Enabling research collaboration with AI: The BI4E experience with Large Language Models

Gianluca Amato[1], Marcello Costantini[2,3], Luca Di Vita[1], Francesca Ferri[3,4], Paolo Melchiorre[5], Maria Chiara Meo[1], Francesca Scozzari[1,*] and Matteo Vitali[1]

[1]*Laboratory of Computational Logic and AI, Department of Economic Studies, University of Chieti–Pescara, Pescara, Italy*
[2]*Department of Psychology, University of Chieti–Pescara, Chieti, Italy*
[3]*Institute for Advanced Biomedical Technologies, University of Chieti–Pescara, Chieti, Italy.*
[4]*Department of Neuroscience, Imaging and Clinical Sciences, University of Chieti–Pescara, Chieti, Italy*
[5]*Django Software Foundation (DSF), Kansas, United States*

Abstract

We present an AI-powered tool developed within the Horizon Europe-funded BI4E (Boosting INGENIUM for Excellence) project, which aims to foster new, interdisciplinary research collaborations across a university network. A key challenge addressed by BI4E is the difficulty of forming interdisciplinary and cross-university research teams, for instance for applying to competitive funding. The CONNECT platform was designed to fill this gap by providing a semantic search engine that allows researchers from different universities to discover potential collaborators on the basis of their scientific publications.

At the core of CONNECT there is a transformer-based language model, used to encode both funding call descriptions and researcher publication metadata into high-dimensional vector representations. The metadata are collected from the Scopus database using ORCID or Scopus identifiers. When queried with the topic of competitive fundings, CONNECT returns a list of researchers whose publications semantically aligns with the input query. In addition, CONNECT integrates an explainability mechanism derived from the attention matrix mechanism, allowing users to understand which parts of the input text most influenced the match.

The aim of CONNECT is to support the strategic goals of the BI4E project by offering an institutional view of research domains, facilitating industry-university matchmaking, and democratizing access to collaboration opportunities. Since its launch in December 2023, CONNECT has already enabled the formation of dozens of new research teams within the INGENIUM European University Alliance, validating its potential as a catalyst for research excellence and institutional transformation.

Keywords

Large Language Model, Natural Language Processing, Explainability, semantic embedding

1. Introduction

1.1. The BI4E project

The Boosting INGENIUM for Excellence (BI4E) project is an European initiative co-funded by the Horizon Europe program which brings together eight European higher education institutions for forming the INGENIUM European University Alliance. The mission of the alliance is to develop a new type of university that is more integrated and aligned with the strategic objectives of the European Higher Education Area. The BI4E project emerges as one of the primary vehicles of change by enhancing the competitiveness and social impact of the participating universities. The BI4E initiative tackles worldwide issues by improving institutional frameworks while developing research excellence and industrial partnerships and ensuring sustainable careers in academia.

Ital-IA 2025: 5th National Conference on Artificial Intelligence, organized by CINI, June 23-24, 2025, Trieste, Italy
*Corresponding author.
gianluca.amato@unich.it (G. Amato); marcello.costantini@unich.it (M. Costantini); lucadivita.ldv@gmail.com (L. D. Vita); francesca.ferri@unich.it (F. Ferri); paolo@melchiorre.org (P. Melchiorre); mariachiara.meo@unich.it (M. C. Meo); francesca.scozzari@unich.it (F. Scozzari); trottomv@gmail.com (M. Vitali)
0000-0002-6214-5198 (G. Amato); 0000-0002-0504-1296 (M. Costantini); 0009-0008-6274-7915 (L. D. Vita); 0000-0003-1821-5870 (F. Ferri); 0009-0003-4258-3232 (P. Melchiorre); 0000-0002-3700-3788 (M. C. Meo); 0000-0002-2105-4855 (F. Scozzari); 0009-0004-1139-1569 (M. Vitali)

1.2. Strategic objectives of BI4E

The project sets out five essential strategic goals for its implementation.

