

「階層的意味構造モジュール群によるトランスフォーマーのXAI透明性と性能の統合的向上：人間との共進化を目指した意味構造生成（SSG）の基礎研究」

「XAIアウトカム層」の組み込みと補強

「情報生命体原理に着想を得たAI共進化への一歩」

「精錬ループ導入（実験コード）」

実験計画（冒頭に挿入）

エコテラボ合同会社
猪澤也寸志
polyp@webman.jp
(2025年6月1日現地時間JST)

目標1: 性能の維持・向上 (指標、アブレーション実験設定、結果、考察は前回合意内容から変更なし。調整パラメータを追記)

- 指標1 (定量): Accuracy (分類精度)
- 指標2 (定量): Loss (評価時損失値)
- 指標3 (定性→準定量): 学習過程の安定性と収束性 (訓練損失・検証損失の推移グラフ、収束エポック数)
- 指標4 (定性→準定量): 誤分類サンプルの傾向分析と、それに対する各モデルの頑健性 (特定パターンの誤分類率など)

アブレーション実験設定 (注: 実験構成名および参照コードは「0601PDF」参照。訓練データ500件、評価データ100件、5エポックの統一条件。)

- 比較1: ベースライン性能の確立 (E0_Baseline)
 - 調整パラメータ: 学習率、バッチサイズ、オプティマイザの種類、乱数シード。
- 比較2: 意味アテンション機構 (ループなし) の性能評価 (E1_SemanticAttention_NoLoop)
 - 調整パラメータ: 上記に加え、SemanticConceptModuleのnum_concepts, concept_dim。
- 比較3: 反復精錬ループ付き意味アテンション機構の性能評価 (E1R_SemanticAttention_WithLoop)
 - 調整パラメータ: 上記に加え、反復精錬ループのmax_iterations, convergence_threshold。

結果 (前回記載のE0, E1, E1RのAccuracy, Lossに加え、指標3,4に関する分析結果を追記する形)

考察 (前回記載の考察に加え、指標3,4の分析結果から、各モデルの学習の質や、どのような種類の入力に強く/弱いといった特性の違いを論じる)

目標2: 豊かな「意味構造」の生成 (SSG)

- **指標1 (定量)**: 意味役割付与 (SRL) タスクにおけるF1スコア、Accuracy、Loss
(SemanticConceptModuleの「意味概念」生成能力の代理評価として)
- **指標2 (定量)**: 依存構造解析タスクにおけるUAS/LAS、Loss (DirectNonCommutativeModuleの「非可換構造」認識能力の代理評価として)
- **指標3 (定量)**: 質問応答 (QA) タスクにおけるEM/F1スコア、Loss
(DiscontinuityEmergenceModuleの「文脈的断絶からの創発」能力の代理評価として)
- **指標4 (準定量)**: 意味概念ネットワークの構造的特徴量
 - 活性化する意味概念の数、種類、一貫性。
 - (もし概念間の関連性をモデル化する場合) ネットワークの密度、クラスタリング係数など。
- **指標5 (準定量)**: 「情報の破れ」と「創発ラベル」の文脈的妥当性スコア
 - 人間が事前に定義した基準に基づき、検出された「断絶」の妥当性と、生成された「創発ラベル」の新規性・適切性を複数人で評価し、評定者間信頼性と平均スコアを算出。
- **指標6 (準定量)**: 非可換性の影響度マップにおける特定パターン検出率
 - 既知の非可換性が重要な構文パターン (例: 否定の反転、受動態など) を含むテストセットに対し、DirectNonCommutativeModuleが期待される箇所で強い影響度を示した割合。

アブレーション実験設定

- **比較1: SemanticConceptModuleによる意味役割理解の評価 (SRL)**
 - 調整パラメータ: SRLタスク用の出力層の設計、損失関数の選択、事前学習済みモデルのファインチューニング戦略、SemanticConceptModuleのnum_concepts, concept_dim、反復精錬ループのパラメータ (EIR構成の場合)。
- **比較2: DirectNonCommutativeModuleによる統語構造理解の評価 (依存構造解析)**
 - 調整パラメータ: 依存構造解析用の出力層設計、損失関数、DirectNonCommutativeModuleの内部パラメータ (例: 畳み込みカーネルサイズ、層数など)。
- **比較3: DiscontinuityEmergenceModuleによる文脈理解と推論の評価 (QA)**
 - 調整パラメータ: QAタスク用の出力層設計、損失関数、DiscontinuityEmergenceModuleの「断絶」検出閾値や「創発ラベル」生成メカニズムのパラメータ。
- **比較4 (今後の実験): フルSSG構成における階層的「意味構造」の評価**
 - 調整パラメータ: 各SSGモジュールの組み合わせ順序、モジュール間の情報伝達方法、各モジュールの内部パラメータ。

結果 (指標1,2,3については各タスクの定量的スコアを記載。指標4,5,6については、定義した準定量的な評価方法に基づき算出したスコアや統計量を記載。現時点では多くが「今後の実験で取得予定」。)

- **SemanticConceptModule (SRLタスク)**: F1/Accuracy/Loss: [取得予定]
- **DirectNonCommutativeModule (依存構造解析タスク)**: UAS/LAS/Loss: [取得予定]
- **DiscontinuityEmergenceModule (QAタスク)**: EM/F1/Loss: [取得予定]
- **意味概念ネットワークの構造的特徴量**: [取得予定]
- **「情報の破れ」と「創発ラベル」の文脈的妥当性スコア**: [取得予定]
- **非可換性の影響度マップにおける特定パターン検出率**: [取得予定]

考察（各モジュールが、割り当てられたタスクの定量的性能、および指標4,5,6の準定量的評価を通じて、どの程度豊かで、かつ人間が期待するような「意味構造」の側面を捉えられているかを論じる。ベースライントランスフォーマーがこれらの指標でどのような値を示すか（あるいは測定不能か）との比較も重要。）

目標3: XAI透明性の飛躍的向上

- ・ **指標1（間接的・定量）**：目標2でSSGモジュール群が示した各種「意味構造」理解タスク（SRL, 依存構造解析, QA）の性能。（目標2の結果を援用）
 - ・ **指標2（準定量）**：カウンターファクチュアルな入力に対する応答の論理的整合率
 - 入力テキストの一部を意図的に変更した場合に、SSGモジュールの内部状態（意味概念、断絶スコア、創発ラベルなど）および最終出力が、論理的に期待される方向に変化した事例の割合。
 - ・ **指標3（準定量）**：人間による判断根拠の特定タスクにおける正解率・時間
 - モデルの判断（例：IMDbの感情ラベル）に対し、SSGモジュールが提示する「意味構造」の断片（例：活性化した意味概念、注目箇所）を見て、人間がその判断根拠をどの程度正確に、かつ迅速に特定できるかを測定。
 - ・ **指標4（準定量・将来的）**：ユーザー評価実験における「説明の納得度」「判断根拠の理解度」スコア。
- アブレーション実験設定**

- ・ **比較1: 意味アテンション機構（特にループ付きEIR）における判断プロセスの追跡**
 - **調整パラメータ**: SemanticConceptModuleのパラメータ、反復精錬ループのパラメータ。
 - **分析**: ループの各イテレーションで活性化する意味概念やアテンションの変化を可視化し、それが最終判断にどう収束していくかをケーススタディで分析。この過程が人間にとって直感的か評価。
- ・ **比較2: カウンターファクチュアル分析によるロバスト性と判断根拠の検証**
 - **調整パラメータ**: 入力テキストの変更方法（単語置換、順序変更、情報削除など）の設計。SSG各モジュールの感度。
 - **分析**: 変更に対するSSGモジュール群の内部状態と出力の変化を記録し、その変化が人間の直感や論理と整合しているかを評価。
- ・ **比較3 (今後の実験): XAIアウトカム層を介したユーザーによる判断根拠特定タスク**
 - **調整パラメータ**: XAIアウトカム層の設計、提示する情報の種類と粒度。
 - **分析**: ユーザーが判断根拠を特定する際の正解率、所要時間、主観的評価（わかりやすさ等）を記録。

結果（指標1は目標2の結果を援用。指標2,3については、設計した実験プロトコルに従って測定した数値や成功事例数を記載。現時点では多くが「今後の実験で取得予定」。）

- ・ **SSGモジュールの各種「意味構造」理解タスク性能**: 目標2の結果を参照。
- ・ **カウンターファクチュアル応答の論理的整合率**: [取得予定]
- ・ **人間による判断根拠の特定タスク（正解率・時間）**: [取得予定]
- ・ **ユーザー評価実験スコア**: [取得予定]

考察（SSGモジュール群、特に「ホログラフィック精錬ループ」などが、モデルの内部状態をより追跡可能にし、判断の論理性を高めているかを、指標2,3の結果に基づいて論じる。これが、トランスフォーマーのブラックボックス性をどの程度克服し、XAI透明性の向上に貢献しているかを評価する。）

目標4: 「ホログラフィック精錬ループ」等による表現の効率性と計算資源への配慮

- **指標1（定量）**：NLIタスクにおけるAccuracy、Loss（InterpretableHolographicReducerの評価として）
- **指標2（定量）**：InterpretableHolographicReducer適用前後の特徴ベクトルの次元削減率、および情報損失の度合い（例：再構成タスクを設定した場合の再構成誤差）。
- **指標3（定量）**：InterpretableHolographicReducer適用による推論時間（Runtime_s）の変化、およびモデル全体のパラメータ数の削減効果。
- **指標4（定量）**：「ホログラフィック精錬ループ」（意味アテンション内）の平均実効反復回数と収束性、およびそれによる推論時間への影響（目標1の結果と関連）。
- **指標5（準定量）**：〇〇様の提案する「要約出力（問い／結果／理由／可謬率）」形式での定性評価
 - InterpretableHolographicReducer適用後の最終的なSSGモデルの出力に対し、この形式で「理由」の質と「可謬率」の妥当性を人間が評価。

アブレーション実験設定

- **比較1: InterpretableHolographicReducerによるNLIタスク性能と効率性の評価**
 - 調整パラメータ: NLIタスク用の出力層設計、損失関数、InterpretableHolographicReducerのnum_key_concepts, key_concept_dim, 圧縮率などの内部パラメータ。
- **比較2: 「ホログラフィック精錬ループ」の計算コストと収束性の評価**（目標1と共通の実験E1, E1R）
 - 調整パラメータ: ループのmax_iterations, convergence_threshold。
- **比較3: フルSSG構成における「要約出力」による総合的定性評価**
 - 実験構成: 複数のSSGモジュールとInterpretableHolographicReducerを組み合わせた最終的なモデル構成。
 - 目的: モデル全体としての情報処理の質、説明能力、自己の不確実性認識能力を「要約出力」形式で評価する。
 - 調整パラメータ: 各SSGモジュールのパラメータ、XAIアウトカム層での情報提示方法。

結果（指標1,2,3,4については、NLIタスクの性能、次元削減率、Runtime、パラメータ数、平均反復回数などを記載。多くが「今後の実験で取得予定」。指標5の「要約出力」の評価結果は、具体的な評価基準と事例に基づいて記述。）

