Attention 機構の熱力学的解釈：geDIG ゲージによる構造評価と F 正則化

Thermodynamic Interpretation of Attention: Structural Evaluation via geDIG Gauge and F-Regularization

宮内 和義 *1
Kazuyoshi Miyauchi

*1独立研究者
Independent Researcher

Transformer attention lacks a unified scalar metric for quality. We propose the geDIG gauge, which interprets attention as a directed graph and operationally bridges the Free Energy Principle and the Minimum Description Length by trading off structural cost (edit-path cost) against information gain (entropy reduction and relative shortest-path gain): $\mathcal{F} = \Delta\mathrm{EPC}_{\mathrm{norm}} - \lambda(\Delta H_{\mathrm{norm}} + \gamma\Delta\mathrm{SP}_{\mathrm{rel}})$. On BERT/GPT-2, real attention is more structured than random baselines and shows a sharp layer-wise transition. The same tendency is reproduced on Llama 3.0/3.1, while it is weaker on Phi-3. Finally, adding $\mathcal{F}$ as a regularization term in DistilBERT fine-tuning on SST-2 yields a small improvement under weak regularization, suggesting the gauge can serve as an actionable factor to intervene in internal structure.

1. はじめに

Transformer[1] は自然言語処理の基盤モデルとして広く普及しているが，Attention 機構の「質」を定量評価する統一指標は確立されていない．既存研究ではヘッドの pruning 基準 [8] や層の役割解釈 [7] が個別に議論されているが,「なぜその Attention パターンが良いのか」を原理的に説明する枠組みは不足している．

本研究では，Attention パターンを有向グラフとして構築し，自由エネルギー原理（FEP）[4] と最小記述長（MDL）[5] を橋渡しする geDIG ゲージで評価する手法を提案する．geDIG は元来，動的知識グラフにおける「いつ新しい知識を受け入れるか（When）」という判断基準として設計された指標である．本研究では，この枠組みを Transformer の内部表現に適用し,「Attention 層は情報の相転移的な遷移を示す可能性がある」という仮説を検証する．

本研究の貢献は以下の 3 点である：

1. geDIG の理論的背景（FEP-MDL 橋渡し）と Attention への適用手法の提示
2. 実 Attention が Random より構造化されていること，および層別の遷移的挙動の観測
3. F 正則化による因果的示唆（性能向上の観測）

2. geDIG の理論的背景

2.1 設計前提と直観

geDIG は「構造を更新するか否か」を判断するためのゲージであり，以下の前提に立つ：

1. **構造はグラフで表現できる**：知識や関係は頂点と辺で表され，編集はエッジ追加・削除として表現できる．
2. **更新にはコストがある**：構造を変えるほど編集コスト（複雑さ）が増え，過剰な更新は不利である．

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

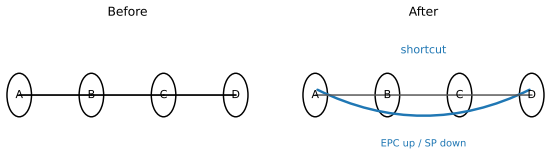

図 1: geDIG の直観的解釈：ショートカット追加で編集コストが増える一方，平均最短路が短縮される．

3. **更新には価値がある**：良い更新は不確実性を減らし，経路を短縮して推論効率を高める．

この 3 点を同時に満たすとき,「コストと利得の釣り合い」を単一スカラーで扱える．geDIG はそのための統一ゲージである．

(1) 直観例：ショートカット追加

鎖状グラフに 1 本のショートカットを追加する場合，編集コスト（$\Delta\mathrm{EPC}_{\mathrm{norm}}$）は増えるが，平均最短路の改善（$\Delta\mathrm{SP}_{\mathrm{rel}}$）が大きければ，総合的には「構造が良くなった」と判断される．一方，無秩序に多数のエッジを追加しても経路短縮が起きなければ，コストだけが増えて不利になる．geDIG はこの直観を形式化し，構造編集の是非を数値で判定できるようにする．

2.2 問題設定：「When」の判断基準

動的に変化する知識グラフ（または情報構造）に新たな要素が注入される場面を考える．このときシステムは，その要素を統合すべきか，棄却すべきかを判断しなければならない．この判断には，**構造編集によるコスト**と**情報利得**の天秤が必要である．