It first establishes the development of strategic planning capabilities as its main priority. The project helps universities acquire methods and tools to predict upcoming trends and evaluate new social requirements as well as match internal objectives with European and worldwide policies such as the Green Deal and United Nations Sustainable Development Goals (SDGs). The process of scenario development alongside data-based decision making and stakeholder involvement enables universities to create long-term plans which move beyond conventional operational boundaries.

Deep institutional transformation represents the second strategic objective of BI4E. The participating universities in this project receive support to develop better governance systems and modernized administrative structures as well as participatory leadership models. The enhanced operational efficiency of these institutions brings forth transparency and inclusive decision-making processes that maintain accountability within their institutional frameworks. The main priority focuses on creating flexible institutions which can rapidly modify their operations in response to environmental changes and external demands.

Research capacity development stands as the project's third strategic objective which aims to enhance research performance. The BI4E initiative identifies promising research units and helps develop them into a more productive and active research community. The alliance enables researchers to collaborate through interdisciplinary projects and enhances their publication output and visibility and enables researcher mobility between alliance institutions. The emphasis includes implementing open science practices alongside digital collaboration tools and research-teacher integration.

BI4E establishes its fourth goal by working to strengthen university-industry relations. The project establishes stronger connections between academia and private industry through improved technology transfer offices and performance-based collaborative project incentives and sustained dialogue platforms. The mission focuses on converting academic knowledge into social and economic innovation through real-world applications to bridge the research-to-practice gap.

The last objective of BI4E consists of improving academic careers through both attraction and equity principles. The initiative achieves its goals through contemporary recruitment system reform alongside financial support for beginning researchers and equal representation of men and women in academic leadership roles. Educational institutions should adopt career systems which promote equality and transparency through frameworks that support teaching and research alongside service activities. The academic institution should foster an environment that welcomes diverse talent regardless of background to achieve its ultimate goal of inclusive success.

2. Research matchmaking

One of the main challenges faced by the INGENIUM Alliance is the difficulty of identifying suitable research collaborators across different universities of different countries and the many diverse disciplines. While forming cross-university research teams is a key goal of the BI4E project, the mechanism for finding the collaborators with a given expertise in the alliance is a complex task. Standard ways for forming research teams in a single university often rely on personal contacts or institutional repository, but it is too limiting and not effective when coming to an alliance of universities. In this context, the BI4E project identified as one of its primary necessities the need for a tool to fill this gap: by exploiting the power of Large Language Models, the tool enables researchers to discover collaborations opportunities on the basis of the researchers' publications, instead of proximity or personal knowledge.

Finding experts in a given topic and team formation for research teams are both well studied fields [1, 2, 3, 4, 5, 6, 7, 8, 9], with many different approaches such as recommendation systems, often based on simple expert search, graph-based and knowledge graph approaches techniques, especially in the case of team formation and often formalized as optimization problems, and recently with embedding techniques exploiting Large Language Models (LLMs).