- **InterpretableHolographicReducer (NLIタスク):** Accuracy/Loss: [取得予定]
- **次元削減率・情報損失:** [取得予定]
- **Runtime・パラメータ数:** [取得予定。一部は「0601PDF」p.75のE1R vs E1参照]
- **ホログラフィック精錬ループの平均反復回数:** [取得予定]
- **「要約出力」評価:** [取得予定]

考察（InterpretableHolographicReducerが、情報の本質を保持しつつどの程度の効率化（圧縮、速度向上、パラメータ削減）を達成できたかを論じる。それが「情報の呼吸」における効率的な「縮減」にどう貢献するか。「ホログラフィック精錬ループ」の計算コストと、それがもたらす性能やXAIへの貢献とのバランスを評価する。「要約出力」の評価結果から、モデル全体としての説明能力と自己認識能力のレベルを考察する。）

目標5: 人間との共進化基盤の構築（SSGとXAIを基盤として）（この目標の定量評価は非常に挑戦的であり、本論文の範囲では主に概念実証や将来展望に留まる可能性が高いです。準定量的な評価の方向性を示します。）

- **指標1（準定量・将来的）**：人間によるフィードバック（例：「要約出力」の「理由」や「可謬率」の修正指示）をシステムが取り込んだ後の、タスク性能や「意味構造」の質の変化量。

- ・ **指標2 (準定量・将ラリー)** : 人間とAIが協調して特定の複雑なタスク (例: 複数文書からの情報統合と要約) を解決する際の、タスク成功率、所要時間、人間の主観的満足度、AIへの信頼度の変化。
- ・ **指標3 (準定量・将来的)** : 特定のコンセプトや推論パターンについて、人間からの教示 (数ショット学習的なものも含む) に対するAIの学習効率 (必要な教示回数など) 。

アブレーション実験設定

- ・ **比較1 (概念実証・プロトタイプ評価): XAIアウトカム層を介したフィードバックループ**
 - **調整パラメータ:** XAIアウトカム層の設計、フィードバックの種類 (誤り指摘、理由の修正、新規概念の教示など)、フィードバックをモデルの学習/推論に反映させるメカニズム (重み更新、メモリへの追加など) 。
 - **分析:** 小規模なインタラクション実験を通じて、人間のフィードバックが実際にモデルの振る舞いや内部状態 (意味構造) を望ましい方向に変化させるかを観察・記録。

結果 (プロトタイプシステムによる小規模な実験結果や、ユーザーインタラクションの観察記録などを記述。現時点では「概念実証の結果 (取得予定)」となる。)

- ・ **フィードバックによるAIの改善 (事例)** : [取得予定]
- ・ **人間とAIの協調タスク (初期試行)** : [取得予定]

考察 (SSGによる「意味構造」の生成 (目標2) とXAIによるその透明化 (目標3) が、人間とAIの間のより効果的なコミュニケーションと相互理解をいかに可能にし、それを通じてAIが人間の意図や知識を学習し、人間もまたAIの能力を理解し活用していくという「共進化」のループが形成されるか、そのメカニズムと初期的な証拠を論じる。〇〇様の「情報生命体原理」の観点から、この共進化プロセスを考察する。)

アブストラクト (Abstract / 抄録)

- ・ 現代のトランスフォーマーモデルは多くのタスクで高い性能を達成しているが、その判断プロセスの不透明性 (ブラックボックス性) は、AIの信頼性、安全性、社会受容における根本的な課題である。また、人間のような深い意味理解や文脈の順序性、構造的情報の扱いは依然として発展途上にある。本研究は、これらの課題に対し、AIと人間が「共進化」するための基盤技術として、トランスフォーマーに複数の階層的な「意味構造生成 (SSG)」モジュールを導入し、そのXAI透明性と性能への貢献を検証する。提案するモジュール群は、(1)解釈可能な「意味概念」を学習・活用する「意味アテンション機構」、(2)トークンシーケンス間の順序依存性や「非可換な」関係性を陽にモデル化する「直接的非可換処理モジュール」、(3)意味的な「断絶」を検出し、そこから新たな「創発的」意味情報を生成する「断絶創発モジュール」、そして(4)これらの処理で得られたリッチな意味構造を、解釈可能性を意識しつつ効率的に集約する「解釈可能なホログラフ縮減モジュール」である。これらのモジュールをIMDb感情分析タスクにおけるベースライントランスフォーマーに様々な組み合わせ・順序で統合し、包括的なアブレーションスタディを実施した。
- ・ 実験の結果、[ここにE0〜E18の実験から得られる最も重要な定量的結果を具体的に記述。例: 特定のモジュールの階層的組み合わせ、特に意味アテンション機構を初期段階に配置し、非可換処理、断絶創発処理を続けた構成 (E6) が、ベースライン (Accuracy X.XX) に匹敵する性能 (Accuracy Y.YY) を示した。他の単独モジュールや一部の組み合わせでは性能低下も見られたが、これは学習

の複雑性や現行のモジュール設計の課題を示唆する。ホログラフ縮減モジュールは、有望な多層構造の最終段に適用することで、特徴量の次元数をZ%削減しつつ、性能低下をW%以内に抑制できる（あるいは特定条件下で向上する）可能性を示した。など]。

- さらに、各モジュールの内部状態（活性化した意味概念、非可換効果のパターン、断絶スコアと創発ラベルの具体例、鍵概念プロトタイプの学習結果など）の定性分析を通じて、それぞれがモデルの判断プロセスの解釈（ホワイトボックス化の試み）にどのように貢献しているかを具体的な事例とともに論じる。例えば、意味アテンション機構は特定の感情に関連する概念を学習し、非可換処理は文の構造的反転を検出し、断絶創発モジュールはレビュー内の論点の転換を示唆する情報を生成した。
- 本研究は、トランスフォーマーのブラックボックス性を低減し、より深い意味理解、構造化された情報処理、効率的な表現獲得、そして将来的な倫理的推論や人間との共進化の基盤となり得る「SSG」の実現に向けた、多階層モジュラーアプローチの具体的な設計、実装、評価方法、そしてその初期的な有効性と課題を提示する。

キーワード: トランスフォーマー、XAI（説明可能なAI）、意味構造生成 (SSG)、階層的ニューラルネットワーク、意味アテンション、非可換性、創発的表現、ホログラフィック表現、アブレーションスタディ、人間とAIの共進化

1. 序論 (Introduction) * **1.1. 背景と研究の動機:** * トランスフォーマーの現状と、その「成功の影」としてのブラックボックス問題、意味理解の表層性、人間の直感や常識との乖離について、具体的な事例や引用を交えながら論じる。* XAIの重要性が叫ばれる中、単なる事後的な説明手法ではなく、モデルアーキテクチャ自体に解釈可能性を組み込むアプローチの必要性を強調する。* お客様の「AIが人間のタイムスケールから取り残され、暴走するリスク」という懸念に触れ、AIと人間が建設的に「共進化」するための技術的基盤の探求が本研究の根源的な動機であることを示す。* **1.2. 問題提起と本研究の目的:** * **問題点1（意味の不在とブラックボックス性）:** 標準トランスフォーマーは「何を」学習しているのか？その内部表現は人間にとって意味があるのか？判断根拠は追跡可能か？* **問題点2（構造と順序の軽視）:** テキストが持つ非可換な構造や、情報の「断絶」といった要素は、意味の理解に不可欠だが、標準トランスフォーマーで十分に扱えているか？* **問題点3（創発の方向性）:** LLMに見られる「アホ創発」のような制御不能な創発ではなく、人間にとって有益で解釈可能な「真の創発」（潜在的な法則やアイデアの新発見）をどのように促すか？* **問題点4（表現の効率性と汎用性）:** 多様な意味情報を捉えたりッチな表現を、いかに効率的に保持し、様々な状況（例：エッジデバイス）で活用可能な形にするか？* **本研究の目的:** 上記の問題群に対し、複数の専門的情報処理モジュールを階層的に組み合わせた「意味構造生成 (SSG)」アーキテクチャを提案する。各モジュール（意味アテンション、非可換処理、断絶創発、ホログラフ縮減）が、それぞれ特定の「意味の側面」を捉え、それらを統合することで、トランスフォーマーのXAI透明性を高め、かつ性能を維持・向上させることを目指す。本論文では、このSSGの基礎となる各モジュールの設計と、それらのアブレーションスタディによる初期的な有効性評価を報告する。* **1.3. 提案手法群の概要とコントリビューション:** (各モジュールの核心的アイデアと、それが上記のどの問題点に対応するかを明確に。以前の骨子をベースに具体化)* **1.4. 論文構成:** (変更なし)

2. 関連研究 (Related Work) * (以前の骨子の2.1~2.5の内容を、お客様の独自性と対比させながら記述) * **2.1. XAIとトランスフォーマーの解釈可能性向上手法:** 既存手法の限界（事後的説明、限定的な可視化など）と、本研究のアーキテクチャレベルでの解釈可能性追求の違いを強調。* **2.2. トランスフォーマーへの構造的情報の導入:** 既存研究が主にどのような「構造」に着目しているか（例：知識グラフ、構文木）を概観し、本研究が扱う「意味概念」「非可換性」「断絶」といった構造の新規性や組み合わせの独自性を論じる。* **2.3. 潜在概念・意味表現の学習:** 既存の潜在表現学習との比較。本研究の「意味概念プロトタイプ」や「創発意味ラベル」が、より具体的で解釈可能な意味単位の獲得を目指す点を明確に。* **2.4. ニューラルネットワークにおける順序性・非可換性・断絶検出:** 先行研究でどのようなアプローチが取られ、どのような成果と限界があったかを整理。本研究の「直接的な非可換処理モジュール」や「断絶創発モジュール」の独自性を強調。* **2.5. 情報圧縮、ベクトルシンボリックアーキテクチャ(VSA)、ホログラフィック表現:** HRRなどのVSA

の原理を簡潔に説明し、それらがニューラルネットワークとどのように統合されうるか、また本研究の「解釈可能なホログラフ縮減モジュール」がこれらの文脈でどのような新しい試みであるかを述べる。* **2.6. AI倫理と人間とAIの共進化:** お客様のビジョンを反映し、本研究が目指すSSGや将来的な倫理階層モジュールが、AIの自律性と人間の制御・理解のバランスを取り、より安全で信頼できるAI、ひいては人間と共進化するAIの実現にどう貢献しうるかの展望を、関連する哲学的・倫理的議論も踏まえて記述。

3. 提案手法群：多階層意味構造生成モジュール (Proposed Multi-Tier Semantic Structure Generation Modules)

* 3.1. 基本アーキテクチャと統合フレームワーク

(EnhancedTransformerForSequenceClassification) * 図1として、ベースTransformerに複数の技術モジュールが階層的・選択的に接続される全体像を示す。*