従来の RAG（Retrieval-Augmented Generation）は「何を取るか（What）」の最適化に長けるが,「いつ受け入れるか（When）」の規範を欠く [11]．geDIG はこのギャップを埋めるために設計された統一ゲージである．

2.3 自由エネルギー原理と MDL の橋渡し

geDIG は 2 つの理論的枠組みを操作的に橋渡しする：

自由エネルギー原理（FEP）：Friston の自由エネルギー原理

表 1: geDIG の各項の意味と理論的対応

項	意味	理論対応	符号
$\Delta\mathrm{EPC}_{\mathrm{norm}}$	編集経路コスト	MDL: $L(M)$	大→悪
ΔH_{norm}	エントロピー差	FEP: 驚き	負→秩序化
$\Delta\mathrm{SP}_{\mathrm{rel}}$	相対経路ゲイン	MDL: 圧縮	正→効率化
λ, γ	重み係数	情報温度	–

[4] は，生物システムが「予測誤差（驚き）」を最小化するように行動・学習するという枠組みである．geDIG では，エントロピー項 ΔH が「情報の曖昧さ・不確実性」に対応し，これを低減することが「秩序化」として解釈される．

最小記述長（MDL）：MDL 原理 [5] は，データとモデルの総記述長 $L(M) + L(D|M)$ を最小化する枠組みである．geDIG では，構造コスト $\Delta\mathrm{EPC}_{\mathrm{norm}}$ が $L(M)$（モデルの複雑さ）に，情報利得 $\Delta\mathrm{IG}_{\mathrm{norm}}$ が $L(D|M)$（データの圧縮効率）に対応する．

この対応により，geDIG は「構造の複雑さ」と「情報の整理度」のトレードオフを単一スカラーで評価する．

2.4 統一ゲージの定義

geDIG ゲージ $\mathcal{F}$ を以下のように定義する：

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

ここで $\Delta\mathrm{IG}_{\mathrm{norm}} = \Delta H_{\mathrm{norm}} + \gamma \cdot \Delta\mathrm{SP}_{\mathrm{rel}}$ とおくと，$\mathcal{F} = \Delta\mathrm{EPC}_{\mathrm{norm}} - \lambda\Delta\mathrm{IG}_{\mathrm{norm}}$ と書ける．各項の意味を表 1 に示す．

解釈：本実装では $\mathcal{F}$ が負の範囲に収まりやすく，**絶対値よりもベースラインとの差**が重要である．実 Attention は Random より $\mathcal{F}$ が 0 に近づき，**より構造化された状態**と解釈する．したがって本稿では $\Delta F = F_{\mathrm{real}} - F_{\mathrm{random}}$ を主要な比較指標として扱う．

2.5 二段ゲート：AG/DG

geDIG は本来，0-hop と multi-hop の二段階で評価を行う：

AG（Attention Gate）：0-hop 評価．局所的な編集直後の状態を評価し，「曖昧さ・不確実性」を検知する．$g_0 > \theta_{\mathrm{AG}}$ ならば探索を深化させる（FEP 的な予測誤差最小化）．

DG（Decision Gate）：Multi-hop 評価．複数ステップ先までの構造効率を評価し，「短絡（ショートカット）」の形成を確認する．$g_{\min} < \theta_{\mathrm{DG}}$ ならば統合を確定する（MDL 的な記述長削減）．

本研究の Attention 適用では，この二段構造を「層をまたいだ情報の流れ」として再解釈する．

3. 関連研究と位置づけ

Attention の解析や重要度評価に関しては，ヘッドの可視化や役割分担の分析 [7]，および剪定（pruning）に向けた重要度推定 [8] が広く用いられてきた．これらは有用だが，**なぜその Attention が良いのか**を説明する原理的指標は明確でない．

一方，FEP[4] や MDL[5] は，それぞれ「予測誤差の最小化」，「記述長の最小化」という強い理論的枠組みを提供する．geDIG はこれらを操作的に橋渡しし，構造編集の価値を単一スカラーで評価する点に特徴がある．本稿の位置づけは，Attention 解析を**構造評価のゲージ**として再定義する試みである．

