Autoregressive Transformers for Disruption Prediction in Nuclear Fusion Plasmas

Lucas Spangher
Plasma Science and Fusion Center
Massachusetts Institute of Technology
spangher@psfc.mit.edu

William F. Arnold
Kim Jaechul School of AI
KAIST
will@mli.kaist.ac.kr

Alexander Spangher
Information Sciences Institute
University of Southern California

Andrew Maris
Plasma Science and Fusion Center
Massachusetts Institute of Technology

Cristina Rea
Plasma Science and Fusion Center
Massachusetts Institute of Technology
crea@psfc.mit.edu

Abstract

The physical sciences require models tailored to specific nuances of different dynamics. In this work, we study outcome predictions in nuclear fusion tokamaks, where a major challenge are *disruptions*, or the loss of plasma stability with damaging implications for the tokamak. Although disruptions are difficult to model using physical simulations, machine learning (ML) models have shown promise in predicting these phenomena. Here, we first study several variations on masked autoregressive transformers, achieving an average of 5% increase in Area Under the Receiving Operating Characteristic metric above existing methods. We then compare transformer models to limited context neural networks in order to shed light on the "memory" of plasma effected by tokamaks controls. With these model comparisons, we argue for the persistence of a memory throughout the plasma *in the context of tokamaks* that our model exploits.

1   Introduction

Nuclear fusion has been long sought as a promising method of generating carbon-free energy without major land-use requirements or high-level radioactive waste. One major approach to obtaining fusion in laboratory plasmas is through *magnetic confinement*, which leverages the tendency of plasmas to remain confined perpendicular to the magnetic field lines. The fusion approach with the highest technological readiness level Wurzel and Hsu [2021] is the *tokamak*, a toroidal chamber with external magnetic coils and a large current that runs toroidally through the plasma with the goal of generating net positive energy (see Figure 1.)

A major problem facing fusion tokamaks are *disruptions*, or loss of plasma stability with damaging implications for the tokamak. Disruptions are difficult to model using physical simulations due in part to the heterogeneity of causes that can lead to instability, the multi-scale nature of plasma dynamics, and to a wealth of unobserved factors that affect known physical dynamics. Thus, there is an emerging literature on methods other than "first principle models" for disruption prediction.

The goal in this prediction problem is to provide sufficient time to engage control actions or emergency mitigation systems, which curb the deleterious consequences of disruptions. Numerous ML models have been proposed to predict the likelihood that a plasma disruption will occur Murari and Contributors [2013], Kates-Harbeck et al. [2019]. Prior work has investigated the use of random forest models to monitor plasma states Rea et al. atlargest tokamak in the US, DIII-D, followed by current state of the art predictors, such as the Hybrid Deep Learner (HDL) proposed in 2021, which utilize classical neural techniques such as a convolutional LSTM network Zhu et al. [2020].

However, none of these models have the desired accuracy for commercial scale tokamaks where disruption prediction accuracy must be nearly flawless (>95% true positive rate) Maris et al. [2023]. Also, we hypothesize, none are tuned to the specific needs of the problem. It is unclear how long the "memory" of the plasma is – while some argue that as little as 200 milliseconds of history is needed to capture the evolution of the plasma temperature and density Zhu et al. [2020], Rea et al., others argue that certain instabilities or tokamak control errors that occur near the start of a seconds-long shot may impact the end behavior of the shotsMurari and Contributors [2013].

In this work, we apply variations on the Transformer architecture Vaswani et al. [2017], Bengio et al. [2009], which has become popular in recent years to model long term, event-based time-series. Our two contributions are:

1. We create a state-of-the-art model on a data from three major tokamaks. We find an average of .05 improvement in Area under the Receiver Operating Characteristic Curve (AUC) over the HDL baseline (Table 1.) We find various pretraining steps help augment performance.

2. We investigate whether long term memory is helpful in modeling the fusion reaction, giving us scientific insight into the dynamics of the process.

2   Dataset Provenance and Preparation

Our dataset is composed of trials from three different tokamaks, r_1, r_2 and r_3.[1] Each r_i differs considerably[2]. Each tokamak is run continuously for a "shot" sequence lasting seconds to minutes, generating gigabytes of multi-modal data, of which we focus on global state variables (the rest include 1D states ("profiles"), 2D cross-sections, and camera photos); we focus on shots' global state variables, listed in Table 3. Our objective is to understand how well this model would deploy to a new tokamak, which is similar to r_1. Thus, we restrict our test data to data generated by r_1. We build off previous work from our lab for the training set composition, which contains different proportions of r_1, r_2, and r_3 data [citation withheld].

