Stabilizing Multi-Nation AI Governance Through Finding a Stable Stochastic Differential N-Player Game between Intermediate Representations of Causeformers using Causal Optimal Transport, Vapnik Chervonenkis and Ergodic Theory

Marvin Koß
Heidelberg University

May 25, 2025

Abstract

The degree of stability of our world hinges on stable competition between systems at multiple scales. We expect a future of LLM governed Nationstates, each governing its subjugate's narratives and control. We should therefore ensure communication between their internal representations through a communication protocol between governing LLMs.

To this end, our first aim is to mathematically design a stable Stochastic Differential N-Player Game between the intermediate representations of neural ordinary differential equation based formulations of Transformers that respect causality.

The second step is to design a practical implementation in which the interacting models communicate intermediate activations.

1 Introduction

In lower saxony there is the old adage "Schiet schickt hin un Schiet krecht wedder."[1], while in German generally there is "Wie man in den Wald hineinruft, so schallt es heraus."[2]. We believe this country lore to be a recommendation for how to behave in an iterated prisoner's dilemma, and indeed the right perspective with which to set up a future world in which we cohabit peacefully with not just AI, but one another. In this view, this research project represents an attempt to design a stable game between governing LLMs, before an unstable game emerges. We turn language-model governance into a provably stable international system. The stack has two main components:

1. Causeformer: A Neural Ordinary Differential Equation [3] based Transformer whose doubly-stochastic attention is hard-masked by a learnable DAG. In the continuous-depth limit, we expect it obeys a heat equation on the Graph Laplacian, so every layer performs a smooth, entropy-increasing diffusion along causal edges.

2. A diplomatic N-Player Game: Each nation-state runs a Causeformer; their latent representations interact through causal optimal transport couplings. Standard existence and uniqueness results for Stochastic Differential N-Player Games [2] hopefully guarantee a unique equilibrium when the coupling cost is monotone. The G-causal Wasserstein distance of [4] provides exactly that property and keeps average treatment effects Lipschitz-stable.

[1] If you send shit, you will get shit in return.
[2] How you call into the forest, so it echoes back.

2 Mathematical Modelling

For the various hypotheses we aim to test and systems we want to build, we make mathematical choices that we expect to interlock well for holistic analyses. In order to be able to talk about the gradient of entropy, we opt for the differential equation based NODE framework. We use Sinkformers in order to have a NODE that is compatible with the optimal transport framework we use to treat causality and abstraction. We finally use Stochastic Differential Game Theory to handle the continuously modelled layers.

2.1 Causal Abstraction

Theory for the phenomenon of grouping instances with causal interactions into concepts at multiple scales has to be developed. DAGs over instances at a given abstraction level frequently become cyclic at the next level as instances are conceptualised. Further, ensembles of instances persist across time, thus having multiple interactions with neighboring nodes (across time and the ensemble), in effect smearing out clear causal directionality until one is left with mere correlation, i.e. as one goes higher up in an ontology, one sacrifices causal clarity for abstraction power. The abstraction difficulty could be treated with Vapnik-Chervonenkis Theory, while the degree of difficulty with keeping track of causal interactions of instances across time should be treatable with ergodic theory. We expect we can prove PAC bounds on the loss in causal accuracy when abstracting using Hoeffding's inequality or similar.

2.2 Causal Optimal Transport

The language data LLMs are trained on are time series of tokens generated by latent processes in our world, namely communications of human agents interacting within games across scales. We investigate what type of causal structures of interactions within which hierarchies of games (that is, narratives) can be inferred using Sinkformers with causal masking using the framework in [4]. More specifically, we make use of their Theorem 3.4 to differentiably check for G-causal coupling between layers (abstraction levels), Proposition 3.6 for convergence of masked sinkhorn attention, and mainly Theorem 1 for the causal heat equation, i.e. diffusion along causal edges.

2.3 Causeformers

Sinkformers [8] are a neural ordinary differential equation based variant of transformers more amenable to the above OT treatment than vanilla transformers, proven to result in a gradient flow in Wasserstein space by the original authors. We extend their analysis to include the multilayer perceptrons and causal masking, ultimately resulting in a practical implementation similar to that of [9].

The original sinkformer flow is shown to be governed by the heat equation by the original authors, i.e. [8, Theorem 1]

$$\partial_t \rho = \Delta \rho \quad \big(= \operatorname{div}(\rho \nabla \log \rho) = \nabla^{\mathrm{nat}} \mathcal{H}(\rho) \big) \tag{1}$$

lending credence to the fact that entropy increases over time. After adding in the causal mask used in language models, we expect the continuous depth limit to be, similarly,

$$\partial_t \rho = \operatorname{div}_G(\rho \nabla_G \log \rho) =: \Delta_G \rho, \tag{2}$$

where G is the causal mask DAG and Δ_G its Graph Laplacian. To see this, we define a masked cost

$$c_G(x_i, x_j) := \begin{cases} (W_Q x_i)^T W_K x_j & \text{if } G_{ij} = 1 \\ -\infty & \text{otherwise.} \end{cases}$$