4. Attention への適用

4.1 Attention からグラフへの Mapping

1 つの Attention ヘッドの重み行列 $\mathbf{A} \in \mathbb{R}^{L\times L}$（$L$ はシーケンス長）から有向グラフ $G = (V, E)$ を構築する．

- **頂点**：$V = \{1, \ldots, L\}$（トークン位置）
- **辺**：$E = \{(i, j) \mid A_{ij} > \tau\}$（閾値 τ を超える接続）

閾値 τ は上位 10 パーセンタイルとし，Pad トークンおよび causal mask（GPT 系）を適用後に評価する．

4.2 各項の計算方法

Attention 適用での各項の計算を以下のように定める：

- $\Delta\mathrm{EPC}_{\mathrm{norm}}$：エッジ密度 $= |E|/L^2$
- ΔH_{norm}：Attention 分布のシャノンエントロピーを $\log L^2$ で正規化
- $\Delta\mathrm{SP}_{\mathrm{rel}}$：最大弱連結成分での平均最短路長の相対ゲイン

パラメータは $\lambda = 1.0$，$\gamma = 0.5$ に固定した．シーケンスが長い場合，最短路計算はサンプリング（200 ペア）で近似する．

4.3 閾値設定と感度

Attention からグラフを構成する際の閾値 τ は結果の符号や分布に影響しうる．本研究では**パーセンタイル閾値**（上位 10%）を採用し，モデル間・層間での比較が安定することを優先した．固定の絶対閾値では，スケール差により ΔF の符号が反転することがあるため，補助的な分析に留める．

4.4 仮説：Transformer 推論は自由エネルギー最小化的挙動

上記の枠組みに基づき，以下の仮説を立てる：

H1：実 Attention は，ランダム/一様ベースラインより高い F 値（0 に近い＝より構造化）を示す．

H2：深層ほど F 値が上昇（0 に接近）し，情報が「探索相」から「構造相」へ移行する．

H3：F を損失に組み込んだ学習は，下流タスク性能に因果的な影響を与える可能性がある．

4.5 ベースライン

比較のため以下のベースラインを設定した：(1) **Random**：ランダム行列，(2) **Uniform**：一様分布，(3) **Local**：窓幅 $w = 5$ の局所 Attention，(4) **Diagonal**：対角成分のみ．

5. 実験 1：記述的分析

5.1 実験設定

モデルとして bert-base-uncased[2]（12 層，12 ヘッド）および gpt2[3]（12 層，12 ヘッド）を使用した．Wikitext[6] から抽出した 200 件の短文サンプル（最大長 512 トークン）で評価し，各層・各ヘッドの F 値を算出した（計 $200 \times 12 \times 12 = 28,800$ サンプル/モデル）．追加検証として Llama 3.0/3.1（各 32 層，32 ヘッド）と Phi-3-mini-4k-instruct を同様に評価した [12, 13, 14]．Wikitext の短文 63 件（最大 120 文字，最大 256 トークン）を用い，計 $63 \times 32 \times 32 = 64,512$ サンプル/モデルを計測した．

表 2: 平均 ΔF（Real - Baseline，数値は小数第 3 位で丸め）

モデル	ΔF_{rand}	ΔF_{uni}	ΔF_{local}	ΔF_{diag}
BERT	+0.11	–	–	–
GPT-2	+0.10	–	–	–
Llama 3.0	+0.116	+0.108	+0.190	+0.074
Llama 3.1	+0.116	+0.108	+0.190	+0.074
Phi-3	+0.060	+0.052	+0.144	+0.018