Each shot is of length n, and we discard shots that are shorter than 125ms. Each tokamak's diagnostics store metrics from each shot at a different sampling rate. We normalize across tokamaks by discretizing at 5ms and interpolating via forward-fill. Roughly 10% of shots end in disruptions. We truncate our time series to $X = 1, ..., n - \nu$, where n is the full length and $\nu = 40ms$, the minimum amount of time for a disruption mitigation system to activate. For a description of the features chosen, please see Table 2 in the Appendix.

3   Model Details

3.1   Modeling Task

We seek to perform label prediction (Figure 1), where X is described above, and y is 1 if the shot ended in disruption and 0 otherwise. We follow Zhu et al. [2020] in several augmentations. First, we generate additional pairs $(\overline{X}, \overline{y})$ from each (X, y) by randomly sampled starting and ending windows of the shot. Next, in order enforce greater reasoning about the start of instabilities, we set a τ_{r_i} variable per tokamak[3]: if $(\overline{X}, \overline{y})$ of length n ends before $n - \tau_{r_i}$, it is labelled 0.

---

[1] r_1 is MIT's Alcator C-Mod, r_2 is General Atomic's DIII-D, and r_3 is the Chinese Academy of Science's EAST tokamak

[2] For a longer discussion on the physical differences between the three, see Granetz et al. [2019]

[3] $\tau(r_i)$ is the average length of time before a disruption occurs which is tokamak-dependent, see Table 3
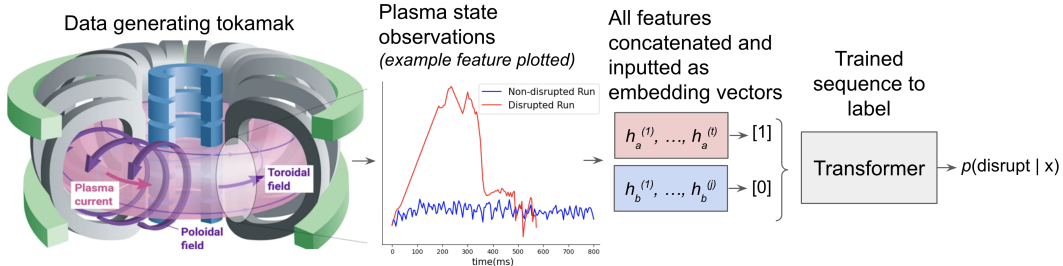
Figure 1: Sequence to Label classification task using GPT2-like transformers.

3.2 Model Architecture

We adapt an autoregressive transformer based on GPT[4] Radford et al. [2018]. GPT-based models, which are typically used to model long sequences of discrete word-tokens, have an embedding layer that is used to project discrete tokens into a continuous embedding space. The dimensions of our data are already continuous variables, so we treat them as embeddings directly. These vectors are normalized, based on training-set mean and variance.

We perform a hyperparameter sweep using a bayesian hyperop strategy Wu et al. [2019] over number of layers, dropout proportions, learning rate schedulers, optimizer β parameters, and other hyperparameters of interest. All training was done on a cluster of 48 A100 Nvidia GPUs and tested on Area Under Receiver Operator Characteristc (AUC).

3.3 Model Augmentations: State Pretraining and Data-cutoff Currica Induction

State Pretraining: We pretrain the model first on a next state prediction (NSP) task. We model NSP with a regression head on the model and Mean Average Error (MAE) loss. After convergence, which is observed to be 50k steps, we replace the regression head with a classification head and perform our sequence classification task.

Data-cutoff Curriculum Learning: We notice the model often misclassifies samples with rapidly evolving end dynamics. We theorize that exposing the model to the full data, without the 40ms cutoff, might improve its predictions for the final 40ms. Using *curriculum learning* Bengio et al. [2009], we sequentially train the model on easier tasks: predicting datasets with cutoffs from 0ms to 30ms, in 10ms increments, for 50k steps each. We then test the model on the main task. For example, a 3 step curriculum trains the model on data with 20ms cutoff, then 30ms, then 40. We experiment with varying curriculum lengths from zero to five steps.

