UNIVERSITÀ
DEGLI STUDI
DI MILANO

University of Milan

Computer Science Department

Natural Language Processing

# Measuring Thematic Alignment of Journal Publications

January 16, 2026

*Student name:*
Anar Mammadov

# 1 Introduction

Academic journals define thematic scopes that describe the research areas they aim to cover. These scope statements guide authors, reviewers and editors and help ensure that published articles are relevant to the intended audience of the journal. However, in broad fields such as artificial intelligence (AI), journal scopes often cover many subfields, making it less obvious how closely individual publications align with the stated scope in practice.

At the same time, recent developments in natural language processing have made it possible to represent text in a numerical form that captures semantic meaning. Sentence embedding models allow texts such as abstracts, descriptions, or short documents to be mapped into vector representations, which can then be compared using similarity measures. This provides a practical way to analyze thematic similarity between texts without relying on simple keyword matching.

In this project, we study the thematic alignment between published articles and the official scope of the Journal of Artificial Intelligence Research (JAIR). JAIR is a journal in the field of artificial intelligence with a broad scope that includes machine learning, reasoning, planning, etc. The amount of research done in this field makes it a suitable choice to test the idea.

The core idea of this project is to treat thematic alignment as a semantic similarity problem. We compare the scope description of the journal with the abstracts of published articles by embedding both texts using sentence embedding models and computing cosine similarity. Each article is assigned a numerical alignment score that reflects how semantically close its abstract is to the stated scope of the journal. By analyzing the distribution of alignment scores and comparing results across multiple embedding models, we aim to provide information on the consistency and robustness of this approach.

# 2 Methodology

## 2.1 Data Loading

We focus on articles published in the Journal of Artificial Intelligence Research (JAIR). To ensure that only articles from the target journal are included, we retrieve publication identifiers using the journal's ISSN through the Crossref API. This allows us to collect Digital Object Identifiers (DOIs) corresponding specifically to JAIR publications within a predefined time window.

For each DOI, we query the Semantic Scholar API to obtain article metadata, including title, publication year, venue, and abstract. Abstracts are used as textual summaries of the content of each article. To allow for reproducibility and robustness against API interruptions, the retrieved records are stored locally and reused when available.

The analysis is restricted to articles published between 2018 and 2026, considering that the number of articles before 2018 was way less and might introduce bias.

## 2.2 Preprocessing

Minimal preprocessing is applied to the collected abstracts. Because sentence embedding models are designed to operate on raw natural language text, aggressive preprocessing such as stopword removal, stemming, or lemmatization is avoided.

The preprocessing steps include:

- filtering articles to the selected publication year range,
- removing entries with missing or extremely short abstracts,
- removing duplicate entries based on DOI and identical abstract text.

These steps improve data quality while preserving the quality of the text most suitable for the sentence embedders.

### 2.3 Scope Representation

The scope of the journal is taken directly from the journal website, as given below:

> The Journal of Artificial Intelligence Research (JAIR) is dedicated to the rapid dissemination of important research results to the global artificial intelligence (AI) community. The journal's scope encompasses all areas of AI, including agents and multi-agent systems, automated reasoning, constraint processing and search, knowledge representation, machine learning, natural language, planning and scheduling, robotics and vision, and uncertainty in AI.

### 2.4 Sentence Embedding Models

To obtain vector representations of text, we use pre-trained sentence embedding models from the Sentence-Transformers library. These models map variable-length text into fixed-dimensional vectors that capture semantic meaning. For this purpose, the state-of-the-art light models are chosen, given the machine restrictions. The main model used is all-MiniLM-L6-v2, although experiments are done through the others as well to confirm the reliability.

The models tested are given below:

- all-MiniLM-L6-v2
- all-mpnet-base-v2
- multi-qa-MiniLM-L6-cos-v1

All embeddings are normalized, enabling cosine similarity to be used directly as a measure of semantic similarity.

### 2.5 Alignment Scoring

Thematic alignment between an article and the journal scope is calculated using cosine similarity between their embedding vectors. For each article abstract, an alignment score is computed as the cosine similarity between the abstract embedding and the scope embedding.

This results in a single score per article, where higher values indicate stronger semantic alignment with the journal's stated scope.