Eq. 2 then follows straightforwardly through similar reasoning as in the proof of [8, Theorem 1]. This simple model therefore induces a heat flow constrained to the causal skeleton.
We deepen our analysis by investigating the spectral gap of $\Delta_G = D - A$[3], namely its second smallest eigenvalue and thus hope to obtain an exponential convergence speed to the limit measure of the form

$$\|\rho_t - \rho_\infty\|_{L^2} \leq \exp(-\lambda_2 t)\|\rho_0 - \rho_\infty\|_{L^2}$$

which is relevant to the coordination problem described in the next section.
In a further step towards the typical transformer architecture used in language modelling, we use piecewise differential equatios to add the MLPs as

$$\dot{\rho}_t = \begin{cases} f_{attn}(\rho_t, \tau), & t \in [l, l + \frac{1}{2}) \\ f_{mlp}(\rho_t), & t \in [l + \frac{1}{2}, l + 1), \end{cases}$$

for integer transformer block indices l. We call the function given by the piecewise ones f_{hybrid}. Hybrid-ODE well-posedness holds since each field is Lipschitz and we have finite blocks.
A preliminary implementation of the final model, which we call *Causeformer*, is available already.

2.4 Entropy

The theoretical analysis we aim to conduct according to the above section contradicts both empirical observations in actual language models and theoretical analyses of them, as one predicts and observes a clustering of tokens [5, 1], so one may actually expect to prove a convergence of the form

$$\rho_t = \frac{1}{N}\sum_{i=1}^{N} \delta_{x_t^i} \underset{t\to\infty}{\overset{\mathcal{D}}{\to}} \delta_{x_\infty}.$$

Indeed, as we let W_Q and W_K depend on time, as they do in real models, and as we add the MLPs, we expect the embedding distribution to not obey a heat equation. Instead, as in [1, Fig. 3], we expect ρ_t to first follow $\nabla^{nat}\mathcal{H}(\rho_t)$, and, after a critical point at some intermediate layer, to go in its opposite direction $-\nabla^{nat}\mathcal{H}(\rho_t)$. If we are able to verify this mathematically and empirically, it would have broad implications, as some operationalisations of consciousness identify it with a critical point of the gradient field of entropy.

2.5 Game Theory

We model decisions made by the LLM agents $p_\theta^i(w_t|w_{<t})$ as interacting in an N-Player stochastic game. The intermediate representations governed by the sinkformer flow are interpreted as pre-decisions, so one can test for possible alignment of decisions across models by investigating solubility of the Forward-Backward SDE ([2]) governing their continuously modeled interactions.

The diplomatic FB-SDE is

$$dX_t^i = \left(f_{hybrid}(X_t^i, \tau) + \lambda \nabla_x W_2(X_t^i, \overline{X}_t)\right) dt + \sigma dW_t^i \tag{3}$$
$$dY_t = -\partial_x H(t, X_t, \mu_t, \alpha_t, Y_t)dt + Z_t dW_t \tag{4}$$
$$\alpha_t = \text{argmin}_\alpha H(t, X_t, \mu_t, \alpha, Y_t), \tag{5}$$

where Y_t is the adjoint process, H the Hamiltonian, and Z_t the martingale representation term.

Even though the drift is nonlinear, [2, Volume I, Chapter 2& 6] may help us find conditions under which an equilibrium exists, is unique, and answer the question of stability to perturbations.

After training each causeformer i, during deployment its cost functional is

[3] With D the diagonal matrix with out-degrees, and A the adjacency matrix of G.

$$J^i := \mathbb{E}\Big[\int_0^T \|u_t^i\|^2 + \beta W_{G,2}(X_t^i, \overline{X}_t)dt\Big]; \tag{6}$$

$$u_t^i := -\lambda \nabla_x W_{G,2}(X_t^i, \overline{X}_t), \tag{7}$$

leading each model to follow the Wasserstein potential surface.

2.6 Training Objective

The training objective will be of the form

$$\mathcal{L} = CE(w_{t+1}, \hat{w}_{t+1}) + \lambda \sum_{l=1}^{L} \big(W_{G,2}(\rho^l, \rho^{(l+1)}) + \gamma CI_G(K_\infty^l) + \alpha \|A^l\|_1 + \beta \Phi_{acyc}(A^l)\big). \tag{8}$$

The explanation of the terms is as follows:

- CE: Standard Crossentropy loss, replaceable by any language modelling objective, such as RLHF.
- G-causal Wasserstein term: Distances between successive hidden states measured along G-causal couplings. Lipschitz continuity of treatment-effect type statistics in this metric means "small modelling error $\to$ small causal-effect error". This keeps the learned flow close to the Δ_G trajectory and is implemented using the JKO scheme [6].
- Conditional Independence penalty: [4, Theorem 3.4] gives that "coupling G-causal $\iff$ future independent of non-parents given parents". We estimate conditional mutual information $I(Y_i; X_i, pa_i | past)$ from attention weights and down-weight deviations.
- ℓ_1-Sparsity of Mask: identifiability; fewer edges, easier to learn.
- Acyclicity Constraint: Keep each A^l a DAG differentiably; $\Phi_{acyc}(A) := tr(e^A) - n$.