4 Results and Discussion

4.1 Performance metrics

Compared to the baseline HDL predictor, our vanilla model achieves favorable AUC across all cases, and our augmented model improves further on the vanilla case (Table 1.) Wall-clock time on forward inference ranges from 1-3ms, which is suitable for real-time deployment.

4.2 Curriculum Training and State Pretraining

Curriculum Learning: We observe, interestingly, that the model achieves the highest accuracy, recall, and precision ($\sim 10\%$ gain in each) following a curriculum in which the models start training 10ms from shots' ends and progress successively – i.e. four curriculum steps. The result matches physical intuition: at the very end of disruptive shots, the plasma has already touched the tokamak's walls, and thus the physics of the fluid have changed.

---

[4]We utilize Huggingface's implementation: `https://github.com/huggingface/transformers/blob/main/src/transformers/models/gpt2/modeling_gpt2.py`

| | Training Set Composition | | | | AUCs | | |
| | Existing: r^2, r^3 | | "New": r^1 | | HDL | GPT | Aug-GPT |
Case	Non-disruptions (ND)	Disruptions (D)	ND	D			
1	None	None	All	20	.640	.781	**.790**
2	None	None	All	All	.799	.836	**.843**
3	None	All	All	All	.808	.831	**.841**
4	All	All	None	None	.588	**.622**	.621
5	All	All	All	20	.703	.699	**.704**
6	All	All	All	All	.794	.803	**.814**
Mean					.72	.76	**.77**

Table 1: A table containing different training cases and output metrics between our transformer and HDL, the baseline. Ours improves prior SOTA by 5% AUC on average. Please note that "GPT" refers to the vanilla GPT, and "Aug-GPT" refers to the augmented GPT.

State pretraining: we observe that model performance consistently improves. We also observe that the transformer model seems to be a very good state predictor, which may be a side benefit of the model for other uses.

4.3   Long Term Memory

We argue that plasma time-series exhibit long-term memory through three ways:

1. A **comparison of model performance** finds the vanilla GPT significantly outperforming the baseline HDL, which explicitly encodes limited memory.

2. **Ablations on data context** show that model performance decreases as available context decreases from 480ms to 80ms, indicating that the model may be using data context from throughout the shot (Figure 3.)

3. **Global attention maps**, which explicitly encode regions of the data to which the model pays attention, does not tend towards zero even when the whole shot is observed (Figure 3)

4.4   Model interpretation.

Generally we may interpret characteristics of the data that the transformer picks out. First, we find the attention spikes maps to loop voltage spikes, of which higher voltage generally correlates to increased system pressure. Second, we find that shaded regions of the attention map correspond to an uptick in β_p and a drop in Li. These characteristics are consistent with those of a cooling plasma prone for edge instabilities Zohm [1996]. Both of these observations give confidence that the transformer is picking up on known physical events.

5   Conclusion and Future Work

Transformers are promising models for the physical sciences. We present a case where of transformers, as well as augmentations, achieve better performance than baseline. We demonstrate that the model picks up on some physics, and benefits from explicit memory of the system.

