Transformerは熱力学的推論を行うか？ ：geDIGゲージによる構造相転移の観測と制御

Does Transformer Perform Thermodynamic Inference?
: Observation and Control of Structural Phase Transitions via geDIG Gauge

宮内 和義 [*1]
Kazuyoshi Miyauchi

[*1]独立研究者
Independent Researcher

Dynamic knowledge acquisition entails a fundamental trade-off: exploring new information vs. integrating verified structures. We propose **geDIG**, a unified gauge bridging the Free Energy Principle (FEP) and Minimum Description Length (MDL) to quantify this trade-off as $\mathcal{F} = \Delta\mathrm{EPC}_{\mathrm{norm}} - \lambda(\Delta H_{\mathrm{norm}} + \gamma\Delta\mathrm{SP}_{\mathrm{rel}})$. This paper demonstrates that $\mathcal{F}$ governs structural optimization across scales. (1) **Macro**: In a partial-observation maze task, a geDIG-driven agent autonomously switches between exploration (Attention Gate) and integration (Decision Gate), achieving 98% success with 95% graph compression. (2) **Micro**: Applying $\mathcal{F}$ to Transformer attention reveals a thermodynamic "phase transition" from entropic interaction to structured sparsity across layers. Furthermore, introducing F-regularization in fine-tuning causally improves performance (+0.33%), suggesting geDIG as a design principle for next-generation, thermodynamically efficient architectures.

1. はじめに

知能システムにおいて，新しい情報を「いつ」構造として受け入れ,「いつ」探索を続けるべきか（When 問題）は，マクロなエージェント行動からミクロな内部推論に至るまで共通の課題である．RAG（検索拡張生成）は「何を取るか（What）」の最適化に成功したが，情報の統合タイミングはヒューリスティックに依存している [4]．一方，Transformer[3] は固定的な層計算を行うが，その内部で情報がどう構造化され，いつ「理解」が完了するのかはブラックボックスのままである．

本研究では，自由エネルギー原理（FEP）[1] と最小記述長（MDL）[2] を操作的に橋渡しする統一ゲージ **geDIG** (graph edit Distance and Information Gain) を提案する．本稿の目的は，このゲージがマクロな探索（迷路）からミクロな推論（Attention）まで，一貫した「構造化の原理」として機能することを示すことである．具体的には以下の 3 点を貢献とする：

1. 構造コストと情報利得のトレードオフを単一スカラー化した geDIG の定式化
2. 迷路環境における探索・統合の自律スイッチングの実証
3. Transformer 内部における構造相転移の発見と，正則化による因果的介入

2. geDIG：統一ゲージの定義

geDIG は,「構造を変えるコスト」と「それによって得られる情報の質」の収支を単一スカラー $\mathcal{F}$ で評価する．

$$\mathcal{F} = \Delta\mathrm{EPC}_{\mathrm{norm}} - \lambda\,(\Delta H_{\mathrm{norm}} + \gamma \cdot \Delta\mathrm{SP}_{\mathrm{rel}}) \qquad (1)$$

ここで各項は以下に対応する（表 1）．

- $\Delta\mathrm{EPC}_{\mathrm{norm}}$：正規化グラフ編集距離．ノードやエッジの追加コストであり，MDL におけるモデル記述長 $L(M)$ の増分に相当する．
- ΔH_{norm}：シャノンエントロピーの差分（負の利得）．FEP における「驚き（変分自由エネルギー）」の最小化に対応し，分布が先鋭化（秩序化）するほど $\mathcal{F}$ を下げる．
- $\Delta\mathrm{SP}_{\mathrm{rel}}$：平均最短路長の相対短縮率．グラフ上のショートカット形成による効率化を表し，MDL におけるデータ記述長 $L(D|M)$ の圧縮に相当する．

表 1: geDIG 構成項の理論的対応

項	物理的意味	FEP/MDL 対応	役割
$\Delta\mathrm{EPC}_{\mathrm{norm}}$	構造仕事	$L(M)$ 増分	コスト（抑制）
ΔH	エントロピー	自由エネルギー	秩序化（負）
$\Delta\mathrm{SP}$	経路短縮	$L(D\|M)$ 圧縮	効率化（正）

パラメータ λ, γ は「情報温度」として機能し，構造の複雑さと情報の圧縮率のバランスを決定する．

2.1 AG/DG 二段ゲート制御

この $\mathcal{F}$ を用いて，エージェントは以下の 2 つのゲートを事象駆動的（Event-Driven）に制御する．

1. **AG (Attention Gate)**: 0-hop（直近）での評価．$g_0 = \mathcal{F}|_{\mathrm{local}} > \theta_{\mathrm{AG}}$ のとき，局所的な「違和感・未知」を検知し，探索モード（Exploration）を起動する．
2. **DG (Decision Gate)**: Multi-hop（数理推論）での評価．$g_{\min} = \min_h \mathcal{F}^{(h)} < \theta_{\mathrm{DG}}$ のとき，大域的な「近道・洞察」を発見したとみなし，統合モード（Integration）としてエッジを確定する．