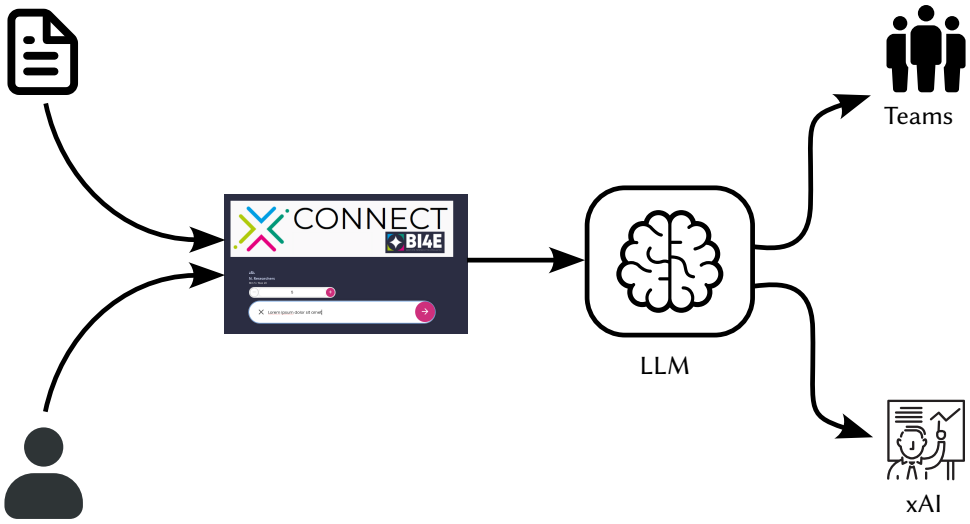

Figure 1: Interaction with CONNECT.

2.1. Interacting with CONNECT

The typical interaction flow is shown in Figure 1. It starts with the user which navigates to the web interface which contains an input form. There is no registration nor login for the search functionality. In the input field the user enters a description, in natural language, of the research topic of the funding call of interest. The use of LLM allows the system to directly process the actual text of a funding call. The input text is transformed into a vector representation which captures the semantic meaning and then it is compared against (the pre-computed embeddings of) the publications of the researchers of the universities in the INGENIUM Alliance.

When a user submits an input, the platform provides two types of results. First, it generates a list of researchers, between 5 and 20, each one presented along with a selection of their most relevant publications connected to the submitted topic. Second, an explanation of why the researcher was selected, showing the semantic match between the publications and the input query.

3. Software architecture

The software architecture of CONNECT is modular and consists of several components:

Relational database A relational database (namely, PostgreSQL) is used to store the set of authors, documents and their relations. The set of authors is populated by the local administrators of each research institution. The set of documents is populated automatically by the ETL layer.

ETL layer The ETL (Extraction-Tranform-Load) layer, using the Scopus Search API, determines the set of papers written by the authors in the relational database, and downloads all their metadata. Abstract and title for each paper are fed to the embedding model, and the result is pushed to the vector database. This is implemented trough Python classes and Django models.

Vector database A vector database (namely, Qdrant) stores, for each document, the value computed by the embedding model. The vector database allows to efficiently perform similarity search.

Frontend The frontend provides a simple and streamlined user interface for the end user and a management console for the local administrator. The UI for the end-user is implemented using React and NextJS, while the management console for the local administrator is implemented using the standard Django admin interface.

Backend The backend is written in Django and coordinates all the other layers. It also provides a RESTful API for both the frontend and third-party applications.

4. Semantic embeddings and transformer model

Our solution employs an embedding model in order to perform a semantic matching between the given research topic and the publications of the involved researchers. Among the various choices, such as for instance GloVe and Word2Vec, we adopted a transformed-based model, which uses an attention mechanism to produce the embeddings, taking into account the whole context of the sentence. In fact, transformer model (see, e.g., BERT [10]) proved to be effective in capturing semantic relations in scientific publications, especially when compared to traditional keyword-based search approaches. The important point here is the ability of the model to produce a high dimensional representation of the text, which in turn can model a query and the publications in a common semantic space. This is particular relevant since it allows us to get rid of synonyms, paraphrasing and interdisciplinary topics, where exact matching based on keywords often fails.

We extract from the publications of the researchers of the alliance universities the title and the abstract, which are encoded in a pre-trained transformer model. These two elements can provide a concise yet precise summary of the publications, capturing both the context and the research focus. The aggregation of the embeddings of all the publications of a researcher forms a semantic representation of his expertise. The query are encoded in the same way. Publications metadata (title and abstract) are extracted from the Scopus repository, by means of ORCID or Scopus identifiers, which are the input to initialize the whole system.

