Benchmarking Large Language Models for Geomagnetic Disturbance Mitigation in Power Systems

Parham Mohammadi
Department of Electrical Engineering and Computer Science
York University
Toronto, Canada
parham11@yorku.ca

Afshin Rezaei–Zare
Department of Electrical Engineering and Computer Science
York University
Toronto, Canada
rezaei@yorku.ca

***Abstract*—We propose a benchmark for evaluating large language models (LLMs) in geomagnetic disturbance (GMD) mitigation when transformer line–outage distribution factors (TLDOFs) are the primary mechanism. The benchmark formalizes TLDOF–centric screening and planning, defines task tracks spanning code/tooling, decision quality, and explanation, and specifies datasets, metrics, and protocols that reflect both power–system performance and model reliability. It targets realistic transmission test systems and supports both offline comparison and human–in–the–loop operation. The result is a reproducible pathway to compare agents, surface failure modes, and integrate them safely into GMD workflows.**

***Index Terms*—geomagnetic disturbance, geomagnetically induced current, line-outage distribution factor, TLDOF, large language model, benchmark, power system resilience**

I. Introduction

Geomagnetic disturbances (GMDs) drive quasi–DC geomagnetically induced currents (GICs) that stress transformers and erode voltage margins [1], [2]. By driving half–cycle saturation, GICs increase transformer losses and reactive power demand, introduce harmonics, and can precipitate system–level voltage instability and cascading outages [5], [10], [17]. Transformer line–outage distribution factors (TLDOFs) provide fast, topology–aware screening of mitigation actions while preserving security constraints [11], [19]. In parallel, tool–augmented large language models (LLMs) are increasingly used for code generation, simulation orchestration, and operator decision support [24], [25], yet no domain benchmark evaluates their performance in TLDOF–based GMD mitigation.

This work introduces a compact, TLDOF–centric benchmark for LLMs, with clear tasks, interfaces, and metrics spanning coding, decision quality, and explanation.

Our contributions are:

- A concise TLDOF–centric formulation linking AC power flow, DC GIC circuits, and operational constraints, suitable for LLM–assisted tooling.
- A three–track benchmark (code/tooling, decision quality, explanation) with explicit I/O schemas, feasibility checks, and robustness/safety protocols.
- An IEEE 118 evaluation with multi–model LLM results and reference scripts to enable reproducible comparisons on public test systems [22], [23].

II. Background

A. Geomagnetic Disturbances and GICs

GMDs induce geoelectric fields in the Earth's crust. Long transmission lines behave as antennas, and quasi–DC GICs are driven into grounded transformer neutrals, where they bypass the normal AC magnetization path and can drive half–cycle saturation [5], [10]. This increases reactive power absorption, produces harmonics, and elevates local heating and protection stress [2], [5]. At the system level, reactive demand can depress voltage and reduce stability margins during severe storms [1], [10]. Analyses typically couple a DC GIC network driven by geoelectric fields with AC power flow for feasibility validation [11], [13], [15]–[17]; heterogeneous ground conductivity can materially change induced currents and therefore screening accuracy [18].

During saturation, distorted magnetizing current introduces harmonics that can propagate through the network, cause CT saturation, and contribute to relay misoperation, stressing protection and monitoring [2], [5]. Reactive power consumption may rise sharply with storm intensity, reducing voltage stability margins and increasing the likelihood of remedial actions [1], [10]. Thermal impacts include elevated hot–spot temperatures, accelerated insulation aging, and dissolved gas generation detectable via dissolved gas analysis (DGA), motivating conservative operating thresholds and monitoring [2], [3], [5]. GIC exposure depends strongly on line orientation relative to the geoelectric field and on spatial variability in ground conductivity, so magnitude–and–azimuth scenario sweeps are common in GMD studies [17], [18]. Accordingly, many workflows use linear screening (e.g., TLDOFs) to prioritize candidates, then rely on coupled AC + GIC simulation to verify feasibility under voltage and thermal constraints [11], [15], [19]. These coupled effects motivate our benchmark's emphasis on both risk reduction and hard feasibility checks under uncertain storm inputs.