この機構により，システムは常に探索し続けるのではなく,「わからないときだけ調べ（AG）」,「わかったときだけ覚える（DG）」という自律的な振る舞いを獲得する（図 1）．

3. 検証 I：迷路による原理検証（マクロ）

まず，物理的な「構造探索」の典型例として，部分観測迷路（15×15〜51×51）を用いた検証を行った．エージェントは自

連絡先: 宮内 和義（独立研究者），所在地：日本，E-mail: miyauchikazuyoshi@gmail.com

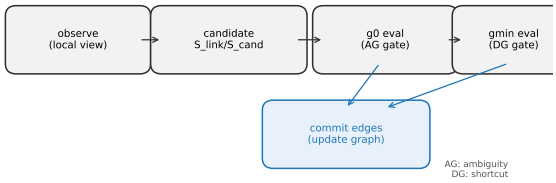

図 1: AG/DG 制御フロー．0-hop で「違和感」を検知して探索し，Multi-hop で「近道」を発見して統合する．

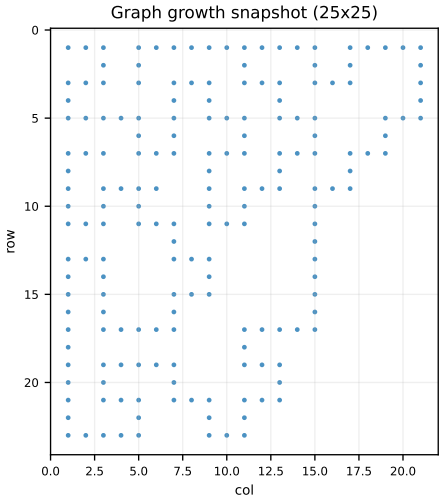

図 2: 迷路におけるグラフ形成（25×25）．空間のトポロジー的に重要な分岐点のみが構造化されている．

身の周囲 1 マスしか見えず，移動しながら内部グラフ（メンタルマップ）を構築する．比較対象として，Random Walk，Greedy DFS（全探索），および geDIG エージェントを設定した．

3.1 結果：探索と統合の自律制御

geDIG エージェントは，AG/DG のダイナミクスのみで迷路探索に成功した．

- **成功率**: 15×15 迷路において，Random Walk が 0.45，Greedy が 0.92 に対し，geDIG は **0.98** を達成した．
- **グラフ圧縮**: 特筆すべきは保持するグラフのサイズである．geDIG は直線通路などの冗長なノードを DG 条件で棄却し，交差点や行き止まりといった「トポロジー的骨格」のみを記憶した．結果，全訪問ノードに対する保持ノードの圧縮率は **95%** に達した（図 2）．

時系列解析からは，行き止まりに遭遇した瞬間に AG が発火（探索開始）し，ループを閉じて既知の経路に合流した瞬間に DG が発火（統合）する様子が確認された．これは，設計者が「行き止まりなら戻れ」とルールを書かずとも，$\mathcal{F}$ の最小化という原理だけで適切な振る舞いが創発することを示している．

4. 検証 II：Transformer の熱力学的解釈（ミクロ）

次に，このマクロな原理が，現代の深層学習モデル（Transformer）の内部でも成立しているかを検証した．Attention 行列

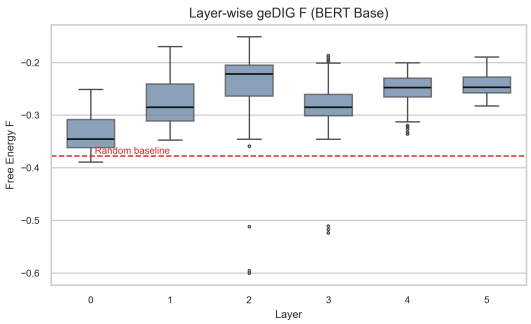

図 3: 層別の F 値分布（BERT）．浅層での高エントロピー状態から，深層での低 F 状態（構造化）への相転移が見られる．

A_{ij} を有向グラフの隣接行列と見なし，閾値処理（上位 10%）を行ってグラフ化し，その $\mathcal{F}$ 値を計測した．対象モデルは BERT, GPT-2, Llama 3 などの代表的な LLM である．

4.1 観測 1：層別の構造相転移

図 3 に BERT の層別 F 値分布を示す．ベースラインとしてランダム行列（同密度）の F 値を算出したところ，実 Attention は有意に低い（$\Delta F \approx 0.11$）値を示した．これは学習済みモデルがランダムな接続ではなく，意味のある構造的接続を獲得していることを定量的に裏付ける．