One shortcoming of the model is that it is optimized for discrete time series, like sentences, where words are some fundamental unit of the process. Ours, meanwhile, is a time series characterizing a dynamic process, where sampling rate is relatively arbitrary and may change, not only between tokamaks, but between diagnostics within a single tokamak. While impressive that the transformer can learn attention on discrete spikes in v_{loop}, we ultimately desire a model that may capture continuous patterns and subtle variations. Thus in future work, we hope to evaluate the performance of an Autoformer Wu et al. [2022], which is a transformer whose discrete attention blocks are replaced with autocorrelational FFT decompositions, making it more suited for continuous sequences.
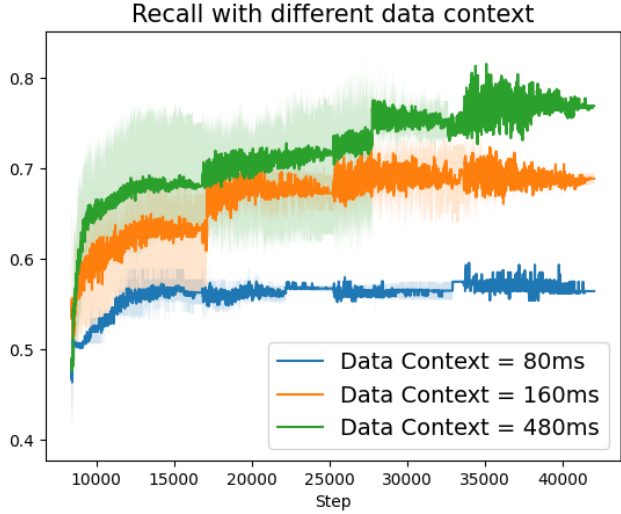
Figure 2: Exploration of transformer behavior. Here, we ablate on limiting data context to test whether performance is directly effected by physical memory.
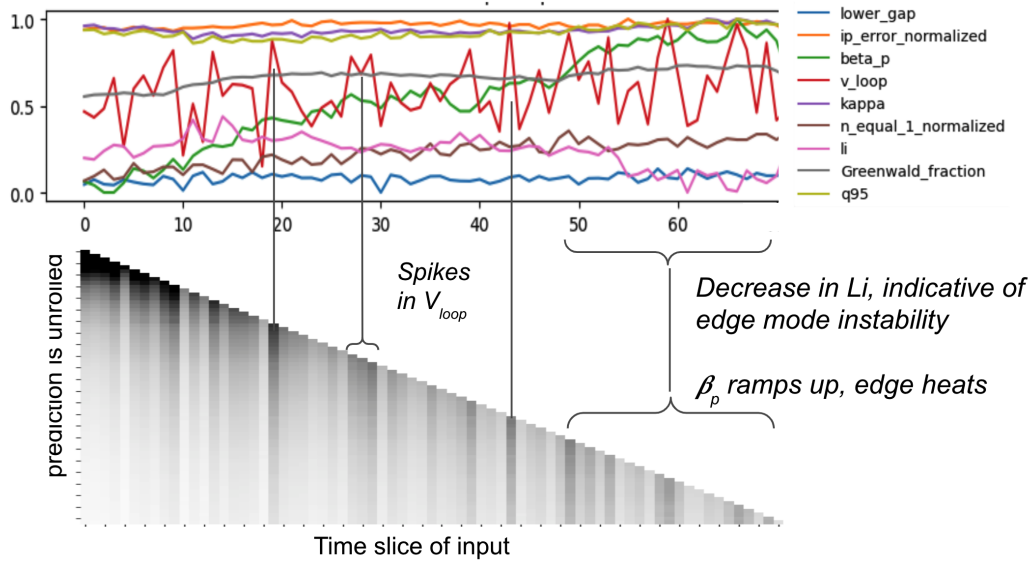

Figure 3: Here we show a further exploration of transformer behavior. In the top plot, we show normalized input features, corresponding to the transformer's attention map below. To read the attention map, an unrolling sequence starts at the top left and then progresses by traversing down.

Acknowledgments and Disclosure of Funding

References

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009.

AJ Creely, Martin J Greenwald, Sean B Ballinger, D Brunner, J Canik, Jeffrey Doody, T Fülöp, DT Garnier, R Granetz, TK Gray, et al. Overview of the sparc tokamak. *Journal of Plasma Physics*, 86(5):865860502, 2020.

MJ Edwards, PK Patel, JD Lindl, LJ Atherton, SH Glenzer, SW Haan, JD Kilkenny, OL Landen, EI Moses, A Nikroo, et al. Progress towards ignition on the national ignition facility. *Physics of Plasmas*, 20(7), 2013.

RS Granetz, N Eidietis, OM Meneghini, DL Chen, BJ Xiao, et al. Machine learning for disruption warnings on alcator c-mod, diii-d, and east. *Nuclear Fusion*, 59(9):096015, 2019.

Julian Kates-Harbeck, Alexey Svyatkovskiy, and William Tang. Predicting disruptive instabilities in controlled fusion plasmas through deep learning. *Nature*, 568(7753):526–531, 2019.

Andrew D Maris, Allen Wang, Cristina Rea, Robert Granetz, and Earl Marmar. The impact of disruptions on the economics of a tokamak power plant. *Fusion Science and Technology*, pages 1–17, 2023.

A Murari and JET EFDA Contributors. Automatic disruption classification based on manifold learning for real-time applications on jet. *Nuclear Fusion*, 53(9):093023, 2013.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

C. Rea, K. J. Montes, K. G. Erickson, R. S. Granetz, and R. A. Tinguely. A real-time machine learning-based disruption predictor in DIII-D. 59(9):096016. ISSN 0029-5515. doi: 10.1088/1741-4326/ab28bf. URL `https://dx.doi.org/10.1088/1741-4326/ab28bf`.