The last two terms can be seen as doing Bayesian inference on the latent DAGs that generated the tokens; each edge explains the conditional log probability of the next token $p_\theta(w_t | w_{<t})$ (We take the logit lens view of [7].), i.e. the likelihood.

Meanwhile, the sparsity and acyclicity penalties implement our prior assumption of sparse causation. Possibly only one of Φ_{acyc} and CI_G is required.

3 Practical Implementation

We will use the python modules `pytorch`, `torchdiffeq`, `POT` and `pyro`. For the masked sinkhorn attention we simply apply the mask A^l between every row- and column-normalisation. The mask is learned and propagated through using the straight-through estimator.

4 Expected Outcomes and Impact

The following are our scientific objectives:

#	Objective	Deliverable
O1	Derive masked Sinkhorn $\to$ causal heat-flow PDE	Formal Proof
O2	Show Lipschitz continuity of key causal estimands	Theorem with constants, extending [4]
O3	Alternate graph heat steps and abstraction jumps	`pytorch` implementation + stability lemma
O4	N-player SDE game, prove monot. cost $\to$ unique equil.	ex. and uniq. according to [2]
O5	Simulate small-N sandbox	public code + convergence plots

We expect the following impacts:

- **Scientific**: first transformer that is provably a gradient flow on causal manifolds; first application of G-causal Optimal Transport to language.
- **Governance**: a mathematically grounded protocol for inter-LLM negotiation with guaranteed unique peaceful equilibrium.
- **Societal**: a blueprint for an automated diplomatic layer that scales better than human protocol and removes incentives for escalation.

5 Timeline

This is the planned timeline for completing our objectives:

Quarter	Work-Package
Q3 2025	WP1 - Formal PDE proof, implement MaskedSinkhorn, synthetic DAG recovery
Q4 2025	WP2 - Hybrid ODE/MLP module, STE-learned masks; run ablations
Q1 2026	WP3 - Game-theoretic layer, prove equilibrium theorem; small-N sandbox
Q2 2026	WP4 - Scale-up pre-training; publish Causeformer + DMFG library
Q3 2026	WP5 - Policy white-paper + live demo; outreach to EU, PRC, OpenAI, UN

6 Project Members

Principal Investigator: Marvin Koß, B.Sc., Universität Heidelberg.
Email koss@cl.uni-heidelberg.de, Website: marvosyntactical.github.io
External Collaborator: Prof. Dr. Jakob Zech, IWR, Universität Heidelberg.
Email jakob.zech@uni-heidelberg.de, Website: jakobzech.com
Expertise: Neural Network Theory, Bayesian Inverse Problems, Differential Equations, Numerics
External Contact: JProf. Dr. Stephan Eckstein, Universität Tübingen.
Email stephan.eckstein@uni-tuebingen.de, Website: sites.google.com/view/stephan-eckstein/startseite
Expertise: Causal Optimal Transport
TODO: Confirm Prof. Dr. Georgia Koppe (Time Series Models, Neuroscience, Dynamical Systems)
TODO: Confirm Prof. Dr. Simon Weissmann (SDE Theory, RL)
TODO: Confirm Purushart Saxen (Causal Inference, Optimal Transport, Biological Data)
TODO: Wait for Dr. Nikolai Köhler's Answer (Time Series Models, Optimal Transport, Biological Data, Topological Data Analysis)
TODO: Wait for Prof. Dr. Artem Sokolov's Answer (Imitation Learning, NLP)
TODO: Wait for Samuel Kiegeland's Answer (RL, NLP, Psycholinguistics)
TODO: Onboard Prof. Dr. Francois Delarue (Game Theory)

References

[1] Riccardo Ali et al. "Entropy-lens: The information signature of transformer computations". In: *arXiv preprint arXiv:2502.16570* (2025).

[2] René Carmona, François Delarue, et al. *Probabilistic theory of mean field games with applications I-II*. Springer, 2018.

[3] Ricky TQ Chen et al. "Neural ordinary differential equations". In: *Advances in neural information processing systems* 31 (2018).

[4] Patrick Cheridito and Stephan Eckstein. "Optimal transport and Wasserstein distances for causal models". In: *Bernoulli* 31.2 (2025), pp. 1351–1376.

[5] Borjan Geshkovski et al. "A mathematical perspective on transformers". In: *arXiv preprint arXiv:2312.10794* (2023).

[6] Richard Jordan, David Kinderlehrer, and Felix Otto. "The variational formulation of the Fokker–Planck equation". In: *SIAM journal on mathematical analysis* 29.1 (1998), pp. 1–17.

[7] nostalgebraist. *interpreting gpt: the logit lens*. LessWrong. 2020. URL: https://www.lesswrong.%20com/posts/AcKRB8wDpdaN6v6ru/%20interpreting-gpt-the-logit-lens (visited on 05/14/2025).

[8] Michael E Sander et al. "Sinkformers: Transformers with doubly stochastic attention". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2022, pp. 3515–3530.

[9] Anh Tong et al. "Neural ODE Transformers: Analyzing Internal Dynamics and Adaptive Fine-tuning". In: *arXiv preprint arXiv:2503.01329* (2025).