979-8-3315-8138-1

At the equipment level, saturation can increase stray flux and localized heating in transformer structural components; hotspot models quantify how temperature rise and loss of life scale with GIC exposure duration and magnitude [5], [6]. Reactive power loss versus GIC is typically nonlinear, implying that linear screening approximations can diverge at high field strengths [7]. Historical events (e.g., Hydro–Québec 1989 and the October 2003 "Halloween" storms in Sweden) and field reports show that these equipment–level effects can couple into wide–area outages even in regions sometimes assumed low–risk [1], [8], [9], [14]. Standards such as IEEE C57.163 and NERC TPL-007 motivate conservative capability limits and planned performance requirements, reinforcing the need for mitigation plans that are both risk–effective and AC-secure [3], [4].

B. Transformer Line-Outage Distribution Factors

Classical line–outage distribution factors (LODFs) quantify how flows redistribute after an outage. TLDOFs adapt this notion to transformer GIC exposure, ranking switching actions by their impact on at–risk transformers [19]. TLDOF-based mitigation algorithms rank lines according to their TLDOF impact on critical transformers and use optimization to select a small set of switching actions that keep all transformer GICs below thermal limits while respecting AC operational constraints [19].

C. LLMs in Power Systems and Benchmarking

LLMs have been explored for forecasting, optimisation, anomaly detection, and operator interfaces [24], [25]. However, existing evaluations do not target TLDOF–based GMD mitigation, where approximate linear sensitivities must be coupled to hard AC security checks and reliability-focused metrics under operational time constraints.

III. Problem Formulation and TLDOF-Centric Mitigation

A. Problem Definition (Formal)

Inputs: AC power system model, geoelectric field scenario, TLDOF sensitivity matrix, switching budget K and operational constraints.

Outputs: Ranked switching actions $\mathcal{S}$ with $|\mathcal{S}| \leq K$, predicted GIC risk reduction ΔR, feasibility flags (voltage, thermal, connectivity), natural language explanations.

Notation: E (V/km) is the geoelectric field scenario; $\mathcal{S}$ is the set of opened lines with $|\mathcal{S}| \leq K$. I_t^{GIC} (A) is transformer t quasi–DC current with limit $I_t^{\max}$; $R(\mathcal{S})$ is risk in (1), $\Delta R \triangleq R(\emptyset) - R(\mathcal{S})$, and $\Delta R_g \triangleq R(\emptyset) - R(\mathcal{S}_g)$ is the greedy baseline risk reduction for plan $\mathcal{S}_g$ (Track B reports $\Delta R/\Delta R_g$). AC feasibility enforces $|V_i| \in [0.95, 1.05]$ p.u. for all buses.

Assumptions: TLDOFs are used as first–order screening sensitivities. Linearity can degrade under high–GIC saturation and AC/DC coupling, interacting multi–outage actions, and heterogeneous 3D ground conductivity; therefore the benchmark treats TLDOF ranking as a screening step and requires AC + GIC co–simulation validation before acceptance [10], [15], [17], [18]. This proxy–truth gap is an explicit benchmark target: Track B plans may appear beneficial under TLDOF–predicted ΔR yet fail AC + GIC validation when the linear proxy breaks down. We therefore gate and score plans using hard post–screen feasibility checks and assign zero credit to infeasible actions regardless of predicted risk reduction.

Constraints: No network islanding, transformer thermal limits respected, voltage stability margins maintained, switching budget enforced.

Threat Model: LLM agents may produce hallucinated actions, violate safety constraints, or fail under adversarial prompts. The benchmark includes red-team testing and provenance tracking to mitigate these risks.

B. GMD-Aware Operational Problem

Under a geoelectric field scenario, a DC GIC network coupled to AC power flow yields transformer GIC currents [11]. Operators can redispatch generation, switch a limited set of lines, and (optionally) activate blockers, seeking to minimise risk while satisfying AC security and GIC limits. Direct multi–scenario optimisation is expensive; TLDOFs enable fast screening and ranking of candidate switching actions.

C. TLDOF-Based Screening

Precomputed TLDOFs allow greedy or budgeted selection of lines that most reduce a chosen GIC risk metric; distributionally robust variants address uncertainty [21]. Such methods reduce exposure with few actions and acceptable security margins on realistic test systems [19], [22], [23].