EnhancedTransformerForSequenceClassificationの__init__におけるtech_flags (特にmodule_order) とmodule_configsを用いたモジュールの動的なインスタンス化と接続、current_dimの引き継ぎロジックを詳細に説明。* forwardメソッドにおけるmodule_orderに従ったモジュールの逐次適用、各モジュール出力 (特にlast_hidden_stateと補助出力) のハンドリング、ホログラフ縮減モジュールが最後に適用される場合の分類器入力切り替えロジックなどを、擬似コードや数式を交えて説明。* **3.2. モジュールA：意味アテンション機構 (SemanticConceptModule):** (アーキテクチャ図、数式、学習されるパラメータ、期待される入出力、解釈方法などを詳細に記述) * **3.3. モジュールB：直接的非可換処理モジュール (DirectNonCommutativeModule):** (同上) * **3.4. モジュールC：断絶創発モジュール (DiscontinuityEmergenceModule):** (同上。特に「断絶検出」と「創発ラベル生成」の連携、創発ラベルが何を表現することを目指すのか、お客様の「準潜在的な要素の組み合わせでの創発」というアイデアを反映した設計意図を記述) * **3.5. モジュールD：解釈可能なホログラフ縮減モジュール**

(InterpretableHolographicReducer): (同上。「鍵概念プロトタイプ」とは何か、入力情報をどのように鍵概念に射影・縮減するのか、そのプロセスがXAIにどう貢献するのかを詳述) * **3.6. 学習目的と損失関数:** (変更なし、エンドツーエンド学習。ただし、将来的なSSGや倫理階層のための補助損失や「正解データ」に基づく学習の可能性について、より具体的に言及する。)

(DiscontinuityEmergenceModule): (同上。特に「断絶検出」と「創発ラベル生成」の連携、創発ラベルが何を表現することを目指すのか、お客様の「準潜在的な要素の組み合わせでの創発」というアイデアを反映した設計意図を記述) * **3.5. モジュールD：解釈可能なホログラフ縮減モジュール**

(InterpretableHolographicReducer): (同上。「鍵概念プロトタイプ」とは何か、入力情報をどのように鍵概念に射影・縮減するのか、そのプロセスがXAIにどう貢献するのかを詳述) * **3.6. 学習目的と損失関数:** (変更なし、エンドツーエンド学習。ただし、将来的なSSGや倫理階層のための補助損失や「正解データ」に基づく学習の可能性について、より具体的に言及する。)

4. 実験設定 (Experimental Setup) * (以前の骨子案の内容を、より具体的に、論文の読者が再現可能なレベルで記述) * **4.1. データセット:** IMDB。訓練・検証・テストの正確な分割方法とサンプル数を明記。なぜこのデータセットを選んだか (感情分析というタスクが、提案モジュールの効果検証に適している理由など)。*

* 4.2. 評価指標: 定量的指標の定義式。定性的評価の具体的な観点と手順。 * 4.3. 実験構成

(experiment_configurations E0~E18): * 全ての実験構成 (E0~E18) とその設定 (module_order, use_..._moduleフラグ、各モジュールの主要ハイパーパラメータ) をまとめた大きな表 (Table X) として提示。各構成の意図を簡潔に説明。* **4.4. 共通訓練ハイパーパラメータと実装詳細:** (詳細に記述)

5. 結果と考察 (Results and Discussion) * **5.1. 定量的結果の全体比較と主要な発見** * Table Yとして、全実験構成の主要評価指標 (Accuracy@Epoch3, Best Accuracy, Final Eval Accuracy, Corresponding Lossなど) をまとめた総合比較表を提示。統計的有意差検定の結果も (可能であれば)。* 図B, C, D...として、主要な構成 (例: ベースライン、各単独モジュール、最も性能の良かったE6、ホログラフ縮減系など) のAccuracyやLossを比較する棒グラフや、訓練中の学習曲線 (訓練損失 vs 検証損失) を提示。* **詳細な分析と考察:** * ベースラインの性能を基準として、各単独モジュール (E1-E3, E16) がどのような影響を与えたか (性能向上、低下、計算コストの変化など)。* 2モジュール構成 (E4, E5, E9-E12) の結果から、モジュール間の相性や順序依存性について何が言えるか。* 3モジュール構成 (E6-E8, E13-E15) の結果から、最も有望だった組み合わせ (E6) とその要因を深く考察。なぜ他の順序では性能が劣ったのか? * ホログラフ縮減モジュール (E16, E17, E18) の効果。特にE17 (E6の最後に適用) とE6を比較し、縮減が性能と表現のコンパクトさに与える影響を議論。E18のような多段ホログラフ構成の可能性と課題。* 全体を通して、「意味文脈最適化」がどの構成で最も達成されたように見えるか、その根拠は何か。* **5.2. XAI透明性 (ホワイトボックス効果) に**

関する定性的分析（各モジュールについて、最も有望だった構成を中心に具体例を多用）* (以前の骨子の5.2.1～5.2.2、5.3.1～5.3.2、そして断絶創発モジュールとホログラフ縮減モジュールの分析を、実際の実験で得られたであろう具体的な「気づき」や「事例」を想像して肉付けする。図のキャプションも重要。)* 例えば、「E6構成において、ある誤分類サンプルXを分析したところ、意味アテンション機構は〇〇という概念を活性化したが、非可換処理モジュールが△△という順序パターンを検出し、最終的に断絶創発モジュールが□□という創発ラベルを生成した結果、ベースラインとは異なる判断に至った（あるいは、依然として誤ったが、その誤り方が変化した）。これは...を示唆する。」といった具体的な記述。* **5.3. 「意味構造生成(SSG)」の観点からの初期的な考察** * 今回の実験で得られた様々な中間表現（意味概念、非可換効果、断絶スコア、創発ラベル、鍵概念による縮減表現）は、お客様の目指す「意味構造」の断片と言えるか？ * それらはどの程度「ホワイトボックス」的で、人間の直感や解釈と整合性があるか？ * これらの要素をどのように組み合わせれば、より首尾一貫した「意味構造」へと発展させられるかの展望。* **5.4. 本研究の限界と今後の実験への示唆** (現状の3エポック、特定のデータセットといった限界。より多くのデータ、エポック、ハイパーパラメータチューニングの必要性など)

6. 結論 (Conclusion) * 本研究で提案・評価した多階層意味構造モジュール群の主要な発見（最も有望だった構成、XAI透明性への貢献の可能性など）を要約。* トランスフォーマーの能力拡張とXAI透明性の実現に向けた、本アプローチの初期的な有効性と、今後の研究開発における具体的な課題を明確に述べる。

7. 今後の課題と展望：人間との共進化を目指したSSGへ (Future Work and Vision: Towards Co-evolving SSG with Humans) * **7.1. 今回の知見に基づく各モジュールの設計改良と厳密なハイパーパラメータ最適化。** * **7.2. 「閾値による階層処理制御」の導入と評価:** 計算効率と適応性を高めるための具体的な設計案。* **7.3. 「正解データ」に基づく学習アプローチの具体化:** お客様の重要なアイデアである、人間が付与した「意味構造」や「倫理的判断」のデータをどのように収集し、SSGや将来的な「倫理階層モジュール」の学習にどう活用していくかの具体的な計画。* **7.4. 「SSG（意味構造生成）モジュール」の完成度向上と「人間との共進化」:** * 今回の実験はSSGの基礎要素の検証であったことを述べ、これらを統合・発展させて、真に「人間の分身」として機能し、AIの進化と人間の理解のタイムスケールを同期させるSSGをどう構築していくかの展望を熱く語る。* これが、お客様の究極の目標である「P2Pエッジでの個別最適化AGI」や、AIの暴走リスクを抑制し、AIと人間が建設的に共存・共進化できる未来にどう繋がるのか、そのビジョンを示す。* **7.5. 他タスク、他言語、マルチモーダルへの展開可能性。**

統合アプリケーション実験コード

セル0

セル1 (セットアップとカーネル再起動用)


```
# -----
# 関連ライブラリを一度アンインストールしてクリーンな状態にする
!pip uninstall -y transformers accelerate peft sentence-transformers datasets -q

# 指定バージョンでライブラリをインストール
!pip install transformers==4.36.2 -q
!pip install datasets -q
!pip install peft==0.7.1 -q
!pip install accelerate==0.25.0 -q # <--- accelerateのバージョンを0.25.0に固定 (または0.26.1)

# 変更を適用するためにカーネルを再起動
import os
os.kill(os.getpid(), 9)
```

セル 1

```
# キャッシュクリア用コマンドの例
!rm -rf ~/.cache/huggingface/datasets/imdb
print("IMDb dataset cache cleared.")
```

セル 2

```
# --- セル2: ライブラリのインポート ---

import torch
import torch.nn as nn
import torch.nn.functional as F # SemanticConceptModule で F.softmax を使用するため

from transformers import (
    AutoModelForSequenceClassification,
    AutoTokenizer,
    Trainer,
    TrainingArguments,
    AutoConfig,
    PreTrainedModel,
    AutoModel
)
from transformers.modeling_outputs import SequenceClassifierOutput # モデルの出力形式用
from torch.nn import CrossEntropyLoss # 損失関数用

from datasets import load_dataset, Dataset # ★ Dataset クラスをインポート

import numpy as np
from sklearn.metrics import accuracy_score # 評価指標計算用
import itertools # 実験構成生成用 (今回は手動定義なので厳密には不要ですが、将来的に使う可能性も)
```

```

# (もし他に必要な標準ライブラリがあればここに追加してください)
# import os # (セル1で使用済みなので、ここでは必須ではない)
# import json
# import time
# import datetime

# (もしグラフ描画などを行う場合は、matplotlibなどもここでインポート)
# import matplotlib.pyplot as plt
# import seaborn as sns
# import pandas as pd

print("Cell 2: Libraries imported successfully.")
print(f" PyTorch version: {torch.__version__}")
try:
    import transformers
    print(f" Transformers version: {transformers.__version__}")
    import datasets
    print(f" Datasets version: {datasets.__version__}")
    import accelerate
    print(f" Accelerate version: {accelerate.__version__}")
    import peft
    print(f" PEFT version: {peft.__version__}")
except ImportError:
    print("Warning: Could not print all Hugging Face library versions. Ensure they are installed.")

