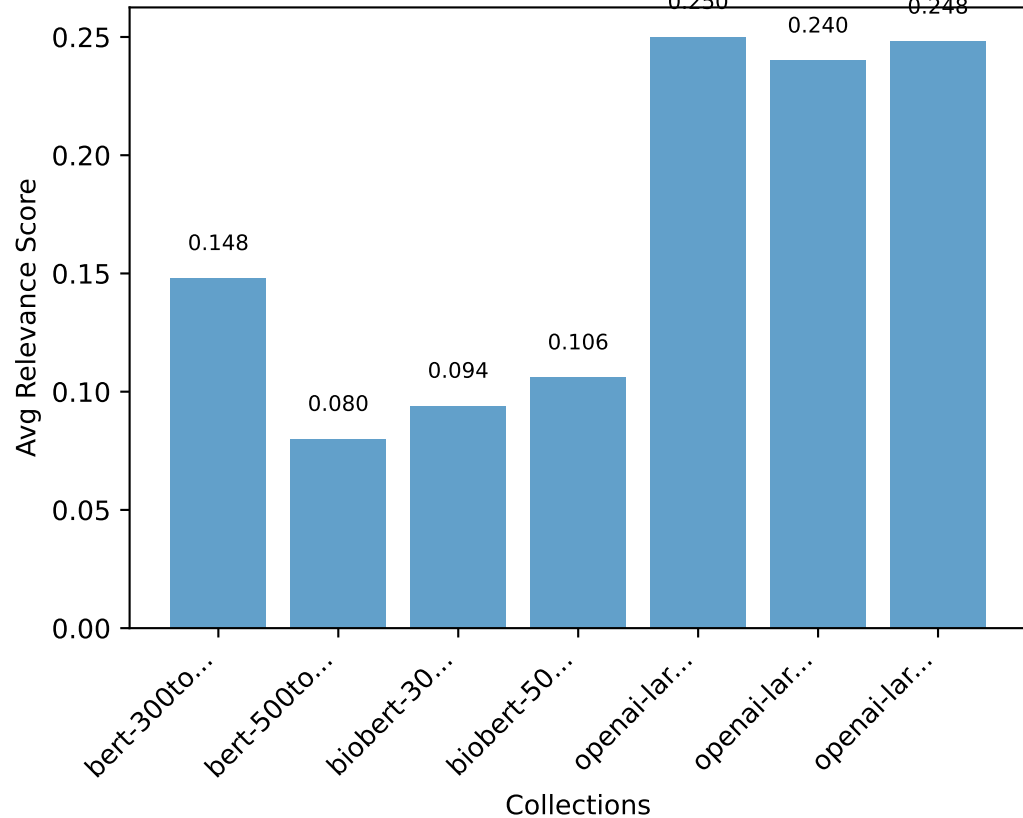|	Search Time (s)	Relevance Score	Topic Coverage	Similarity Score
bert-300tokens	0.0104	0.177	0.44	0.9092
bert-500tokens	0.011	0.106	0.22	0.8993
biobert-300tokens	0.0105	0.096	0.315	0.8675
biobert-500tokens	0.0105	0.125	0.35	0.8548
openai-large-1000tokens	0.0349	0.304	0.53	0.5987
openai-large-300tokens	0.0338	0.319	0.49	0.6509
openai-large-500tokens	0.0309	0.328	0.46	0.6328

Performance by Query Type

Factual_Comprehension Queries

Mechanistic Queries

Embedding Method Comparison Report
Generated: 2025-06-01 13:28:22

Collections Compared:
- bert-500tokens
- bert-300tokens
- biobert-500tokens
- biobert-300tokens
- openai-large-1000tokens
- openai-large-500tokens
- openai-large-300tokens

Overall Performance Summary:

collection	search_time mean	std	avg_similarity_score mean	std	relevance_score mean	std	topic_coverage mean	std
bert-300tokens	0.0104	0.0006	0.9092	0.0105	0.177	0.1491	0.440	0.3635
bert-500tokens	0.0110	0.0006	0.8993	0.0091	0.106	0.1484	0.220	0.2751
biobert-300tokens	0.0105	0.0007	0.8675	0.0080	0.096	0.0971	0.315	0.3019
biobert-500tokens	0.0105	0.0007	0.8548	0.0093	0.125	0.1011	0.350	0.3055
openai-large-1000tokens	0.0349	0.0036	0.5987	0.1101	0.304	0.2718	0.530	0.3924
openai-large-300tokens	0.0338	0.0023	0.6509	0.1262	0.319	0.2960	0.490	0.3872
openai-large-500tokens	0.0309	0.0026	0.6328	0.1169	0.328	0.3307	0.460	0.3978

Key Findings:
- **Fastest Search**: bert-300tokens
- **Highest Relevance**: openai-large-500tokens
- **Best Topic Coverage**: openai-large-1000tokens
- **Highest Similarity Scores**: bert-300tokens

Performance by Query Type:

query_type	collection	relevance_score	topic_coverage	search_time
factual_comprehension	bert-300tokens	0.206	0.61	0.0101
	bert-500tokens	0.132	0.24	0.0110
	biobert-300tokens	0.098	0.36	0.0107
	biobert-500tokens	0.144	0.39	0.0105
	openai-large-1000tokens	0.358	0.61	0.0365
	openai-large-300tokens	0.398	0.66	0.0337
	openai-large-500tokens	0.408	0.56	0.0321
mechanistic	bert-300tokens	0.148	0.27	0.0108
	bert-500tokens	0.080	0.20	0.0110
	biobert-300tokens	0.094	0.27	0.0103
	biobert-500tokens	0.106	0.31	0.0104
	openai-large-1000tokens	0.250	0.45	0.0333
	openai-large-300tokens	0.240	0.32	0.0338
	openai-large-500tokens	0.248	0.36	0.0296

Recommendations:
Based on the analysis, consider the following:
1. For speed-critical applications: Use the fastest performing model
2. For accuracy-critical applications: Use the highest relevance scoring model
3. For domain-specific queries: Consider how well each model handles medical terminology

Files Generated:
- PDF Report: experiments/results/embedding_comparison_report_20250601_132821.pdf

Total Results: 70 comparisons across 10 queries