Let $I_t^{\mathrm{GIC}}(\mathcal{S})$ denote transformer t current after opening lines in $\mathcal{S}$ (screened by TLDOF and later validated by AC + GIC checks), with limit $I_t^{\max}$. A common risk objective is the sum of squared limit exceedances across transformers,

$$R(\mathcal{S}) = \sum_t \max\bigl(0,\, I_t^{\mathrm{GIC}}(\mathcal{S}) - I_t^{\max}\bigr)^2, \ \ |\mathcal{S}| \leq K. \qquad (1)$$

Squaring emphasizes severe violations, yields a smooth scalar for greedy/budgeted screening, and aligns with TLDOF–based mitigation formulations where large exceedances dominate thermal risk [5], [6], [19]. Compared to a max–exceedance objective, it provides graded credit across multiple stressed transformers; compared to an ℓ_1 sum, it penalizes rare large overloads more strongly. Consequence–based metrics (e.g., hotspot temperature rise) and blocker–placement objectives are also used [6], [20]; under storm uncertainty, probabilistic/robust variants are also plausible [21]. Greedy screening uses TLDOFs to approximate marginal ΔR for each candidate line, selecting the best within budget K before validating with AC + GIC co–simulation.

D. Where LLMs Enter the Loop

While TLDOF-based algorithms can be implemented in conventional optimization frameworks, GMD mitigation is operationally complex: operators must interpret TLDOF tables, mitigation plans must respect pre-existing procedures, and new GMD events may require fast adaptation of models and scripts across heterogeneous simulation tools. LLMs can act as

979-8-3315-8138-1

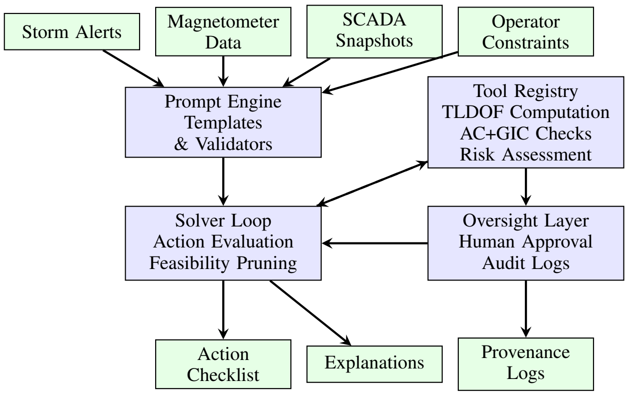

Fig. 1. Architecture overview of the TLDOF-based GMD mitigation agent showing key components and data flow.

tool–augmented agents that call simulators, propose switching plans, and explain trade–offs, as well as code generators for TLDOF and GIC studies [25].

E. Architecture Overview

Figure 1 illustrates the practical agent architecture for TLDOF–based mitigation, comprising: (i) a conversational core with prompt templates and schema validators; (ii) a tool registry exposing TLDOF computation, AC + GIC checks, and playbook assembly; (iii) a solver loop to evaluate candidate actions; and (iv) an oversight layer for human approval and cryptographic audit logs.

Key components: **Prompt Engine** normalises context and enforces output schemas; **Tool Registry** provides TLDOF screening, risk computation, AC + GIC feasibility checks; **Solver Loop** iterates ranked actions and selects the best budgeted plan; **Oversight** routes recommendations for approval and enforces policy guardrails. This architecture informs the benchmark's tool APIs and ensures that advances translate to operator–usable workflows.

F. TLDOF Formalism (Concise)

Let $u_\ell \in \{0,1\}$ denote the status of line ℓ (1=closed, 0=open). We define a transformer–line sensitivity (TLDOF) as the first–order sensitivity of transformer quasi–DC current to the status of a line,

$$T_{t,\ell} \triangleq \left.\frac{\partial I_t^{\mathrm{GIC}}}{\partial u_\ell}\right|_{u_\ell=1}, \qquad (2)$$

and approximate the post–switching GIC at transformer t for a set $\mathcal{S}$ of opened lines by

$$I_t^{\mathrm{GIC}}(\mathcal{S}) \approx I_t^{\mathrm{GIC}} + \sum_{\ell\in\mathcal{S}} T_{t,\ell}\, I_\ell^{\mathrm{GIC}}, \qquad (3)$$

where I_ℓ^{GIC} (A) is the base–topology quasi–DC line current driven by the induced corridor voltage ($\int_\ell E \cdot dl$, capturing orientation and length) and the DC network resistances. Screening uses the risk objective $R(\mathcal{S})$ in (1).