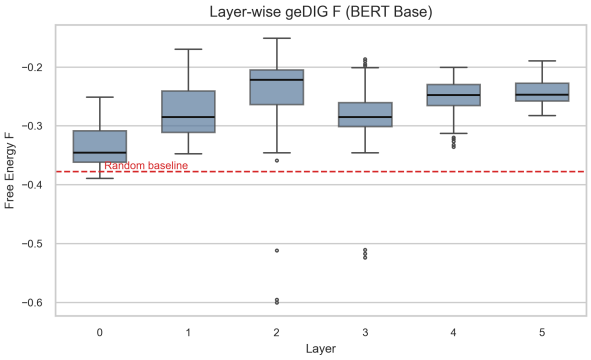

図 2: BERT の層別 F 値分布（箱ひげ）．赤点線は Random baseline の平均．

5.2 結果 1：Real vs Random（H1 の検証）

表 2 に平均 ΔF（Real - Baseline）を示す．BERT/GPT-2 では Random のみ計測したが，両モデルで実 Attention は有意に高い F 値（0 に近い＝より構造化）を示す傾向があり，効果量は $d > 2.0$ と大きかった（$p < 0.001$）．ここでは F 値そのものを示すが，評価は ΔF（Real-Random）の符号と大きさに基づいて行う．

F 値の差 $\Delta F \approx 0.11$ は，操作的には「仕事」に相当する量として解釈できるが，本稿では比喩的解釈に留める．これはモデルがランダムノイズから「秩序」を獲得したことを示す一つの定量的示唆である．

追加検証では，Llama 3.0/3.1 で $\Delta F_{\text{random}} \approx +0.116$（正の比率 ≈ 0.995）となり，BERT/GPT-2 と同傾向が再現された．Phi-3 では $\Delta F_{\text{random}} \approx +0.060$（正の比率 ≈ 0.73）と効果が弱いが，Uniform/Local/Diagonal の各ベースラインでも符号は維持された．

H1 は概ね支持された．

5.3 結果 2：層別の遷移（H2 の検証）

図 2 に BERT の層別 F 値分布を示す．

- **Layer 0**（$F \approx -0.34$）：高エントロピー,「探索相」
- **Layer 1–2**（$F \approx -0.24$）：急激な上昇,「遷移」
- **Layer 3+**（$F \approx -0.25$）：プラトー,「構造相」

これは物理における相転移（気体 → 結晶）に類似した変化とみなせるが，ここでは比喩として扱う．浅層では多様な情報を広く収集し（探索），深層では関連情報を効率的に統合する（構造化）と解釈できる．

追加検証では，Llama 3.0/3.1 は全層で $\Delta F_{\text{random}} > 0$ だった一方，Phi-3 は Layer 0 で $\Delta F_{\text{random}} < 0$ となり，浅層の探索性が強い可能性がある．

H2 は概ね支持された．

表 3: F 正則化の結果（3 シード平均）

α	Accuracy (%)	Final F
0 (baseline)	86.00 ± 0.53	-0.448
0.001	$\mathbf{86.33 \pm 0.46}$	-0.447
0.01	86.27 ± 0.58	-0.452
0.1	85.93 ± 1.01	-0.466
1.0	78.93 ± 2.00	-0.515

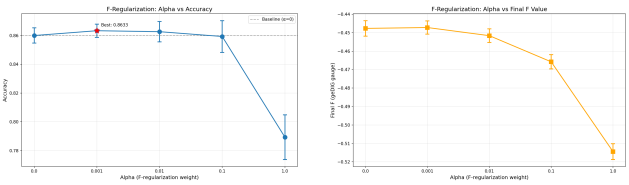

図 3: F 正則化の強さ α と指標の関係．左：精度（弱い正則化で改善）．右：最終 F（強い正則化ほど低下）．

5.4 結果 3：ヘッド多様性

同一層内でもヘッド間で F 値に差異が見られた．これは各ヘッドが異なる熱力学的状態を持ち，**マルチエージェント的**に振る舞う可能性を示唆する．geDIG はヘッドの「役割」を定量的に診断する指標となりうる．

6. 実験 2：因果的示唆（F 正則化）

6.1 動機

実験 1 は相関を示すが因果は示さない．そこで，**F を損失項に組み込んだ学習**が性能に与える影響を検証し，geDIG の因果的示唆を検討する．

6.2 実験設定

DistilBERT[9] を SST-2[10]（感情分析）で fine-tuning し，損失関数を以下のように設定した：

$$L_{\text{total}} = L_{\text{CE}} + \alpha \cdot F_{\text{mean}} \tag{2}$$

ここで F_{mean} は全層・全ヘッドの F 値の平均である．$\alpha \in \{0, 0.001, 0.01, 0.1, 1.0\}$ を sweep し，各設定で 3 シード（42, 123, 456）の平均を報告する．学習データ 2000 件，評価データ 500 件，3 エポック，バッチサイズ 16 で実験した．