さらに重要な発見は，深さ方向の推移である．

- **Layer 0-1**: F 値が高く（−0.34），エントロピーが高い．これは迷路における「探索相（Exploration）」に対応する．
- **Layer 2-3**: F 値が急激に低下し（−0.24），構造化が進む．
- **Layer 4+**: 低い値で安定する．これは「構造相（Structure）」への**相転移**と解釈できる．

この結果は，Transformer が層を重ねるごとに情報を「探索」から「結晶化」へと移行させていることを示唆する．

4.2 観測 2：ヘッドの機能分化

同一層内でもヘッドごとに F 値のばらつきが見られた．一部のヘッドは極めて低い F 値（強い構造化，文法処理など）を示し，他方は高い F 値（広範な探索）を示した．これは Attention Head が均質なコピーではなく，**「探索担当」と「統合担当」のマルチエージェント**として機能分化していることを熱力学的に裏付けるものである．

4.3 観測 3：F 正則化による因果的介入

以上の観測は相関関係に過ぎない．$\mathcal{F}$ が真に計算効率や性能に関与しているなら，この値を直接最適化することで性能が変化するはずである．DistilBERT を用いた SST-2 感情分析タスクの Fine-tuning において，損失関数に F 項による正則化を追加した．

$$L_{\text{total}} = L_{\text{CE}} + \alpha \cdot \mathcal{F}$$

実験の結果（図 4），微弱な正則化（$\alpha = 0.001$）を与えた場合に，ベースライン（$\alpha = 0$）に比べて Test Accuracy が 86.00% $\rightarrow$ **86.33%** へと改善した．一方で，強すぎる正則化（$\alpha = 1.0$）は構造を固定化しすぎ，性能を悪化させた．この逆 U 字型の特性は，適切な「構造化圧」を与えることで，モデルの汎化性能を引き出せる可能性（因果性）を示している．

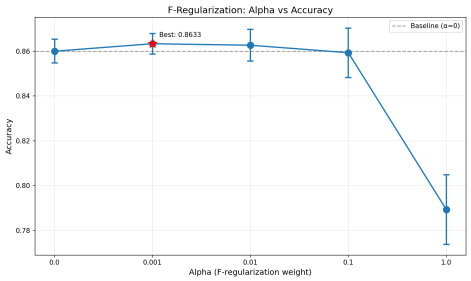

図 4: F 正則化強度 α と精度の関係．微弱な介入が性能を改善する一方，過剰な介入は阻害する．

5. 関連研究

5.1 Adaptive Computation Time

推論コストを動的に制御する試みとして，DeeBERT[5] や PonderNet[6] がある．これらはエントロピーや分類信頼度を停止基準とするが，モデル内部の「構造獲得」自体を評価指標にはしていない．geDIG はグラフ構造としての成熟度（F 値）を基準とする点で新規性があり，特に生成タスクのような信頼度定義が難しい場面での応用が期待される．

5.2 Free Energy Principle in AI

FEP を AI に応用する研究は強化学習分野で盛んだが，Transformer のような静的モデルの内部解析に適用した例は少ない．本研究は，Attention 機構自体を「能動的推論を行うエージェント群」とみなす新たな視点を提供する．

6. 考察とおわりに

本研究では，マクロ（迷路）とミクロ（Attention）という異なるスケールにおいて，geDIG ゲージ $\mathcal{F}$ が共通の「構造化原理」として機能することを示した．

Transformer は熱力学的推論を行うか？我々の観測結果は Yes を示唆している．Attention 層は単なる行列演算ではなく，入力情報の不確実性（エントロピー）を減じ，意味的な近道（最短路）を発見・固定化する熱力学的プロセスとして解釈できる．

次世代アーキテクチャへの展望現在の Transformer は，情報が既に構造化された後（Layer 4 以降）も，固定的に最終層まで計算を続けている．geDIG に基づけば，$\mathcal{F}$ がある閾値を下回って安定した時点で計算を打ち切る **Dynamic Depth (Early Exit)** や，推論していない時間にキャッシュを再構築する **Sleep Phase** の導入が可能になる．本研究で提案した統一ゲージは，そのような「計算資源を自律的・適応的に配分する次世代 AI」の基礎理論となりうると考える．

参考文献

[1] Friston, K.: The free-energy principle: a unified brain theory?, Nat. Rev. Neurosci., 11, 127–138 (2010).

[2] Grünwald, P. D.: The Minimum Description Length Principle, MIT Press (2007).

[3] Vaswani, A., et al.: Attention Is All You Need, NeurIPS (2017).

[4] Lewis, P., et al.: Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, NeurIPS (2020).

[5] Xin, J., et al.: DeeBERT: Dynamic Early Exiting for Accelerating BERT Inference, ACL (2020).

[6] Banino, A., et al.: PonderNet: Learning to Ponder, ICML (2021).