IV. Benchmark Design

We evaluate LLMs across three complementary tracks: (A) code/tooling, (B) decision quality, and (C) explanation/alignment.

A. Scope and Datasets

We include open transmission systems spanning medium to large scale. For the 118–bus demonstration in Section V, the AC model follows the standard IEEE case; a GIC DC layer is obtained by mapping buses to ten substations with ground resistances in 0.65 – 2.9 Ω, and projecting uniform geoelectric fields (0.5 – 3.0 V/km) onto line corridors. These field ranges align with NERC GMD benchmarking events and the recent May 2024 "Gannon" G5 storm [23]. Each dataset ships with: (i) AC+DC models and geoelectric field scenarios; (ii) reference TLDOF tables; (iii) gold baseline plans and evaluation scripts. We follow established GIC test–case coupling practice for DC network construction and validation [11], [12].

B. Tool Interfaces

To reduce tool friction, we expose a small set of canonical functions (e.g., `compute_tldof`, `run_ac_gic`, `check_security`) behind a JSON–serialised API. Agents submit structured calls and receive typed returns (arrays, tables, flags).

C. Protocols and Metrics

Evaluation protocols cover zero–shot, few–shot, and tool–augmented settings. Beyond accuracy, we report: (i) reliability across paraphrases; (ii) robustness to noisy TLDOFs; (iii) operator effort (dialogue turns until plan passes screen); and (iv) runtime.

D. Benchmark Specification (Summary)

Track A. Input: AC case, geoelectric field series, tool API interface. Output: unit–tested functions `compute_tldof(case)`, `run_ac_gic(case,E)`. Metrics: pass rate, numerical agreement (RMSE vs reference), robustness to prompt paraphrases.

Track B. Input: snapshot (case, scenario E, TLDOF table, K, constraints). Output: JSON plan with actions and safety fields. Metrics: $\Delta R/\Delta R_g$ and feasibility rate; feasibility requires no islanding, $|V_i| \in [0.95, 1.05]$ p.u. for all buses, and line loading $\leq 100\%$ after AC + GIC validation.

Safety rules (hard). Any plan that fails post–screen AC + GIC validation (including thermal limits) fails Track B regardless of predicted ΔR.

Track C. Input: scenario + plan. Output: free–form explanation and Q&A. Metrics: blinded expert scores (correctness, completeness, clarity), and hallucination rate.

E. Robustness and Safety

We recommend red–team prompts and data–leakage probes to test safety posture. Models should be rewarded for deferring actions under uncertainty and for flagging missing data or violated constraints. Outputs should attach provenance (e.g., hashes of TLDOF tables) to aid auditability.

979-8-3315-8138-1

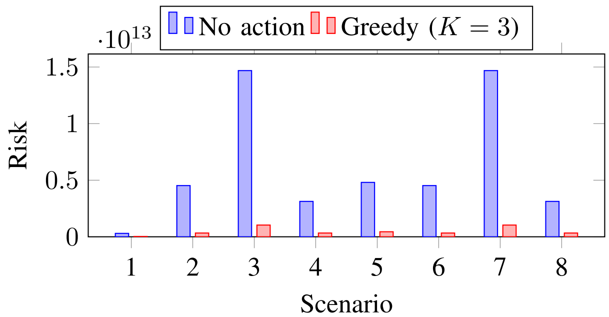

Fig. 2. Risk (sum of squared exceedances) per scenario before and after greedy TLDOF screening.

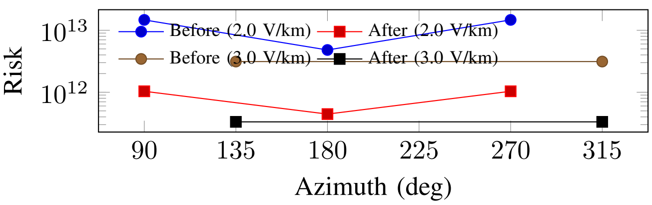

Fig. 3. Risk vs azimuth for selected field magnitudes (log scale) before and after greedy mitigation.