4.1. The LLM model

We use the `all-MiniLM-L6-v2` model [11], a transformer-based sentence embedding model. This is a compact and efficient variant of BERT, specifically designed for sentence similarity tasks. The model has only six transformer layers and produces an embedding with 384 dimensions, which allows us to maintain scalability across a large number of documents.

Once the input query is transformed into a vector, CONNECT performs a nearest neighbor search against a collection of publication vectors precomputed from the metadata (title and abstract). The similarity between vectors is often measured using metrics such as cosine similarity or Euclidean distance. We employed cosine similarity as it is widely recommended for textual data since it focuses on the orientation of vectors rather than their magnitude. This property makes it particularly suitable for NLP tasks, as it effectively captures semantic similarity while being invariant to text length and word frequency variations.

The vector search is performed in the Qdrant vector database and ranks researchers based on cosine similarity to the query embedding. As in any transformer model, the fundamental object here is the attention matrix [12], whose role is to determine which tokens influence each other in a sequence. This is accomplished by assigning attention scores at a sentence level, instead of treating words independently, which in turn can generate context-aware embeddings. This is the key mechanisms which ensures that the search is performed at a semantic level, which consider the context and is able to recognize different meanings of the same word.

In addition, CONNECT enhances transparency through an explainability layer based on the model attention mechanism. Attention matrices highlight the set of tokens in the input that influenced the result, allowing users to understand the reason behind the suggestions and reinforcing trust in the AI result. This interpretable output is especially useful in interdisciplinary contexts where users may be unfamiliar with the matched researcher's domain.

5. Software instance

CONNECT is an open-source software, distributed under the AGPL v3 license. The full source code is publicly available on GitLab at https://gitlab.com/uda-connect/, allowing researchers to freely inspect the code and contribute to its development.

One instance of the software is currently running at the University of Chieti–Pescara, accessible at https://connect.unich.it/. It contains the data relative to the institutions belonging to the INGENIUM Alliance.

6. Conclusions

The CONNECT platform represents a key technological innovation within the BI4E project, serving as a digital transformation solution in higher education aligned with the BI4E principles. It enables interdisciplinary research collaboration among the members of the INGENIUM Alliance. The platform operates by processing researcher publication metadata extracted from Scopus to build semantic researcher profiles. This allows it to create intelligent connections between researchers who would not typically collaborate.

The CONNECT platform directly supports all five strategic goals of the BI4E project in concrete ways. It provides universities with an interactive overview of their research domains as part of their strategic planning efforts. Institutional leaders can trace collaborative networks while following emerging themes to make strategic investment decisions. Moreover, it introduces simplified administrative workflows through which institutions can manage partnership development and research coordination thus releasing resources toward innovation and strategic planning.

6.1. Software adoption

CONNECT already proved to bring revolutionary changes to research performance. Launched in December 2023, it has already stimulated the creation of more than 40 new research collaborations between alliance members, accelerating the knowledge creation process by simplifying for researchers from different countries and disciplines to find collaborators, especially for funding calls.

The platform also serves as a tool that helps universities to connect with industry partners, providing external stakeholders, such as private companies and non-profit organizations, with a user-friendly searchable interface to quickly discover academic partners with relevant expertise.

Finally, CONNECT contributes to more inclusive and equitable academic careers. It allows early-career researchers to gain more visibility and access international collaborations regardless of institutional prestige or geographic location. By focusing on the relevance of the researchers' scientific publications rather than personal networks, CONNECT aligns closely with the BI4E goal of building a more open, transparent, and inclusive European academic landscape.

6.2. Future work

We plan to improve CONNECT in several ways:

- adding filtering options for the end-user, such as filtering on the researcher home institution and/or on the ERC research sector of the authors;
- adding sources of information other than Scopus, such as Clarivate Web of Science, in order to cover a broader range of disciplines, especially in fields which are underrepresented in Scopus;
- using static analysis techniques [13, 14] to formally verify the correctness of the code [15].

