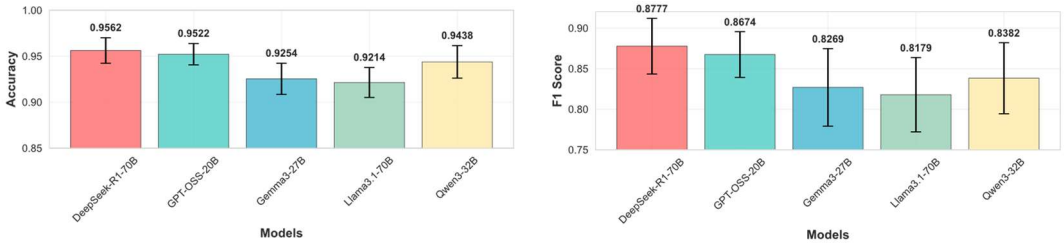
A. Comparative analysis of large language model performance

In the context of rapid development of current large language models, different models exhibit significant performance variations in specific tasks. To ensure the reliability and effectiveness of literature screening and structured text generation tasks in this paper, we conducted performance evaluations on mainstream open-source large language models with similar parameter scales.

We selected five representative large language models for comparative evaluation: (1) DeepSeek-R1-Distill-Llama-70B; (2) GPT-OSS-20B; (3) Qwen3-32B; (4) Gemma3-27B; (5) Llama3.1-70B. We evaluated the models using the article screening dataset constructed in Section III.B. This dataset contains 500 randomly selected documents annotated by expert human reviewers. The article screening task requires models to determine whether documents meet research requirements based on preset criteria. The models output binary classification results. Table IX and Fig. 8 presented the experimental results.

TABLE IX: Performance comparison of large language models on article screening task.

Model	Accuracy	F1 Score
DeepSeek-R1-70B	**0.9562 ± 0.0139**	**0.8777 ± 0.0343**
GPT-OSS-20B	0.9522 ± 0.0116	0.8674 ± 0.0282
Qwen3-32B	0.9438 ± 0.0177	0.8382 ± 0.0438
Gemma3-27B	0.9254 ± 0.0169	0.8269 ± 0.0478
Llama3.1-70B	0.9214 ± 0.0163	0.8179 ± 0.0458



(a) Accuracy comparison across different models. (b) F1 score comparison across different models.

Fig. 8: Performance comparison of five large language models.

The five models showed significant performance differences in the literature screening task. DeepSeek-R1-Distill-Llama-70B ranked first with an accuracy of 0.9562 ± 0.0139. Its small standard deviation indicates good output stability. GPT-OSS-20B follows closely with an accuracy of 0.9522 ± 0.0116. Although its average performance is slightly lower, it has the smallest standard deviation. This shows excellent consistency. Qwen3-32B achieved an accuracy of 0.9438 ± 0.0177.

Gemma3-27B and Llama3.1-70B achieved accuracies of 0.9254 ± 0.0169 and 0.9214 ± 0.0163, respectively. Although these results are within acceptable ranges, they show significant gaps compared to the first three models. DeepSeek-R1-Distill-Llama-70B achieved the highest F1 score of 0.8777 ± 0.0343 among all tested models. GPT-OSS-20B achieved an F1 score of 0.8674 ± 0.0282. Considering accuracy, F1 score, computational efficiency, and open-source accessibility, we selected DeepSeek-R1-Distill-Llama-70B as the base model for subsequent research.