Lucas Spangher, J Scott Vitter, and Ryan Umstattd. Characterizing fusion market entry via an agent-based power plant fleet model. *Energy Strategy Reviews*, 26:100404, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting, 2022.

Jia Wu, Xiu-Yun Chen, Hao Zhang, Li-Dong Xiong, Hang Lei, and Si-Hao Deng. Hyperparameter optimization for machine learning models based on bayesian optimization. *Journal of Electronic Science and Technology*, 17(1):26–40, 2019.

Samuel E. Wurzel and Scott C. Hsu. Progress toward fusion energy breakeven and gain as measured against the lawson criterion. 2021. URL `https://api.semanticscholar.org/CorpusID:235166318`.

J. X. Zhu, C. Rea, K. Montes, R. S. Granetz, R. Sweeney, and R. A. Tinguely. Hybrid deep-learning architecture for general disruption prediction across multiple tokamaks. *Nuclear Fusion*, 61(2):026007, December 2020. ISSN 0029-5515. doi: 10.1088/1741-4326/abc664.

Hartmut Zohm. Edge localized modes (elms). *Plasma Physics and Controlled Fusion*, 38(2):105, 1996.

A Data Feature Decsription

Feature	Definition	Relevant instab.
Locked mode indicator	Locked mode mag. field normalized to toroidal field	MHD
Rotating mode indicator	Std. dev. of Mirnov array normalized by toroidal field	MHD
β_p	Plasma pressure normalized by poloidal magnetic field	MHD
ℓ_i	Normalized plasma internal inductance	MHD
q_{95}	Safety factor at 95th flux surface	MHD
n/n_G	Electron density normalized by Greenwald density	Dens. limit
Δz_{center}	Vertical position error of plasma current centroid	Vert. Stab.
Δz_{lower}	Gap between plasma and lower divertor	Shaping
κ	Plasma elongation	Shaping
$P_{\text{rad}}/P_{\text{input}}$	Radiated power normalized by input power	Impurities
$I_{\text{p,error}}/I_{\text{p,prog}}$	Plasma current error normalized by programmed current	Impurities
V_{loop}	Toroidal "loop" voltage	Impurities

Table 2: The input features of the model, their definitions, and a categorization of the type of instability the signal indicates.

B Further Experimental Details

tokamak	τ	Number of Shots	Average Shot Length	Initial Sampling Rate
C-Mod	50 ms	4000	.52s	.005ms
DIII-D	150 ms	8000	3.7s	.01ms
EAST	400 ms	11000	5.3s	.025ms

Table 3: Metrics on a dataset composed of multiple tokamaks.

C Supplemental Background on Fusion Devices and Plasma Dynamics

The two major approaches to generation are *inertial confinement*, which uses lasers or motion to heat up plasmaEdwards et al. [2013], and *magnetic confinement*, which leverages extreme magnetic fields to pressurize gas into plasma, and then pressurize plasma into "burning plasma", i.e. ones that undergo nuclear fusion Creely et al. [2020].

Tokamaks exert magnetic fields of roughly 5 Tesla in a toroidal vacuum chamber filled with a gas of hydrogren, deuterium and tritium isoptopes[5]. Under extreme conditions, this gas enters a plasma state, which is characterized by the highly charged flow of subatomic particles (ions and electrons.) Tokamaks achieve some degree of magnetic-field symmetry, so tokamak plasmas are stable under certain conditions. However, magnetohydrodynamic (MHD) instabilities may still arise. For example, a majority of electrons may slow in toroidal rotations relative to poloidal rotations, creating areas where the induced magnetic force of the plasma pressures certain points of the toroid and not others. Left untreated, instabilities like this may grow out of control and lead to a loss of plasma confinement, causing an end to the reaction and a fast quench of the temperature and current onto the walls of the tokamak. Given that the generated plasma has a core temperature on par with the center of the sun, this may be a serious problem, resulting in damage to the walls and downtime, making plasma disruptions a fundamental hurdle facing tokamak commercializability Spangher et al. [2019].

---

[5]Deuterium and tritium isotopes are hydrogen atoms that have one and two extra neutrons, respectively