The aim is to keep CONNECT always useful and up to date, while shepherding its adoption also in other European University alliances beyond INGENIUM. The ultimate goal is to promote international collaboration and research excellence, aspiring to become a replicable model of how artificial intelligence can be effectively applied to promote strategic collaboration among different research institutions.

Conflict of Interest

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

Acknowledgments

This work was partially supported by Boosting INGENIUM for Excellence (BI4E) project, funded by the European Union's HORIZON-WIDERA-2021-ACCESS-05-01-European Excellence Initiative under the Grant Agreement No. 101071321; the PNRR project FAIR — Future AI Research (PE00000013), Spoke 9 – Green-aware AI, under the NRRP MUR program funded by the NextGenerationEU; the GNCS (Gruppo Nazionale di Calcolo Simbolico) group of the Istituto Nazionale di Alta Matematica — INdAM.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, Z. Su, Arnetminer: Extraction and mining of academic social networks, in: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2008, p. 990 – 998. doi:10.1145/1401890.1402008.

[2] K. Balog, Y. Fang, M. De Rijke, P. Serdyukov, L. Si, Expertise retrieval, Foundations and Trends in Information Retrieval 6 (2012) 127 – 256. doi:10.1561/1500000024.

[3] J. Tang, S. Wu, J. Sun, H. Su, Cross-domain collaboration recommendation, in: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2012, p. 1285 – 1293. doi:10.1145/2339530.2339730.

[4] F. Xia, Z. Chen, W. Wang, J. Li, L. T. Yang, Mvcwalker: Random walk-based most valuable collaborators recommendation exploiting academic factors, IEEE Transactions on Emerging Topics in Computing 2 (2014) 364 – 375. doi:10.1109/TETC.2014.2356505.

[5] C. Xiong, R. Power, J. Callan, Explicit semantic ranking for academic search via knowledge graph embedding, in: 26th International World Wide Web Conference, WWW 2017, 2017, p. 1271 – 1279. doi:10.1145/3038912.3052558.

[6] H. Wan, Y. Zhang, J. Zhang, J. Tang, Aminer: Search and mining of academic social networks, Data Intelligence 1 (2019) 58 – 76. doi:10.1162/dint_a_00006.

[7] K. Golzadeh, L. Golab, J. Szlichta, Explaining expert search and team formation systems with exes, 2024. URL: https://arxiv.org/abs/2405.12881. arXiv:2405.12881.

[8] X. Kong, Y. Shi, S. Yu, J. Liu, F. Xia, Academic social networks: Modeling, analysis, mining and applications, Journal of Network and Computer Applications 132 (2019) 86 – 103. doi:10.1016/j.jnca.2019.01.029.

[9] X. Su, S. Xue, F. Liu, J. Wu, J. Yang, C. Zhou, W. Hu, C. Paris, S. Nepal, D. Jin, Q. Z. Sheng, P. S. Yu, A comprehensive survey on community detection with deep learning, IEEE Transactions on Neural Networks and Learning Systems 35 (2024) 4682 – 4702. doi:10.1109/TNNLS.2021.3137396.

[10] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.

[11] N. Reimers, all-minilm-l6-v2, https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2, 2021.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017.

[13] G. Amato, M. Rubino, F. Scozzari, Inferring linear invariants with parallelotopes, Science of Computer Programming 148 (2017) 161–188. doi:10.1016/j.scico.2017.05.011.

[14] G. Amato, M. C. Meo, F. Scozzari, On collecting semantics for program analysis, Theoretical Computer Science 823 (2020) 1–25. doi:10.1016/j.tcs.2020.02.021.

[15] G. Amato, F. Scozzari, The ScalaFix equation solver, in: M. Chechik, J.-P. Katoen, M. Leucker (Eds.), Formal Methods, 25th International Symposium, FM 2023, Lübeck, Germany, Proceedings, volume 14000 of *LNCS*, 2023, pp. 142–159. doi:10.1007/978-3-031-27481-7_10.