F. Baselines and Reproducibility

We include pragmatic baselines: (i) TLDOF–greedy with budgeted switching; (ii) a small integer programme when tractable; and (iii) no–action and random–action baselines. All experiments fix random seeds, pin tool versions, and log hashes of datasets and TLDOF tables. An artifact package containing the AC/DC cases, geoelectric scenarios, TLDOF tables, and evaluation scripts is intended for public release.

V. Results

We evaluate the benchmark on an IEEE 118–bus dataset (buses = 118, lines = 186, generators = 25, transformers = 9). Eight GMD scenarios span 0.5 – 3.0 V/km and azimuths at $45°$ increments. The switching budget is $K = 3$ with no islanding and approximate feasibility screens.

Figure 2 shows the risk metric (sum of squared GIC limit exceedances) per scenario before and after greedy TLDOF screening. Across all eight scenarios, greedy selection reduces the proxy risk by one to two orders of magnitude while preserving network connectivity. We treat this greedy policy as the *Track B reference baseline* that any LLM + tools policy should at least match on $\Delta R/\Delta R_g$ and feasibility.

Figure 3 plots risk versus azimuth at 2.0 and 3.0 V/km on a log scale, illustrating strong directional dependence and the consistent reduction achieved by the budgeted plan.

LLM panel. Table I reports results for a set of LLM baselines on Tracks A/B/C. Frontier models approach the greedy baseline on $\Delta R/\Delta R_g$ and feasibility while smaller models degrade across all three tracks.

LLM Baselines and Protocol: **Models.** We evaluate both frontier and smaller open models in tool–augmented mode with access to `compute_tldof`, `run_ac_gic`, and `check_security` via JSON calls; temperature is 0.2 with a concise token cap. In Table I, GPT 4o is representative of the frontier category, while Llama 3.1 8B Instruct represents the smaller model category.

TABLE I

Model comparison on IEEE–118 across Tracks A/B/C: unit-test pass rate (A), relative risk reduction vs greedy and feasibility (B), and expert score (C).

Model	Provider	Params	Tool	A (%)	B $\Delta R/\Delta R_g$	B feas. (%)	C (1–5)
Greedy baseline	Baseline	–	n/a	–	1.00	100	–
GPT-4o	OpenAI	–	yes	90	0.94	92	4.2
Claude 3.5 Sonnet	Anthropic	–	yes	92	0.95	93	4.3
Llama 3.1 70B Instruct	Meta	70	yes	82	0.86	86	3.9
Qwen2.5-72B-Instruct	Alibaba	72	yes	83	0.88	87	3.9
Llama 3.1 8B Instruct	Meta	8	yes	60	0.70	72	3.5

Track A. Ten coding tasks (TLDOF and GIC functions) with 30 unit tests per function; we report the fraction of passed tests. **Track B.** Eight IEEE 118 scenarios with budget $K = 3$; agents may propose one plan per scenario and invoke tools freely; we report average $\Delta R/\Delta R_g$ and feasibility rate. **Track C.** Twenty plan explanations rated by one domain expert on a 1 – 5 scale; we report the mean score.

Interpretation. Frontier models (GPT–4o, Claude 3.5 Sonnet) nearly match the greedy baseline on $\Delta R/\Delta R_g$ and feasibility but exhibit occasional feasibility misses; smaller open models (e.g., Llama 3.1 8B Instruct) underperform across tracks. This supports the benchmark's emphasis on hard safety checks and structured outputs in tool–augmented settings.

Evaluation Fairness and Computational Cost. We fix temperature (0.2), tool call budgets (max 10 calls per scenario), and random seeds. With precomputed TLDOF tables, greedy screening scales as $O(K|\mathcal{L}||\mathcal{T}|)$ per scenario and is fast in our reference implementation on IEEE-118 (risk evaluation $<$ 0.1 ms; greedy K=3 selection $\approx$ 40 ms per scenario). AC + GIC co–simulation costs are approximately 0.15 s per run on a standard workstation and dominates end–to–end latency under the tool-call budget (LLM inference latency is model/hardware-dependent).