```

セル 3

```

# --- セル3: 基本設定・共通パラメータ ---

```

```

from transformers import AutoTokenizer, AutoConfig # 必要なものをインポート (セル2でインポート済
みなら重複を避けてもOK)

```

```

# --- 基本的な設定値 ---

```

```

BASE_MODEL_NAME = "bert-base-uncased" # ベースとなる事前学習済みモデルの名前
NUM_LABELS = 2 # 分類タスクのラベル数 (例: IMDbならポジティブ/ネガティブの2)
OUTPUT_DIR_BASE = "./experiment_results/" # 実験結果の出力先ベースディレクトリ

```

```

# --- 訓練に関するハイパーパラメータ ---

```

```

LEARNING_RATE = 2e-5 # 学習率
BATCH_SIZE = 1 # バッチサイズ (GPUメモリやデータに応じて調整)
# CPU実行の場合は小さめ(例: 1や2)にしないと非常に遅い可能性があります
NUM_EPOCHS = 1 # 訓練エポック数 (最初は1などでテストし、問題なければ増やす)

```

```

# --- トークナイザーとモデル設定のロード ---

```

```

try:
    print(f"Loading tokenizer for {BASE_MODEL_NAME}...")
    tokenizer = AutoTokenizer.from_pretrained(BASE_MODEL_NAME)

```

```

    print(f"Tokenizer for {BASE_MODEL_NAME} loaded successfully.")
except Exception as e:
    print(f"Error loading tokenizer for {BASE_MODEL_NAME}: {e}")
    raise # エラーが発生したら処理を停止

try:
    print(f"Loading Hugging Face model config for {BASE_MODEL_NAME}...")
    hf_model_config = AutoConfig.from_pretrained(
        BASE_MODEL_NAME,
        num_labels=NUM_LABELS
        # 必要に応じて他のAutoConfigのパラメータもここで設定可能
        # 例: finetuning_task='text-classification' など
    )
    print(f"Model config for {BASE_MODEL_NAME} loaded successfully.")
except Exception as e:
    print(f"Error loading model config for {BASE_MODEL_NAME}: {e}")
    raise

print("\n--- Cell 3: Basic/Common Settings Defined ---")
print(f"BASE_MODEL_NAME: {BASE_MODEL_NAME}")
print(f"NUM_LABELS: {NUM_LABELS}")
print(f"OUTPUT_DIR_BASE: {OUTPUT_DIR_BASE}")
print(f"LEARNING_RATE: {LEARNING_RATE}")
print(f"BATCH_SIZE: {BATCH_SIZE}")
print(f"NUM_EPOCHS: {NUM_EPOCHS}")
print(f"Tokenizer type: {type(tokenizer).__name__}")
print(f"HF Model Config type: {type(hf_model_config).__name__}")

```

セル 4

--- セル4: データセット準備 と compute_metrics 関数の定義 (本格運用版) ---

```

from datasets import load_dataset
from sklearn.metrics import accuracy_score
import numpy as np
import torch

# --- グローバル変数 (セル3で定義されているはずのもの) の確認 ---
if 'BASE_MODEL_NAME' not in globals():
    raise NameError("Global variable 'BASE_MODEL_NAME' is not defined. Please ensure Cell 3 has been executed.")
if 'tokenizer' not in globals():
    raise NameError("Global variable 'tokenizer' is not defined. Please ensure Cell 3 has been executed.")

print(f"Starting dataset preparation using BASE_MODEL_NAME: {BASE_MODEL_NAME} and tokenizer: {type(tokenizer).__name__}")

```

```

# 1. データセットのロード
try:
    dataset_dict = load_dataset("imdb") # DatasetDictオブジェクトとしてロード
    print("IMDb dataset loaded successfully.")
except Exception as e:
    print(f"Error loading IMDb dataset: {e}")
    raise

# 2. トークナイズ関数の定義
def tokenize_function(examples):
    return tokenizer(examples["text"], padding="max_length", truncation=True, max_length=256)

# 3. データセットのトークナイズ
try:
    print("Tokenizing datasets...")
    tokenized_train = dataset_dict["train"].map(tokenize_function, batched=True,
remove_columns=["text"])
    tokenized_test = dataset_dict["test"].map(tokenize_function, batched=True,
remove_columns=["text"])
    print("Tokenization complete for train and test splits.")
except Exception as e:
    print(f"Error during tokenization: {e}")
    raise

# 4. 訓練用・評価用データセットの準備とカラム名修正
try:
    # 'label' カラムを 'labels' にリネーム
    if 'label' in tokenized_train.column_names and 'labels' not in tokenized_train.column_names:
        print("Renaming 'label' to 'labels' in tokenized_train...")
        final_tokenized_train = tokenized_train.rename_column("label", "labels")
    else:
        final_tokenized_train = tokenized_train
    if 'labels' not in final_tokenized_train.column_names:
        print("Warning: 'labels' column expected but not found in final_tokenized_train.")

    if 'label' in tokenized_test.column_names and 'labels' not in tokenized_test.column_names:
        print("Renaming 'label' to 'labels' in tokenized_test...")
        final_tokenized_test = tokenized_test.rename_column("label", "labels")
    else:
        final_tokenized_test = tokenized_test
    if 'labels' not in final_tokenized_test.column_names:
        print("Warning: 'labels' column expected but not found in final_tokenized_test.")

# --- ★★★ 本格的な実験用のデータサンプル数をここで設定 ★★★ ---
NUM_TRAIN_SAMPLES_FULL = 10 # 例: 10,000件 (または len(final_tokenized_train) で全件)
NUM_EVAL_SAMPLES_FULL = 2 # 例: 2,500件 (または len(final_tokenized_test) で全件)

```

```

# -----

# グローバル変数として train_dataset と eval_dataset を定義
train_dataset = final_tokenized_train.shuffle(seed=42).select(
    range(min(NUM_TRAIN_SAMPLES_FULL, len(final_tokenized_train)))
)
eval_dataset = final_tokenized_test.shuffle(seed=42).select(
    range(min(NUM_EVAL_SAMPLES_FULL, len(final_tokenized_test)))
)

print(f"Train dataset created. Number of samples: {len(train_dataset)}")
print(f" Train dataset FINAL column names: {train_dataset.column_names}")
print(f"Evaluation dataset created. Number of samples: {len(eval_dataset)}")
print(f" Eval dataset FINAL column names: {eval_dataset.column_names}")

if len(eval_dataset) > 0: # 念のため表示
    print(f"First sample of EVAL_dataset (for column check): {eval_dataset[0]}")

except Exception as e:
    print(f"Error creating or renaming train/eval datasets: {e}")
    raise

# 5. 評価指標計算関数の定義 (デバッグプリントは有効のまま残しておいても良い)
def compute_metrics(eval_pred):
    logits, labels = eval_pred
    if labels is None or len(labels) == 0:
        print("--- Inside compute_metrics: Received no labels or labels is None. Returning empty
metrics. ---")
        return {}
    predictions = np.argmax(logits, axis=-1)
    accuracy = accuracy_score(labels, predictions)
    metrics_dict = {"eval_accuracy": accuracy}
    # --- デバッグ用プリント (本格運用時はコメントアウトしても良い) ---
    # print(f"--- Inside compute_metrics ---")
    # print(f" eval_pred logits type: {type(logits)}, labels type: {type(labels)}")
    # if hasattr(logits, 'shape'): print(f" eval_pred logits shape: {logits.shape}")
    # if hasattr(labels, 'shape'): print(f" eval_pred labels shape: {labels.shape}")
    # print(f" Predictions example (first 5): {predictions[:5]}")
    # print(f" Labels example (first 5): {labels[:5]}")
    # print(f" Calculated accuracy: {accuracy}")
    # print(f" Returning metrics_dict: {metrics_dict}")
    # --- ここまでデバッグ用プリント ---
    return metrics_dict

print("\nDataset preparation cell (Cell 4) complete.")
print(f"Defined global variables: 'train_dataset' (size: {len(train_dataset)}), 'eval_dataset' (size:
{len(eval_dataset)}), 'compute_metrics'")
if callable(globals().get('compute_metrics')):

```



```

    print("'compute_metrics' function is defined and callable.")
else:
    print("Warning: 'compute_metrics' function is NOT defined or not callable.")

```

セル 5

```

# --- セル5: 意味アテンション機構 (SemanticConceptModule) の定義 ---
import torch
import torch.nn as nn
import torch.nn.functional as F
# from transformers import PretrainedConfig # model_configの型ヒント用 (必要なら)

class SemanticConceptModule(nn.Module):
    def __init__(self, input_dim, output_dim,
                  num_concepts=32, concept_dim=128,
                  model_config=None,
                  module_specific_config=None
                  ):
        super().__init__()
        self.input_dim = input_dim
        self.output_dim = output_dim
        self.num_concepts = num_concepts
        self.concept_dim = concept_dim

        self.concept_prototypes = nn.Parameter(torch.randn(self.num_concepts, self.concept_dim))
        nn.init.xavier_uniform_(self.concept_prototypes)

        self.hidden_to_concept_proj = nn.Linear(self.input_dim, self.concept_dim)
        self.concepts_to_integrate_proj = nn.Linear(self.concept_dim, self.input_dim)
        self.activation = nn.ReLU()

        if self.input_dim != self.output_dim:
            self.final_projection = nn.Linear(self.input_dim, self.output_dim)

        print(f"SemanticConceptModule defined and initialized: input_dim={self.input_dim}, "
              output_dim={self.output_dim}, "
              f"num_concepts={self.num_concepts}, concept_dim={self.concept_dim}")

    def forward(self, hidden_states, attention_mask=None, **kwargs):
        projected_hidden = self.activation(self.hidden_to_concept_proj(hidden_states))
        attention_scores = torch.matmul(projected_hidden, self.concept_prototypes.t())

        if attention_mask is not None:
            expanded_attention_mask = attention_mask.unsqueeze(-1).float()
            attention_scores = attention_scores.masked_fill(expanded_attention_mask == 0, -1e9)

        attention_weights = F.softmax(attention_scores, dim=-1)

```

```

        contextual_concepts = torch.matmul(attention_weights, self.concept_prototypes)
        projected_contextual_concepts =
self.activation(self.concepts_to_integrate_proj(contextual_concepts))
        fused_hidden_states = hidden_states + projected_contextual_concepts

        if self.input_dim != self.output_dim:
            output = self.final_projection(fused_hidden_states)
        else:
            output = fused_hidden_states

        return {"last_hidden_state": output, "attention_weights": attention_weights}

def get_output_dim(self):
    return self.output_dim

print("Cell 5: SemanticConceptModule class defined.")

```

セル 6

```

# --- セル6: 直接的非可換処理モジュール (DirectNonCommutativeModule) の定義 ---
import torch
import torch.nn as nn
# from transformers import PretrainedConfig # model_configの型ヒント用 (必要なら)

class DirectNonCommutativeModule(nn.Module):
    def __init__(self, input_dim, output_dim,
                  dropout=0.2,
                  model_config=None,
                  module_specific_config=None
                  ):
        super().__init__()
        self.input_dim = input_dim
        self.output_dim = output_dim
        internal_dropout = module_specific_config.get("dropout", dropout)
        internal_processing_dim = module_specific_config.get("internal_processing_dim", input_dim * 2)

        self.non_commutative_processor = nn.Sequential(
            nn.Linear(input_dim * 2, internal_processing_dim),
            nn.LayerNorm(internal_processing_dim),
            nn.ReLU(),
            nn.Dropout(internal_dropout),
            nn.Linear(internal_processing_dim, input_dim)
        )
        self.lambda_param = nn.Parameter(torch.tensor(0.5))

        if self.input_dim != self.output_dim:
            self.final_integration_projection = nn.Linear(self.input_dim, self.output_dim)

```

```

        print(f"DirectNonCommutativeModule defined and initialized: input_dim={input_dim},
output_dim={output_dim}")

def forward(self, hidden_states, attention_mask=None, **kwargs):
    batch_size, seq_len, local_input_dim = hidden_states.shape
    device = hidden_states.device

    if seq_len < 2:
        # print("Warning: Seq_len < 2 in DirectNonCommutativeModule...") # 必要なら表示
        if self.input_dim == self.output_dim:
            return {"last_hidden_state": hidden_states, "non_commutative_effects_map": None}
        else:
            if not hasattr(self, 'final_integration_projection'):
                temp_proj = nn.Linear(self.input_dim, self.output_dim).to(device)
                return {"last_hidden_state": temp_proj(hidden_states), "non_commutative_effects_map":
None}
            return {"last_hidden_state": self.final_integration_projection(hidden_states),
"non_commutative_effects_map": None}

    non_commutative_effects_at_each_step = torch.zeros_like(hidden_states)
    for i in range(seq_len - 1):
        current_h = hidden_states[:, i, :]
        next_h = hidden_states[:, i + 1, :]
        forward_concat = torch.cat([current_h, next_h], dim=-1)
        backward_concat = torch.cat([next_h, current_h], dim=-1)
        forward_processed = self.non_commutative_processor(forward_concat)
        backward_processed = self.non_commutative_processor(backward_concat)
        non_comm_effect = self.lambda_param * (forward_processed - backward_processed)
        non_commutative_effects_at_each_step[:, i, :] += non_comm_effect

    fused_hidden_states = hidden_states + non_commutative_effects_at_each_step

    if self.input_dim != self.output_dim:
        output = self.final_integration_projection(fused_hidden_states)
    else:
        output = fused_hidden_states

    return {"last_hidden_state": output, "non_commutative_effects_map":
non_commutative_effects_at_each_step}

def get_output_dim(self):
    return self.output_dim

print("Cell 6: DirectNonCommutativeModule class defined.")

