9th Asia-Pacific Conference on Plasma Physics, 21-26 Sep, 2025 at Fukuoka

# Foundation Models for Structured Knowledge and Predictive Modelling in Nuclear Fusion Research

A. Loreti[1,2], K. Chen[2], R. George[2], R. Firth[2], A. Agnello[2], S. Tanaka[3]
T. Boschi[4], R. Ordonez-Hurtado[4], C. Rousseau[4], A. Pascale[4], M. Zayats[4],
N. Amorisco[1], S. Jackson[1], S. Pamela[1].

[1]United Kingdom Atomic Energy Authority (UKAEA), Culham Centre for Fusion Energy, Culham Science Centre, Abingdon OX14 3EB, [2] STFC Hartree Centre, Daresbury, [3]WA4 4AD, UK, IBM Research, STFC Hartree Centre, [4]Trinity Business School, Trinity College, 182 Pearse Street, Dublin 2, D02 F6N2, Ireland.

e-mail andrea.loreti@ukaea.uk

The increasing complexity and heterogeneity of data in nuclear fusion research demands novel approaches for knowledge representation and predictive modelling. In this work, we explore the application of AI models, particularly foundation models (FMs), to two different challenges in the field of nuclear fusion research:

(1) Automated construction of a Knowledge Graph (KG) of nuclear fusion energy for effective elicitation and retrieval of information.
(2) The development of a data-driven framework for forecasting plasma states in tokamaks.

For the first task, we present a multi-step approach for structuring and representing domain-specific knowledge derived from large corpora of documents [1]. We apply our method to build the first KG of nuclear fusion energy. In the graph, knowledge is stored in nodes (entities) linked together by edges (relationships).

The heterogeneity and vast scope of the nuclear fusion research field make it an ideal benchmark to test the key features of our automated pipeline, including named entity recognition and entity resolution. We show how pre-trained Large Language Models (LLMs) can be used to address these challenges, and we evaluate their performance against Zipf's law, which characterizes human-generated natural language. The KG supports Retrieval-Augmented Generation (RAG) using a multi-prompt strategy to answer complex, multi-hop scientific queries. We demonstrate how KG-RAG architectures extend beyond standard RAG systems by introducing an additional layer of abstraction during the answer generation. For instance, the tasks of summarizing the content of the whole KG or extracting information within a specific time range require multi-hop capabilities to retrieve information that goes beyond the content of individual documents. KG-RAG enables this level of abstraction by leveraging the entities and relationships stored in the graph.

For the second task, we discuss the development of a predictive modelling framework aimed at forecasting plasma states in tokamak-like devices.

In this study, we use the FAIR-MAST data service [2] an open dataset of historical diagnostic measurements from the Mega Ampere Spherical Tokamak (MAST) experiment.

Our framework is developed to address several challenges associated with the dataset such as:

(1) sparsity: the model must cope with missing data and sensors failures.

(2) multi-modality: the model should be able to integrate different kinds of signals, e.g., 1D time-series, N-dimensional arrays and videos in its predictive framework.

We propose to use existing pre-trained foundation models and adapting them to our specific use case in plasma modelling. Additionally, we plan to explore variational auto encoder architectures to embed plasma features into a latent space of reduced dimensionality. These latent representations will serve as basis for generating predictions across a range of plasma scenarios in downstream tasks of interest, e.g., [3].

Together, these contributions demonstrate the potential of foundation models to bridge structured knowledge representation and predictive analytics in fusion science.

[1] A. Loreti et al., arXiv:2504.07738.

[2] S. Jackson et al., SoftwareX, **27**, 101869, (2024).

[3] Y. Wei et al., Nuclear Fusion, **61**, 12, 126063, (2012).