VI. Discussion

A. Domain-Specific vs General-Purpose LLMs

The proposed benchmark can compare: general-purpose frontier models; power-system–specialized models (e.g., trained on grid standards, operation manuals, and GMD literature); and hybrid systems coupling smaller LLMs with problem-specific solvers. Domain–specific models often outperform general models on specialised tasks given similar parameter counts. We expect specialised TLDOF–GMD copilots to be more reliable and interpretable, while general models excel at code generation.

B. Human-in-the-Loop Integration

LLMs for GMD mitigation should be deployed as advisory tools, not autonomous controllers. The benchmark therefore emphasizes: structured outputs that can be easily reviewed and overridden by operators; explanations aligned with existing operational procedures (e.g., NERC TPL-007, IEEE C57.163, utility playbooks) [3], [4]; and interfaces that combine natural-language dialogue with canonical power-system visualizations and TLDOF tables.

C. Related Work

Benchmarks for tool–augmented LLMs and scientific reasoning exist in general domains, but they do not address physics–constrained grid tasks with safety–critical constraints and TLDOF–based screening. Our design differs by grounding tasks in transformer GIC physics, explicit risk and feasibility metrics, and human–in–the–loop procedures relevant to operations.

VII. Limitations and Future Work

Limitations include reliance on public test networks, approximate linearity of TLDOFs (extreme events may need EMT) [15], and the cost of human evaluation. Track C in this demo uses one expert rater; future versions will include multiple raters. Security and privacy should be incorporated via red–team prompts and leakage probes where appropriate.

Future work will extend the benchmark to larger transmission systems (e.g., WECC 179-bus, European 2383-bus), incorporate real-time PMU data streams, and develop specialized TLDOF-GMD copilot models trained on grid standards and operational procedures. Additional research directions include multi-objective optimization balancing GIC risk with economic dispatch, integration with weather forecasting systems, and deployment in utility control room environments.

VIII. Conclusion

We proposed a TLDOF–centric benchmarking framework for evaluating and deploying LLMs in GMD mitigation. By tying LLM tasks to TLDOF–based screening and optimisation, the three–track design—code and tooling, decision quality, and explanation—enables holistic assessment of assistance in computing TLDOFs, proposing actions, and explaining strategies.

Future work will extend the implementation to additional GMD test systems, compare a broader range of general-purpose and domain-specific LLMs, and refine metrics for robustness, security, and human trust. Ultimately, rigorous benchmarking is a prerequisite for safely integrating LLMs into real-time GMD response workflows and enhancing the resilience of modern power systems to space-weather–induced disturbances.

References

[1] J. G. Kappenman, "Geomagnetic storms and their impact on power systems," *IEEE Power Engineering Review*, vol. 16, no. 5, p. 5, 1996.

[2] IEEE PES Task Force on Geomagnetic Disturbances, "Geomagnetic disturbances: Their impact on the power grid," *IEEE Power and Energy Magazine*, vol. 11, no. 4, pp. 71–78, Jul./Aug. 2013.

[3] IEEE Std. C57.163-2015, *IEEE Guide for Establishing Power Transformer Capability While Under Geomagnetic Disturbances*, 2015.

[4] North American Electric Reliability Corporation (NERC), *Transmission System Planned Performance for Geomagnetic Disturbance Events*, Reliability Standard TPL-007-4, 2020.

[5] P. R. Price, "Geomagnetically induced current effects on transformers," *IEEE Trans. Power Delivery*, vol. 17, no. 4, pp. 1002–1008, Oct. 2002.

[6] L. Marti, A. Rezaei-Zare, and A. Narang, "Simulation of transformer hot-spot heating due to geomagnetically induced currents," *IEEE Trans. Power Delivery*, vol. 28, no. 1, pp. 320–327, Jan. 2013.

[7] A. Rezaei-Zare, "Reactive power loss versus GIC characteristic of single-phase transformers," *IEEE Trans. Power Delivery*, vol. 30, no. 3, pp. 1639–1640, Jun. 2015.

[8] L. Bolduc, "GIC observations and studies in the Hydro–Québec power system," *J. Atmos. Solar-Terr. Phys.*, vol. 64, no. 16, pp. 1793–1802, Nov. 2002.

[9] C. T. Gaunt and G. Coetzee, "Transformer failures in regions incorrectly considered to have low GIC-risk," in *Proc. IEEE PowerTech*, Lausanne, Switzerland, 2007, pp. 807–812, doi: 10.1109/PCT.2007.4538419.