6.3 結果

表 3 に F 正則化の結果を示す．

精度は逆 U 字カーブを描き，$\alpha = 0.001$ で最高精度 **86.33%**（+0.33%），$\alpha = 1.0$ で 78.93%（-7%）となった．

この結果は以下を示す：(1) **弱い F 正則化は改善傾向**：損失に F を組み込むことで性能向上に寄与，(2) **強すぎると有害**：過度の正則化はタスク固有の情報を損なう，(3) **geDIG は訓練目的の候補**：F の寄与量を制御することで性能に影響を与えうる．

H3 を支持する傾向が得られた．

7. 考察

7.1 理論的含意

本研究の結果は，Transformer 推論を「自由エネルギー最小化的プロセス」として解釈する視点を支持する可能性を示す．

- **各層**：情報の遷移（探索相 → 構造相）を実現

- **各ヘッド**：異なる熱力学的状態を持つマルチエージェント的挙動
- **geDIG F**：内部状態を追跡するゲージ

これは Attention を「情報の流れ」としてだけでなく,「構造の形成プロセス」として理解する新しい視点を提供する．Phi-3 で効果が弱いのは，蒸留や学習データ特性により注意分布の初期エントロピーが低く，ΔF が縮小した可能性がある．

7.2 限界と今後の課題

本研究には以下の限界がある：

- **規模**：DistilBERT + SST-2 という小規模設定
- **効果量**：+0.33%は統計的に有意だが実用的には小さい
- **統計**：層/ヘッド/文が入れ子構造であり，独立性仮定が厳密ではない
- **理論**：F 正則化が「なぜ」効くかの厳密な説明は未確立

今後の課題として，大規模モデル（GPT-4 クラス）での検証，多様なタスクでの再現性確認，および geDIG と情報理論的指標（mutual information など）との関係解明が挙げられる．

8. 結論

本研究では，geDIG ゲージによる Attention 品質の熱力学的評価手法を提案した．FEP-MDL 橋渡しという理論的背景のもと，(1) 実 Attention が Random より構造化されていること（$d \approx 2.3$），(2) 層別の遷移的挙動，(3) F 正則化による微小な性能向上（+0.33%）を示した．さらに Llama 3.0/3.1 で同傾向が再現される一方，Phi-3 では効果が弱いことを確認した．

これらの結果は，Transformer を「自由エネルギー最小化的に解釈する枠組み」を補強し，Attention 機構の設計・最適化に新しい視点を与えるものである．

参考文献

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

[2] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, NAACL-HLT (2019).

[3] Radford, A., et al.: Language Models are Unsupervised Multitask Learners, OpenAI Technical Report (2019).

[4] Friston, K.: The free-energy principle: a unified brain theory?, Nature Reviews Neuroscience, Vol.11, pp.127–138 (2010).

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

[6] Merity, S., Xiong, C., Bradbury, J., Socher, R.: Pointer Sentinel Mixture Models, ICLR (2017).

[7] Clark, K., et al.: What Does BERT Look At? An Analysis of BERT's Attention, BlackboxNLP (2019).

[8] Voita, E., et al.: Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned, ACL (2019).

[9] Sanh, V., Debut, L., Chaumond, J., Wolf, T.: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, arXiv:1910.01108 (2019).

[10] Socher, R., et al.: Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank, EMNLP (2013).

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

[12] Meta AI: Llama 3 8B Model Card, Hugging Face (2024), `https://huggingface.co/meta-llama/Meta-Llama-3-8B`.

[13] Meta AI: Llama 3.1 8B Model Card, Hugging Face (2024), `https://huggingface.co/meta-llama/Llama-3.1-8B`.

[14] Microsoft: Phi-3 Mini 4K Instruct Model Card, Hugging Face (2024), `https://huggingface.co/microsoft/Phi-3-mini-4k-instruct`.