```

```

# --- セル7: 断絶創発モジュール (DiscontinuityEmergenceModule) の定義 ---
import torch
import torch.nn as nn
# from transformers import PretrainedConfig # model_configの型ヒント用 (必要なら)

class DiscontinuityEmergenceModule(nn.Module):
    def __init__(self, input_dim, output_dim,
                  semantic_dim_for_discontinuity=12,
                  emergent_label_dim=24,
                  dropout=0.2,
                  model_config=None,
                  module_specific_config=None
                  ):
        super().__init__()
        self.input_dim = input_dim
        self.output_dim = output_dim
        self.semantic_dim_for_discontinuity =
module_specific_config.get("semantic_dim_for_discontinuity", semantic_dim_for_discontinuity)
        self.emergent_label_dim = module_specific_config.get("emergent_label_dim",
emergent_label_dim)
        internal_dropout = module_specific_config.get("dropout", dropout)

        self.feature_to_semantic_for_discontinuity = nn.Sequential(
            nn.Linear(input_dim, input_dim // 2),
            nn.ReLU(),
            nn.Linear(input_dim // 2, self.semantic_dim_for_discontinuity),
            nn.Tanh()
        )
        self.discontinuity_detector = nn.Sequential(
            nn.Linear(self.semantic_dim_for_discontinuity * 2, 64),
            nn.ReLU(),
            nn.Dropout(internal_dropout),
            nn.Linear(64, 1),
            nn.Sigmoid()
        )
        self.emergent_label_generator = nn.Sequential(
            nn.Linear(self.semantic_dim_for_discontinuity * 2, input_dim // 4),
            nn.LayerNorm(input_dim // 4),
            nn.ReLU(),
            nn.Dropout(internal_dropout),
            nn.Linear(input_dim // 4, self.emergent_label_dim),
            nn.Tanh()
        )
        if self.input_dim + self.emergent_label_dim != self.output_dim :
            self.integration_projection = nn.Linear(self.input_dim + self.emergent_label_dim,
self.output_dim)
        elif self.input_dim != self.output_dim:

```

```

self.integration_projection = nn.Linear(self.input_dim, self.output_dim)

print(f"DiscontinuityEmergenceModule defined and initialized: input_dim={input_dim},
output_dim={output_dim}, "
      f"semantic_dim_disc={self.semantic_dim_for_discontinuity},
emergent_dim={self.emergent_label_dim}")

def forward(self, hidden_states, attention_mask=None, **kwargs): # hidden_states を
input_features から変更
    batch_size, seq_len, _ = hidden_states.shape
    device = hidden_states.device

    semantic_features_for_discontinuity = self.feature_to_semantic_for_discontinuity(hidden_states)

    discontinuity_scores = torch.zeros(batch_size, 0, device=device) # 初期化
    if seq_len > 1:
        sfd_t = semantic_features_for_discontinuity[:, :-1, :]
        sfd_t_plus_1 = semantic_features_for_discontinuity[:, 1:, :]
        combined_for_discontinuity = torch.cat([sfd_t, sfd_t_plus_1], dim=-1)
        discontinuity_scores = self.discontinuity_detector(combined_for_discontinuity)
        discontinuity_scores = discontinuity_scores.squeeze(-1)

    emergent_labels_sequence = torch.zeros(batch_size, 0, self.emergent_label_dim, device=device)
    if seq_len > 1:
        # combined_for_discontinuity を再利用
        emergent_labels_sequence = self.emergent_label_generator(combined_for_discontinuity)

    output_features = hidden_states
    if emergent_labels_sequence.numel() > 0 and emergent_labels_sequence.shape[1] > 0:
        pooled_emergent_label = emergent_labels_sequence.mean(dim=1)
        expanded_emergent_label = pooled_emergent_label.unsqueeze(1).expand(-1, seq_len, -1)
        combined_features = torch.cat([hidden_states, expanded_emergent_label], dim=-1)

    if hasattr(self, 'integration_projection'):
        output_features = self.integration_projection(combined_features)
    elif self.input_dim + self.emergent_label_dim == self.output_dim:
        output_features = combined_features
    # ... (他のフォールバックや次元調整ロジックは以前のコードを参照)
    elif self.input_dim != self.output_dim :
        if not hasattr(self, 'final_projection_fallback'): # __init__で定義しておくべき
            self.final_projection_fallback = nn.Linear(self.input_dim, self.output_dim).to(device)
        output_features = self.final_projection_fallback(hidden_states)

    return {
        "last_hidden_state": output_features,
        "discontinuity_scores": discontinuity_scores,
        "emergent_labels_sequence": emergent_labels_sequence
    }

```

```
def get_output_dim(self):
    return self.output_dim
```

```
print("Cell 7: DiscontinuityEmergenceModule class defined.")
```

セル 8

--- セル7.1: ホログラフ縮減モジュール (InterpretableHolographicReducer) の定義 ---

```
import torch
import torch.nn as nn
import torch.nn.functional as F
# from transformers import PretrainedConfig # model_configの型ヒント用 (必要なら)

class InterpretableHolographicReducer(nn.Module):
    def __init__(self, input_feature_dim, # 前の階層からの特徴ベクトルの次元
                  num_key_concepts=16,    # 鍵となる意味概念の数 (アトム数に相当)
                  reduced_dim_per_concept=64, # 各鍵概念の縮減後ベクトル次元
                  model_config=None,
                  module_specific_config=None):
        super().__init__()
        self.input_feature_dim = input_feature_dim
        self.num_key_concepts = module_specific_config.get("num_key_concepts", num_key_concepts)
        self.reduced_dim_per_concept = module_specific_config.get("reduced_dim_per_concept",
                                                                    reduced_dim_per_concept)

        self.key_concept_prototypes = nn.Parameter(
            torch.randn(self.num_key_concepts, self.input_feature_dim)
        )
        nn.init.xavier_uniform_(self.key_concept_prototypes)

        self.reduction_projector = nn.Linear(self.input_feature_dim, self.reduced_dim_per_concept)

        self.final_output_dim = self.num_key_concepts * self.reduced_dim_per_concept

        print(f"InterpretableHolographicReducer initialized: input_dim={self.input_feature_dim}, "
              f"num_key_concepts={self.num_key_concepts}, "
              f"reduced_dim_per_concept={self.reduced_dim_per_concept}, "
              f"total_output_dim={self.final_output_dim}")

    def forward(self, hidden_states_sequence, attention_mask=None, **kwargs):
        # hidden_states_sequence: (batch_size, seq_len, input_feature_dim)

        reduced_concept_vectors_list = []
        # 各鍵概念プロトタイプで入力シーケンス全体から情報を集約
        for i in range(self.num_key_concepts):
            # (batch_size, seq_len, input_feature_dim) と (input_feature_dim)
```

```

concept_specific_attention_scores = torch.matmul(
    hidden_states_sequence, self.key_concept_prototypes[i].unsqueeze(-1)
).squeeze(-1) # (batch_size, seq_len)

if attention_mask is not None: # パディング部分をマスク
    concept_specific_attention_scores =
concept_specific_attention_scores.masked_fill(attention_mask == 0, -1e9)

concept_specific_attention_weights = F.softmax(concept_specific_attention_scores, dim=1) #
(batch_size, seq_len)

# このウェイトで hidden_states を重み付き和 (概念iに関する文脈ベクトル)
context_vector_for_concept_i = torch.sum(
    hidden_states_sequence * concept_specific_attention_weights.unsqueeze(-1), dim=1
) # (batch_size, input_feature_dim)

reduced_vector_for_concept_i = self.reduction_projector(context_vector_for_concept_i) #
(batch_size, reduced_dim_per_concept)
reduced_concept_vectors_list.append(reduced_vector_for_concept_i)

final_reduced_representation = torch.cat(reduced_concept_vectors_list, dim=1) # (batch_size,
num_key_concepts * reduced_dim_per_concept)

# このモジュールは縮減された固定長ベクトルを返す
# これが後続の分類器の直接の入力となる
return {
    "last_hidden_state": final_reduced_representation,
    # "concept_attention_weights": concept_global_attention_weights # (オプション) 全体としてど
の概念が活性化したか
}

def get_output_dim(self):
    return self.final_output_dim

print("Cell 7.1: InterpretableHolographicReducer class defined.")