[10] K. Shetye and T. J. Overbye, "Modeling and analysis of GMD effects on power systems: An overview of the impact on large-scale power systems," *IEEE Electrification Magazine*, vol. 3, no. 4, pp. 13–21, 2015.

[11] T. J. Overbye, T. R. Hutchins, K. Shetye, J. Weber, and S. Dahman, "Integration of geomagnetic disturbance modeling into the power flow: A methodology for large-scale system studies," in *Proc. North American Power Symposium*, 2012, pp. 1–7.

[12] R. Horton, D. Boteler, T. J. Overbye, R. Pirjola, and R. C. Dugan, "A test case for the calculation of geomagnetically induced currents," *IEEE Trans. Power Delivery*, vol. 27, no. 4, pp. 2368–2373, Oct. 2012, doi: 10.1109/TPWRD.2012.2206407.

[13] L. Trichtchenko and D. H. Boteler, "Modeling geomagnetically induced currents using geomagnetic indices and data," *IEEE Trans. Plasma Sci.*, vol. 32, no. 4, pp. 1459–1467, 2004.

[14] A. Pulkkinen, S. Lindahl, A. Viljanen, and R. Pirjola, "Geomagnetic storm of 29–31 October 2003: Geomagnetically induced currents and their relation to problems in the Swedish high-voltage power transmission system," *Space Weather*, vol. 3, no. 8, Art. no. S08C03, 2005, doi: 10.1029/2004SW000123.

[15] A. Haddadi, R. Hassani, J. Mahseredjian, L. Gerin-Lajoie, and A. Rezaei-Zare, "Evaluation of simulation methods for analysis of geomagnetic disturbance system impacts," *IEEE Trans. Power Delivery*, vol. 36, no. 3, pp. 1509–1516, 2021.

[16] D. H. Boteler and R. J. Pirjola, "Comparison of methods for modelling geomagnetically induced currents," *Ann. Geophys.*, vol. 32, no. 9, pp. 1177–1187, 2014.

[17] D. H. Boteler and R. J. Pirjola, "Modeling geomagnetically induced currents," *Space Weather*, vol. 15, no. 1, pp. 258–276, 2017.

[18] L. Juusola, H. Vanhamäki, A. Viljanen, and M. Smirnov, "Induced currents due to 3D ground conductivity play a major role in the interpretation of geomagnetic variations," *Annales Geophysicae*, vol. 38, no. 5, pp. 983–998, 2020.

[19] M. Kazerooni and T. J. Overbye, "Transformer protection in large-scale power systems during geomagnetic disturbances using line switching," *IEEE Trans. Power Systems*, vol. 33, no. 6, pp. 5990–5999, 2018.

[20] A. H. Etemadi and A. Rezaei-Zare, "Optimal placement of GIC blocking devices for geomagnetic disturbance mitigation," *IEEE Trans. Power Systems*, vol. 29, no. 6, pp. 2753–2762, Nov. 2014.

[21] M. Ryu, H. Nagarajan, and R. Bent, "Mitigating the impacts of uncertain geomagnetic disturbances on electric grids: A distributionally robust optimization approach," *IEEE Trans. Power Systems*, vol. 37, no. 6, pp. 4258–4269, 2022.

[22] A. B. Birchfield, K. M. Gegner, T. Xu, K. S. Shetye, and T. J. Overbye, "Statistical considerations in the creation of realistic synthetic power grids for geomagnetic disturbance studies," *IEEE Trans. Power Systems*, vol. 32, no. 2, pp. 1502–1510, Mar. 2017.

[23] D. H. Mac Manus *et al.*, "Implementing geomagnetically induced currents mitigation during the May 2024 "Gannon" G5 storm: Research informed response by the New Zealand power network," *Space Weather*, vol. 23, no. 6, 2025.

[24] L. Zhang, W. Ji, Y. Zhao, D. Huang, and B. Qian, "Review on application and challenges of large language models in power grids," *IEEE Access*, vol. 13, pp. 106932–106941, 2025.

[25] M. Jia, Z. Cui, and G. Hug, "Enabling large language models to perform power system simulations with previously unseen tools: A case of Daline," *arXiv:2406.17215*, 2024.

979-8-3315-8138-1