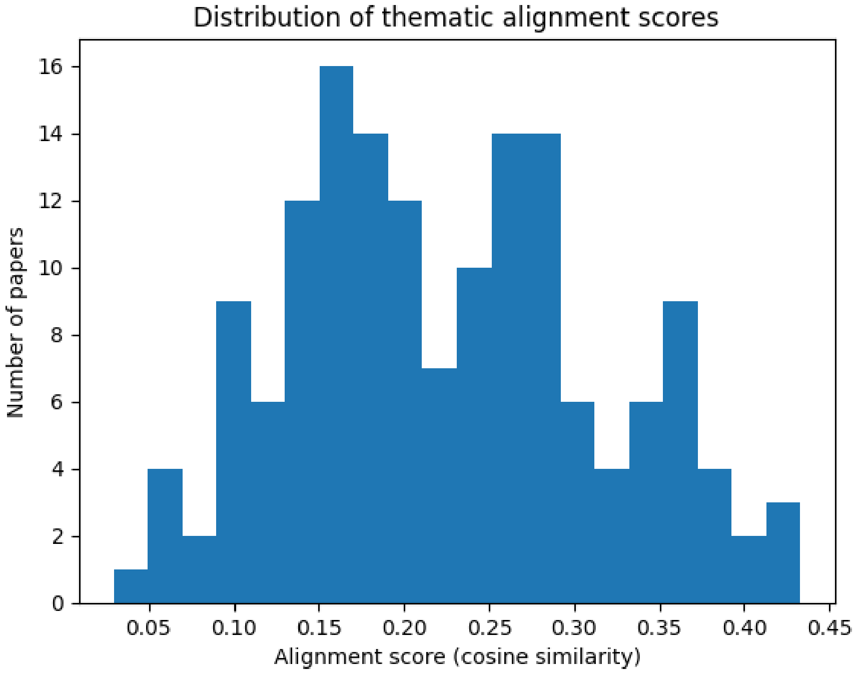

Figure 1: Distribution of thematic alignment scores between article abstracts and the journal scope using the all-MiniLM-L6-v2 embedding model.

## 3 Experimental Results

The experiments are conducted on a dataset of approximately 180 article abstracts published in the Journal of Artificial Intelligence Research between 2018 and 2026. Each data point corresponds to one article abstract and its publication year. The dataset is obtained after filtering and preprocessing steps described in Section 2.

For all three models, the obtained scores are approximately between 0 and 0.55.

Table 1: Summary statistics of alignment scores across embedding models

| Model | Count | Mean | Std | Min | 25% | Median | Max |
|---|---|---|---|---|---|---|---|
| all-MiniLM-L6-v2 | 155 | 0.226 | 0.090 | 0.029 | 0.157 | 0.221 | 0.433 |
| all-mpnet-base-v2 | 155 | 0.323 | 0.085 | 0.122 | 0.266 | 0.320 | 0.537 |
| multi-qa-MiniLM-L6-cos-v1 | 155 | 0.177 | 0.091 | -0.026 | 0.117 | 0.168 | 0.451 |

Table 1 reports summary statistics of alignment scores obtained using three different sentence embedding models. While the absolute values of the scores differ across models, all three exhibit similar patterns in terms of score distribution and variability.

The all-MiniLM-L6-v2 model yields a mean alignment score of approximately 0.23, indicating

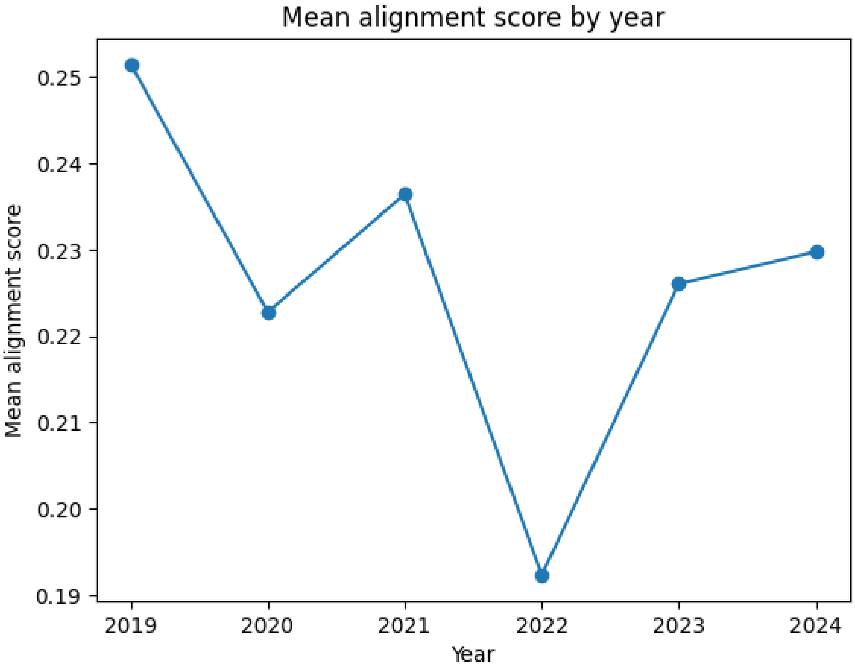

Figure 2: Mean alignment score by publication year using the all-MiniLM-L6-v2 embedding model.

moderate semantic alignment between article abstracts and the journal scope. This result reflects the broad nature of the journal, where many articles focus on specific subtopics within artificial intelligence rather than restating general AI themes.