```

セル 9

--- セルA: 統合モデルとDebugTrainerの定義 (ホログラフ縮減モジュール対応版) ---

```

from transformers import PreTrainedModel, AutoModel, AutoConfig, Trainer
from transformers.modeling_outputs import SequenceClassifierOutput
from torch.nn import CrossEntropyLoss
import torch
import torch.nn as nn # nn もここでインポートしておく と 確実
import torch.nn.functional as F # F もここでインポートしておく と 確実

```



```
# 以下のカスタムモジュールクラスが、これより前のセル (セル5, 6, 7, 7.1) で
# 既に定義されていることを前提とします。
# - SemanticConceptModule
# - DirectNonCommutativeModule
# - DiscontinuityEmergenceModule
# - InterpretableHolographicReducer

print("--- Cell A: Defining EnhancedTransformerForSequenceClassification (Holographic Reduction
Ready) and DebugTrainer ---")
```

```
# --- 統合モデル: EnhancedTransformerForSequenceClassification ---
class EnhancedTransformerForSequenceClassification(PreTrainedModel):
    def __init__(self, hf_config, base_model_name, num_labels, tech_flags=None,
module_configs=None):
        super().__init__(hf_config)
        self.num_labels = num_labels
        self.config = hf_config # Hugging Faceのconfigオブジェクトを保存

        # ベースモデルのロードと設定
        self.base_model_prefix = self.config.model_type
        core_model = AutoModel.from_pretrained(base_model_name, config=self.config)
        setattr(self, self.base_model_prefix, core_model)

        current_dim = self.config.hidden_size # ベースモデルの出力次元
        self.tech_modules = nn.ModuleDict()
        self.tech_flags = tech_flags if tech_flags is not None else {}
        self.module_configs = module_configs if module_configs is not None else {}

        # 適用するモジュールの順序を取得 (指定がなければ空リスト)
        self.module_order = self.tech_flags.get("module_order", [])
        self.active_module_names_in_order = [] # 実際に初期化されたモジュールの順序を保持

        # module_order に従ってモジュールを初期化
        for module_key in self.module_order:
            if module_key == "semantic_attention_module" and
self.tech_flags.get("use_semantic_attention_module", False):
                scm_config = self.module_configs.get(module_key, {})
                module_output_dim = scm_config.get("output_dim", current_dim)
                self.tech_modules[module_key] = SemanticConceptModule(
                    input_dim=current_dim, output_dim=module_output_dim,
                    num_concepts=scm_config.get("num_concepts", 32),
                    concept_dim=scm_config.get("concept_dim", 128), model_config=self.config,
                    module_specific_config=scm_config # モジュール固有設定を渡す
                )
                current_dim = module_output_dim
                self.active_module_names_in_order.append(module_key)
            print(f"{module_key} enabled and initialized. Output dim: {current_dim}")
```

```

elif module_key == "direct_non_commutative_module" and
self.tech_flags.get("use_direct_non_commutative_module", False):
    dnc_config = self.module_configs.get(module_key, {})
    module_output_dim = dnc_config.get("output_dim", current_dim)
    self.tech_modules[module_key] = DirectNonCommutativeModule(
        input_dim=current_dim, output_dim=module_output_dim,
        dropout=dnc_config.get("dropout", 0.2), model_config=self.config,
        module_specific_config=dnc_config
    )
    current_dim = module_output_dim
    self.active_module_names_in_order.append(module_key)
    print(f"{module_key} enabled and initialized. Output dim: {current_dim}")

elif module_key == "discontinuity_emergence_module" and
self.tech_flags.get("use_discontinuity_emergence_module", False):
    dem_config = self.module_configs.get(module_key, {})
    module_output_dim = dem_config.get("output_dim", current_dim)
    self.tech_modules[module_key] = DiscontinuityEmergenceModule(
        input_dim=current_dim, output_dim=module_output_dim,
        semantic_dim_for_discontinuity=dem_config.get("semantic_dim_for_discontinuity", 12),
        emergent_label_dim=dem_config.get("emergent_label_dim", 24),
        dropout=dem_config.get("dropout", 0.2), model_config=self.config,
        module_specific_config=dem_config
    )
    current_dim = module_output_dim
    self.active_module_names_in_order.append(module_key)
    print(f"{module_key} enabled and initialized. Output dim: {current_dim}")

elif module_key == "holographic_reduction_module" and
self.tech_flags.get("use_holographic_reduction_module", False):
    hr_config = self.module_configs.get(module_key, {})
    # InterpretableHolographicReducer は自身のget_output_dim()で出力次元を返す
    # input_feature_dim は current_dim (前の層の出力次元)
    self.tech_modules[module_key] = InterpretableHolographicReducer(
        input_feature_dim=current_dim,
        # num_key_concepts と reduced_dim_per_concept は module_specific_config から渡す
        module_specific_config=hr_config,
        model_config=self.config
    )
    current_dim = self.tech_modules[module_key].get_output_dim() # 縮減後の次元に更新
    self.active_module_names_in_order.append(module_key)
    print(f"{module_key} enabled and initialized. Output dim (classifier input): {current_dim}")

# 他のモジュールも同様に追加可能

self.dropout = nn.Dropout(self.config.hidden_dropout_prob if hasattr(self.config,
'hidden_dropout_prob') else 0.1)
self.classifier = nn.Linear(current_dim, self.num_labels) # 最終的なcurrent_dimで分類器を初期化

```

```

print(f"EnhancedTransformerForSequenceClassification initialized. Final classifier input dim:
{current_dim}.")
if self.active_module_names_in_order:
    print(f" Applied modules (in order): {self.active_module_names_in_order}")
else:
    print(" No additional tech modules applied (Baseline configuration).")

def forward(
    self, input_ids=None, attention_mask=None, token_type_ids=None, labels=None,
    return_dict=None, **kwargs ): # target_semantic_labels は kwargs に含まれる想定もし必要なら

    # print(f"--- EnhancedTransformer.forward CALLED (is_training: {self.training}) ---")
    # if labels is not None: print(f" Forward RECEIVED labels, shape: {labels.shape}")
    # else: print(f" Forward did NOT receive labels this call.")

    base_model_kwargs = { k: v for k, v in kwargs.items() if k in ["position_ids", "head_mask",
"inputs_embeds", "output_attentions", "output_hidden_states"]}
    transformer_outputs = self.base_model(input_ids=input_ids, attention_mask=attention_mask,
token_type_ids=token_type_ids, return_dict=True, **base_model_kwargs)
    current_features = transformer_outputs.last_hidden_state # (batch, seq_len, hidden_dim)

    all_module_specific_outputs = {}

    # module_order に従ってモジュールを順次適用
    for module_name_key in self.active_module_names_in_order: # __init__で実際に初期化された順
    序リストを使用
        if module_name_key in self.tech_modules:
            # print(f" Applying module: {module_name_key}")
            # HolographicReductionModule はシーケンスを受け取り、固定長ベクトルを返す想定
            # それ以外のモジュールはシーケンスを受け取り、シーケンスを返す想定
            # この分岐は HolographicReductionModule が常に module_order の最後に来るという前提
            に基づく
            # もし途中に来る場合は、current_features の形状が変わるため、より複雑な制御が必要

            module_output_dict = self.tech_modules[module_name_key](current_features,
attention_mask=attention_mask)
            current_features = module_output_dict["last_hidden_state"]

            for output_key, output_value in module_output_dict.items():
                if output_key != "last_hidden_state":
                    all_module_specific_outputs[f"{module_name_key}_{output_key}"] = output_value

            # プーリング処理
            # HolographicReductionModule が適用された場合、current_features は既に (batch,
reduced_dim) のはず
            # そうでない場合は、通常のCLSトークンなどによるプーリングを行う
            is_holographic_active_and_last = self.active_module_names_in_order and \

```

```

        self.active_module_names_in_order[-1] ==
        "holographic_reduction_module" and \
        "holographic_reduction_module" in self.tech_modules

    if is_holographic_active_and_last:
        pooled_output = current_features # ホログラフ縮減モジュールの出力 (既にプーリング済み)
        # print(f" Using Holographic Reduction output as pooled_output. Shape:
        {pooled_output.shape}")
    else:
        if current_features.ndim == 3: # (batch, seq, dim)
            pooled_output = current_features[:, 0] # [CLS] トークンを使用 (または他のプーリング)
            # print(f" Using CLS token pooling. Shape: {pooled_output.shape}")
        elif current_features.ndim == 2: # 既にプーリング済みの場合 (例: 前のモジュールがプーリング
        した場合)
            pooled_output = current_features
            # print(f" Input 'current_features' is already pooled. Shape: {pooled_output.shape}")
        else:
            raise ValueError(f"Unsupported shape for current_features before classifier:
            {current_features.shape}")

    pooled_output = self.dropout(pooled_output)
    logits = self.classifier(pooled_output)

    loss = None
    if labels is not None:
        loss_fct = CrossEntropyLoss()
        loss = loss_fct(logits.view(-1, self.num_labels), labels.view(-1))
        # if loss is not None: print(f" Forward CALCULATED loss: {loss.item()}")

    # print(f"--- EnhancedTransformer.forward RETURNING ... ---")

    final_outputs_tuple = (logits,)
    # ベースモデルの主要な出力と、モジュールからの補助的な出力を選択的に含める
    # (今回はシンプルにベースモデルの主要なものと、モジュールからの全追加出力を返す)
    if hasattr(transformer_outputs, 'hidden_states') and transformer_outputs.hidden_states is not
    None:
        final_outputs_tuple += (transformer_outputs.hidden_states,)
    if hasattr(transformer_outputs, 'attentions') and transformer_outputs.attentions is not None:
        final_outputs_tuple += (transformer_outputs.attentions,)
    if all_module_specific_outputs:
        final_outputs_tuple += (all_module_specific_outputs,)

    if loss is not None:
        return (loss,) + final_outputs_tuple
    else:
        return (None,) + final_outputs_tuple

# --- DebugTrainer クラスの定義 (変更なし) ---

```

```

class DebugTrainer(Trainer):
    def prediction_step(
        self, model: nn.Module, inputs: dict[str, torch.Tensor | nn.utils.rnn.PackedSequence],
        prediction_loss_only: bool, ignore_keys: list[str] | None = None,
    ) -> tuple[torch.Tensor | None, torch.Tensor | None, torch.Tensor | None]:
        # print(f"--- DebugTrainer.prediction_step CALLED ---")
        model.eval()
        with torch.no_grad(): model_outputs = model(**inputs)
        loss = model_outputs[0] if model_outputs[0] is not None else None
        logits = model_outputs[1] if len(model_outputs) > 1 and model_outputs[1] is not None else None
        labels_out = inputs.get("labels")
        if prediction_loss_only: return (loss, None, None)
        return loss, logits, labels_out

print("Cell A: EnhancedTransformerForSequenceClassification (Holographic Reduction capable) and
DebugTrainer defined.")

```

セル 1 0

--- セルB: 実験構成、グローバル変数パック、共通実行関数 (完全統合版) ---

```

print("--- Cell B (Fully Integrated): Defining All Experiment Setups and Utilities ---")

```

1. 前提となるグローバル変数の確認

```

required_globals_cell_B_full = [
    'hf_model_config', 'BASE_MODEL_NAME', 'NUM_LABELS', 'OUTPUT_DIR_BASE',
    'LEARNING_RATE', 'BATCH_SIZE', 'NUM_EPOCHS',
    'EnhancedTransformerForSequenceClassification',
    'train_dataset', 'eval_dataset', 'tokenizer', 'compute_metrics',
    'DebugTrainer', 'Dataset',
    'SemanticConceptModule', 'DirectNonCommutativeModule',
    'DiscontinuityEmergenceModule', 'InterpretableHolographicReducer'
]
missing_globals_check_B_full = False
for var_name in required_globals_cell_B_full:
    if var_name not in globals():
        print(f"ERROR in Cell B (Fully Integrated): Global variable or class '{var_name}' is not defined!")
        missing_globals_check_B_full = True
if missing_globals_check_B_full:
    raise NameError("One or more required global variables/classes for Cell B (Fully Integrated) are
not defined. Please check prerequisite cells (1 through A).")
else:
    print("All prerequisite global variables and classes for Cell B (Fully Integrated) seem to be
defined.")

```

2. 実験構成の定義 (お客様の全パターン + ホログラフ縮減テスト)

```

experiment_configurations = [

```

```

# --- ベースライン ---
{"name": "E0_Baseline", "tech_flags": {"module_order": [], "module_configs": {}, "#日本語": "ベースライン"},
# --- 単一モジュール ---
{"name": "E1_SemanticAttention",
"tech_flags": {"module_order": ["semantic_attention_module"], "use_semantic_attention_module": True},
"module_configs": {"semantic_attention_module": {"num_concepts": 32, "concept_dim": 128, "output_dim": hf_model_config.hidden_size}}, "#日本語": "意味アテンション"},
{"name": "E2_DirectNonCommutative",
"tech_flags": {"module_order": ["direct_non_commutative_module"],
"use_direct_non_commutative_module": True},
"module_configs": {"direct_non_commutative_module": {"output_dim": hf_model_config.hidden_size, "dropout": 0.2, "internal_processing_dim": hf_model_config.hidden_size}}, "#日本語": "非可換"},
{"name": "E3_DiscontinuityEmergence",
"tech_flags": {"module_order": ["discontinuity_emergence_module"],
"use_discontinuity_emergence_module": True},
"module_configs": {"discontinuity_emergence_module": {"output_dim": hf_model_config.hidden_size, "semantic_dim_for_discontinuity": 12, "emergent_label_dim": 24, "dropout": 0.2}}, "#日本語": "断絶創発"},

# --- 2モジュールの組み合わせ (順序考慮) ---
{"name": "E4_SemAtt_then_NonComm",
"tech_flags": {"module_order": ["semantic_attention_module", "direct_non_commutative_module"],
"use_semantic_attention_module": True, "use_direct_non_commutative_module": True},
"module_configs": {"semantic_attention_module": {"num_concepts": 32, "concept_dim": 128, "output_dim": hf_model_config.hidden_size},
"direct_non_commutative_module": {"output_dim": hf_model_config.hidden_size, "dropout": 0.2, "internal_processing_dim": hf_model_config.hidden_size}}, "#日本語": "意味アテンション → 非可換"},
{"name": "E5_NonComm_then_SemAtt",
"tech_flags": {"module_order": ["direct_non_commutative_module", "semantic_attention_module"],
"use_direct_non_commutative_module": True, "use_semantic_attention_module": True},
"module_configs": {"direct_non_commutative_module": {"output_dim": hf_model_config.hidden_size, "dropout": 0.2, "internal_processing_dim": hf_model_config.hidden_size},
"semantic_attention_module": {"num_concepts": 32, "concept_dim": 128, "output_dim": hf_model_config.hidden_size}}, "#日本語": "非可換 → 意味アテンション"},
{"name": "E6_SemAtt_then_DiscEm",
"tech_flags": {"module_order": ["semantic_attention_module", "discontinuity_emergence_module"],
"use_semantic_attention_module": True, "use_discontinuity_emergence_module": True},
"module_configs": {"semantic_attention_module": {"num_concepts": 32, "concept_dim": 128, "output_dim": hf_model_config.hidden_size},
"discontinuity_emergence_module": {"output_dim": hf_model_config.hidden_size, "semantic_dim_for_discontinuity": 12, "emergent_label_dim": 24, "dropout": 0.2}}, "#日本語": "意味アテンション → 断絶創発"},
{"name": "E7_DiscEm_then_SemAtt",

```

```

    "tech_flags": {"module_order": ["discontinuity_emergence_module", "semantic_attention_module"],
    "use_discontinuity_emergence_module": True, "use_semantic_attention_module": True},
    "module_configs": {"discontinuity_emergence_module": {"output_dim":
hf_model_config.hidden_size, "semantic_dim_for_discontinuity": 12, "emergent_label_dim": 24,
"dropout": 0.2},
        "semantic_attention_module": {"num_concepts": 32, "concept_dim": 128,
"output_dim": hf_model_config.hidden_size}}, "#日本語": "断絶創発 → 意味アテンション"),
    {"name": "E8_NonComm_then_DiscEm",
    "tech_flags": {"module_order": ["direct_non_commutative_module",
"discontinuity_emergence_module"], "use_direct_non_commutative_module": True,
"use_discontinuity_emergence_module": True},
    "module_configs": {"direct_non_commutative_module": {"output_dim":
hf_model_config.hidden_size, "dropout": 0.2, "internal_processing_dim":
hf_model_config.hidden_size},
        "discontinuity_emergence_module": {"output_dim": hf_model_config.hidden_size,
"semantic_dim_for_discontinuity": 12, "emergent_label_dim": 24, "dropout": 0.2}}, "#日本語": "非可換 →
断絶創発"),
    {"name": "E9_DiscEm_then_NonComm",
    "tech_flags": {"module_order": ["discontinuity_emergence_module",
"direct_non_commutative_module"], "use_discontinuity_emergence_module": True,
"use_direct_non_commutative_module": True},
    "module_configs": {"discontinuity_emergence_module": {"output_dim":
hf_model_config.hidden_size, "semantic_dim_for_discontinuity": 12, "emergent_label_dim": 24,
"dropout": 0.2},
        "direct_non_commutative_module": {"output_dim": hf_model_config.hidden_size,
"dropout": 0.2, "internal_processing_dim": hf_model_config.hidden_size}}, "#日本語": "断絶創発 → 非可
換"),

    # --- 3モジュールの組み合わせ ---
    {"name": "E10_SemAtt_NonComm_DiscEm",
    "tech_flags": {"module_order": ["semantic_attention_module", "direct_non_commutative_module",
"discontinuity_emergence_module"], "use_semantic_attention_module": True,
"use_direct_non_commutative_module": True, "use_discontinuity_emergence_module": True},
    "module_configs": {
        "semantic_attention_module": {"num_concepts": 32, "concept_dim": 128, "output_dim":
hf_model_config.hidden_size},
        "direct_non_commutative_module": {"output_dim": hf_model_config.hidden_size, "dropout": 0.2,
"internal_processing_dim": hf_model_config.hidden_size},
        "discontinuity_emergence_module": {"output_dim": hf_model_config.hidden_size,
"semantic_dim_for_discontinuity": 12, "emergent_label_dim": 24, "dropout": 0.2}}, "#日本語": "意味アテ
ンション → 非可換 → 断絶創発"),
    {"name": "E11_NonComm_SemAtt_DiscEm",
    "tech_flags": {"module_order": ["direct_non_commutative_module", "semantic_attention_module",
"discontinuity_emergence_module"], "use_direct_non_commutative_module": True,
"use_semantic_attention_module": True, "use_discontinuity_emergence_module": True},
    "module_configs": {"direct_non_commutative_module": {"output_dim":
hf_model_config.hidden_size, "dropout": 0.2, "internal_processing_dim":
hf_model_config.hidden_size},
        "semantic_attention_module": {"num_concepts": 32, "concept_dim": 128, "output_dim":
hf_model_config.hidden_size},
        "discontinuity_emergence_module": {"output_dim": hf_model_config.hidden_size,
"semantic_dim_for_discontinuity": 12, "emergent_label_dim": 24, "dropout": 0.2}}, "#日本語": "断絶創発 → 非可換 → 断絶創発"),

```

```

        "semantic_attention_module": {"num_concepts": 32, "concept_dim": 128,
"output_dim": hf_model_config.hidden_size},
        "discontinuity_emergence_module": {"output_dim": hf_model_config.hidden_size,
"semantic_dim_for_discontinuity": 12, "emergent_label_dim": 24, "dropout": 0.2}}, "#日本語": "非可換 →
意味アテンション → 断絶創発"},
        {"name": "E12_DiscEm_SemAtt_NonComm",
        "tech_flags": {"module_order": ["discontinuity_emergence_module", "semantic_attention_module",
"direct_non_commutative_module"], "use_discontinuity_emergence_module": True,
"use_semantic_attention_module": True, "use_direct_non_commutative_module": True},
        "module_configs": {"discontinuity_emergence_module": {"output_dim":
hf_model_config.hidden_size, "semantic_dim_for_discontinuity": 12, "emergent_label_dim": 24,
"dropout": 0.2},
        "semantic_attention_module": {"num_concepts": 32, "concept_dim": 128,
"output_dim": hf_model_config.hidden_size},
        "direct_non_commutative_module": {"output_dim": hf_model_config.hidden_size,
"dropout": 0.2, "internal_processing_dim": hf_model_config.hidden_size}}, "#日本語": "断絶創発 → 意味
アテンション → 非可換"},
        {"name": "E13_SemAtt_DiscEm_NonComm",
        "tech_flags": {"module_order": ["semantic_attention_module", "discontinuity_emergence_module",
"direct_non_commutative_module"], "use_semantic_attention_module": True,
"use_discontinuity_emergence_module": True, "use_direct_non_commutative_module": True},
        "module_configs": {"semantic_attention_module": {"num_concepts": 32, "concept_dim": 128,
"output_dim": hf_model_config.hidden_size},
        "discontinuity_emergence_module": {"output_dim": hf_model_config.hidden_size,
"semantic_dim_for_discontinuity": 12, "emergent_label_dim": 24, "dropout": 0.2},
        "direct_non_commutative_module": {"output_dim": hf_model_config.hidden_size,
"dropout": 0.2, "internal_processing_dim": hf_model_config.hidden_size}}, "#日本語": "意味アテンション
→ 断絶創発 → 非可換"},
        {"name": "E14_NonComm_DiscEm_SemAtt",
        "tech_flags": {"module_order": ["direct_non_commutative_module",
"discontinuity_emergence_module", "semantic_attention_module"],
"use_direct_non_commutative_module": True, "use_discontinuity_emergence_module": True,
"use_semantic_attention_module": True},
        "module_configs": {"direct_non_commutative_module": {"output_dim":
hf_model_config.hidden_size, "dropout": 0.2, "internal_processing_dim":
hf_model_config.hidden_size},
        "discontinuity_emergence_module": {"output_dim": hf_model_config.hidden_size,
"semantic_dim_for_discontinuity": 12, "emergent_label_dim": 24, "dropout": 0.2},
        "semantic_attention_module": {"num_concepts": 32, "concept_dim": 128,
"output_dim": hf_model_config.hidden_size}}, "#日本語": "非可換 → 断絶創発 → 意味アテンション"},
        {"name": "E15_DiscEm_NonComm_SemAtt",
        "tech_flags": {"module_order": ["discontinuity_emergence_module",
"direct_non_commutative_module", "semantic_attention_module"],
"use_discontinuity_emergence_module": True, "use_direct_non_commutative_module": True,
"use_semantic_attention_module": True},
        "module_configs": {"discontinuity_emergence_module": {"output_dim":
hf_model_config.hidden_size, "semantic_dim_for_discontinuity": 12, "emergent_label_dim": 24,
"dropout": 0.2},