The all-mpnet-base-v2 model produces higher alignment scores overall, with a mean of approximately 0.32. This model is known to capture broader semantic relationships, which leads to higher similarity values even when abstracts do not closely match the exact wording of the scope description. Despite the higher absolute scores, the relative distribution of alignment values remains comparable to that of the baseline model.

In contrast, the multi-qa-MiniLM-L6-cos-v1 model yields lower average alignment scores and a wider range of values. This behavior is expected, as the model is optimized for question–answer retrieval tasks rather than general semantic similarity, making it more conservative when comparing abstract text to a broad scope description.

Overall, the results show that although different embedding models operate on different similarity scales, they lead to consistent conclusions. Across all models, JAIR articles exhibit moderate thematic alignment with the journal's stated scope, with natural variation reflecting the diversity of research topics published by the journal.

### 3.1 Understanding scores by year

Figure 2 shows the mean alignment score of article abstracts with the journal scope for each publication year between 2019 and 2024. The alignment scores remain relatively stable over time, with only small fluctuations across years.

A noticeable decrease in mean alignment is observed in 2022, followed by a recovery in subsequent years. However, the overall range of variation is limited, and no strong upward or downward trend is present. This suggests that the thematic relationship between published articles and the journal's scope has remained largely consistent over time.

Similar temporal patterns are observed when using alternative embedding models, indicating that the year-level alignment trends are not strongly dependent on the choice of embedding model.

### 3.2 Qualitative Examples of Alignment

To complement the quantitative analysis, we present examples of article abstracts with high and low thematic alignment scores using the *all-MiniLM-L6-v2* embedding model. These examples help illustrate how the alignment metric relates to the content and framing of the abstracts.

**Highly Aligned Abstracts.** The following abstracts received some of the highest alignment scores. They explicitly discuss broad artificial intelligence topics such as machine learning, reasoning, planning, agents, safety, which are directly mentioned in the journal's scope.

- An article examining the attitudes of AI and machine learning researchers toward ethics, governance, safety, and military applications of AI, based on survey data collected from major AI/ML conferences.
- An essay analyzing the safety and risks of artificial intelligence systems using the framework of "normal accidents," with a focus on complexity, system interactions, and societal impact.
- A planning-focused study proposing a machine learning–based approach to reduce state uncertainty and improve agent performance in domains with incomplete information.
- An article introducing a knowledge representation and reasoning architecture for robots that combines declarative programming and probabilistic models to handle uncertainty and non-monotonic reasoning.
- A study on decentralized multi-agent path finding that analyzes conditions under which agents can reach their goals without central planning or communication.

**Low Alignment Abstracts.** The following abstracts received lower alignment scores. These articles focus on more specialized theoretical or technical problems, where the connection to the journal's broad AI scope is less explicitly stated in the abstract.

- A paper proposing improvements to t-SNE visualization by replacing the Gaussian kernel with an Isolation kernel to improve scalability and representation quality.

- A theoretical study on fair division of indivisible goods, focusing on ordinal approximations of maximin share allocations.
- A complexity-theoretic analysis of the network satisfaction problem for relation algebras, providing classification results under specific algebraic conditions.
- A theoretical contribution on risk bounds for halfspace learning using random projections in high-dimensional settings.
- A study proposing a new adversarial attack framework for time-series data based on statistical feature constraints.

These examples show that higher alignment scores are typically associated with abstracts that explicitly reference general artificial intelligence concepts and applications, while lower scores correspond to articles that emphasize narrow technical contributions or mathematical theory without explicit reference to broader AI themes.

## 4 Conclusion

This project explored the thematic alignment between articles in JAIR and the journal's stated scope using sentence embedding models. By representing article abstracts and the journal scope in a shared semantic space, we computed alignment scores that provide a overview of how published research relates to the journal's thematic focus.

The results show that JAIR articles exhibit moderate and consistent alignment with the journal scope. While alignment scores vary across articles, this variation reflects the broad and diverse nature of the journal rather than a lack of coherence. The analysis across multiple embedding models further indicates that the observed alignment patterns are stable and not strongly dependent on the choice of a specific model.

This study has several limitations. The analysis is based on abstracts rather than full article texts, which may not capture all aspects of each paper's contribution. In addition, the journal scope is represented by a single textual description, which may not fully reflect editorial interpretations over time. Finally, sentence embedding models may introduce biases based on their training data.

Future work could extend this analysis by incorporating full-text articles, comparing multiple journals, or validating alignment scores through human evaluation. Despite these limitations, this project demonstrates that modern NLP techniques can be effectively applied to analyze thematic alignment in real-world academic publishing data in a simple and interpretable way.

## AI Usage Disclosure

This project was developed with the assistance of an AI-based language model (ChatGPT). The tool was used to support brainstorming, clarify methodological choices, and assist with editing text. All implementation decisions, experimental design choices, result interpretations, and final writing decisions were made by the author. The author reviewed, adapted, and validated all content to ensure correctness and understanding.