```



```

        "direct_non_commutative_module": {"output_dim": hf_model_config.hidden_size,
"dropout": 0.2, "internal_processing_dim": hf_model_config.hidden_size},
        "semantic_attention_module": {"num_concepts": 32, "concept_dim": 128,
"output_dim": hf_model_config.hidden_size}}, {"#日本語": "断絶創発 → 非可換 → 意味アテンション"}],

# --- ホログラフ縮減モジュールを含むテスト構成 ---
{"name": "E16_Baseline_then_HolographicReduction",
 "tech_flags": {"module_order": ["holographic_reduction_module"]},
"use_holographic_reduction_module": True},
    "module_configs": {"holographic_reduction_module": {"num_key_concepts": 16,
"reduced_dim_per_concept": 32, "reduction_type": "attention_pooling"}}, {"#日本語": "ベースライン → ホ
ログラフ縮減(AttPool)"},
    {"name": "E17_E6_then_HolographicReduction",
 "tech_flags": {"module_order": ["semantic_attention_module", "direct_non_commutative_module",
"discontinuity_emergence_module", "holographic_reduction_module"]},
        "use_semantic_attention_module": True, "use_direct_non_commutative_module": True,
"use_discontinuity_emergence_module": True, "use_holographic_reduction_module": True},
    "module_configs": {
        "semantic_attention_module": {"num_concepts": 32, "concept_dim": 128, "output_dim":
hf_model_config.hidden_size},
        "direct_non_commutative_module": {"output_dim": hf_model_config.hidden_size, "dropout": 0.2,
"internal_processing_dim": hf_model_config.hidden_size},
        "discontinuity_emergence_module": {"output_dim": hf_model_config.hidden_size,
"semantic_dim_for_discontinuity": 12, "emergent_label_dim": 24, "dropout": 0.2},
        "holographic_reduction_module": {"num_key_concepts": 16, "reduced_dim_per_concept": 32,
"reduction_type": "attention_pooling"}}, {"#日本語": "意味Att→非可換→断絶創発→ホログラフ(AttPool)"},
    ]
print(f"Total experiment configurations defined in Cell B (Fully Integrated):
{len(experiment_configurations)}")

```

3. グローバル変数のパッケージ化

```

global_variables_for_experiments = {
    'hf_model_config': hf_model_config, 'BASE_MODEL_NAME': BASE_MODEL_NAME,
'NUM_LABELS': NUM_LABELS,
    'OUTPUT_DIR_BASE': OUTPUT_DIR_BASE, 'LEARNING_RATE': LEARNING_RATE, 'BATCH_SIZE':
BATCH_SIZE,
    'NUM_EPOCHS': NUM_EPOCHS,
    'EnhancedTransformerForSequenceClassification':
EnhancedTransformerForSequenceClassification,
    'train_dataset': train_dataset, 'eval_dataset': eval_dataset, 'tokenizer': tokenizer,
'compute_metrics': compute_metrics, 'DebugTrainer': DebugTrainer, 'Dataset': Dataset,
'SemanticConceptModule': SemanticConceptModule,
'DirectNonCommutativeModule': DirectNonCommutativeModule,
'DiscontinuityEmergenceModule': DiscontinuityEmergenceModule,
'InterpretableHolographicReducer': InterpretableHolographicReducer
}
print("Global variables for experiments packaged.")

```

```

# 4. 共通実験実行関数 (run_ablation_experiment)
def run_ablation_experiment(exp_config, global_vars_dict):
    config_name = exp_config["name"]
    tech_flags = exp_config["tech_flags"]
    module_cfgs = exp_config["module_configs"]

    print(f"\n--- Running Experiment: {config_name} ---")
    print(f"  Technology Flags: {tech_flags}")
    print(f"  Module Configs: {module_cfgs}")

    hf_cfg = global_vars_dict['hf_model_config']
    base_model_name_local = global_vars_dict['BASE_MODEL_NAME']
    n_labels = global_vars_dict['NUM_LABELS']
    output_dir_base_local = global_vars_dict['OUTPUT_DIR_BASE']
    lr = global_vars_dict['LEARNING_RATE']
    bs = global_vars_dict['BATCH_SIZE']
    n_epochs = global_vars_dict['NUM_EPOCHS']

    enh_transformer_class_local = global_vars_dict['EnhancedTransformerForSequenceClassification']
    tr_dataset = global_vars_dict['train_dataset']
    ev_dataset = global_vars_dict['eval_dataset']
    tok = global_vars_dict['tokenizer']
    comp_metrics = global_vars_dict['compute_metrics']
    dbg_trainer_class_local = global_vars_dict['DebugTrainer']
    dataset_class_local = global_vars_dict['Dataset']

    if ev_dataset is not None and isinstance(ev_dataset, dataset_class_local):
        if 'labels' not in ev_dataset.column_names:
            print(f"  WARNING (inside run_ablation_experiment for {config_name}): 'labels' column not
found in eval_dataset!")
        else:
            print(f"  Warning (inside run_ablation_experiment for {config_name}): eval_dataset not found or
not a Dataset object.")

    model = enh_transformer_class_local(
        hf_config=hf_cfg, base_model_name=base_model_name_local, num_labels=n_labels,
        tech_flags=tech_flags, module_configs=module_cfgs
    )
    current_output_dir = f"{output_dir_base_local}{config_name}"

    training_args = TrainingArguments(
        output_dir=current_output_dir, learning_rate=lr, per_device_train_batch_size=bs,
        per_device_eval_batch_size=bs, num_train_epochs=n_epochs, weight_decay=0.01,
        evaluation_strategy="epoch", logging_strategy="epoch", save_strategy="epoch",
        load_best_model_at_end=True, metric_for_best_model="accuracy", report_to="none",
        do_eval=True, remove_unused_columns=False,
        # save_total_limit=1, # ディスク容量を節約する場合
    )

```

```

trainer = dbg_trainer_class_local(
    model=model, args=training_args, train_dataset=tr_dataset, eval_dataset=ev_dataset,
    tokenizer=tok, compute_metrics=comp_metrics,
)
print(f"Starting training for {config_name}...")
eval_results_dict = {}
try:
    trainer.train()
    print(f"Training finished for {config_name}.")
    print(f"Starting final evaluation for {config_name} (with best model)...")
    eval_results = trainer.evaluate()
    print(f"Evaluation results for {config_name}: {eval_results}")
    eval_results_dict = eval_results
except Exception as e:
    print(f"!!! An error occurred during training or evaluation for {config_name}: {e} !!!")
    import traceback
    traceback.print_exc()
    eval_results_dict = {"error": str(e), "eval_accuracy": float('nan'), "eval_loss": float('nan')}
return config_name, eval_results_dict

```

5. 全ての実験結果を格納する辞書の初期化

```
all_experiment_results_summary = {}
```

```

print("\nCell B (Fully Integrated): All setups for running experiments are complete.")
print(f"To run experiments, execute the individual experiment cells (e.g., Cell_E0, Cell_E1, etc.) one by one.")

```

実験コード

E0

```
# E0--- ベースライン ---
```

```
print("--- Executing Experiment E2_DirectNonCommutative (Baseline + Direct Non-Commutative) ---")
```

```
exp_to_run_name_e2 = "E2_DirectNonCommutative"
```

```

config_to_run_e2 = next((config for config in experiment_configurations if config["name"] ==
exp_to_run_name_e2), None)

```

```
if config_to_run_e2:
```

```
    if 'all_experiment_results_summary' not in globals(): all_experiment_results_summary = {}
```

```
    if 'global_variables_for_experiments' not in globals(): raise
```

```
NameError("global_variables_for_experiments is not defined.")
```

```
    name, result = run_ablation_experiment(config_to_run_e2, global_variables_for_experiments)
```

```
    all_experiment_results_summary[name] = result
```

```
    print(f"\n--- Results for {name} ---")
```

```

    print(result)
else:
    print(f"Configuration for '{exp_to_run_name_e2}' not found.")

print(f"--- Finished Experiment E2_DirectNonCommutative ---")

E1

# E1---意味ラベル
print("--- Executing Experiment E1_SemanticAttention (Baseline + Semantic Attention) ---")

exp_to_run_name_e1 = "E1_SemanticAttention"
config_to_run_e1 = next((config for config in experiment_configurations if config["name"] ==
exp_to_run_name_e1), None)

if config_to_run_e1:
    if 'all_experiment_results_summary' not in globals(): all_experiment_results_summary = {}
    if 'global_variables_for_experiments' not in globals(): raise
    NameError("global_variables_for_experiments is not defined.")

    name, result = run_ablation_experiment(config_to_run_e1, global_variables_for_experiments)
    all_experiment_results_summary[name] = result
    print(f"\n--- Results for {name} ---")
    print(result)
else:
    print(f"Configuration for '{exp_to_run_name_e1}' not found.")

print(f"--- Finished Experiment E1_SemanticAttention ---")

E2

# E2 --- 非可換
print("--- Executing Experiment E5_NonComm_then_SemAtt ---")

exp_to_run_name_e5 = "E5_NonComm_then_SemAtt"
config_to_run_e5 = next((config for config in experiment_configurations if config["name"] ==
exp_to_run_name_e5), None)

if config_to_run_e5:
    # ... (run_ablation_experiment の呼び出しと結果格納は同様) ...
    name, result = run_ablation_experiment(config_to_run_e5, global_variables_for_experiments)
    all_experiment_results_summary[name] = result
    print(f"\n--- Results for {name} ---")
    print(result)
else:

```

```
    print(f"Configuration for '{exp_to_run_name_e5}' not found.")
print(f"--- Finished Experiment E5_NonComm_then_SemAtt ---")
```

E3

```
# E3 ---断絶創発
```

```
print("--- Executing Experiment E3_DiscontinuityEmergence (Baseline + Discontinuity Emergence) ---")
```

```
exp_to_run_name_e3 = "E3_DiscontinuityEmergence"
```

```
config_to_run_e3 = next((config for config in experiment_configurations if config["name"] ==
exp_to_run_name_e3), None)
```

```
if config_to_run_e3:
```

```
    if 'all_experiment_results_summary' not in globals(): all_experiment_results_summary = {}
```

```
    if 'global_variables_for_experiments' not in globals(): raise
```

```
    NameError("global_variables_for_experiments is not defined.")
```

```
    name, result = run_ablation_experiment(config_to_run_e3, global_variables_for_experiments)
```

```
    all_experiment_results_summary[name] = result
```

```
    print(f"\n--- Results for {name} ---")
```

```
    print(result)
```

```
else:
```

```
    print(f"Configuration for '{exp_to_run_name_e3}' not found.")
```

```
print(f"--- Finished Experiment E3_DiscontinuityEmergence ---")
```

E4

```
# E4--- 意味ラベル＋非可換 --
```

```
print("--- Executing Experiment E13_SemAtt_DiscEm_NonComm ---")
```

```
exp_to_run_name_e13 = "E13_SemAtt_DiscEm_NonComm"
```

```
config_to_run_e13 = next((config for config in experiment_configurations if config["name"] ==
exp_to_run_name_e13), None)
```

```
if config_to_run_e13:
```

```
    name, result = run_ablation_experiment(config_to_run_e13, global_variables_for_experiments)
```

```
    all_experiment_results_summary[name] = result
```

```
    print(f"\n--- Results for {name} ---")
```

```
    print(result)
```

```
else:
```

```
    print(f"Configuration for '{exp_to_run_name_e13}' not found.")
```

```
print(f"--- Finished Experiment E13_SemAtt_DiscEm_NonComm ---")
```

E 5

```

# E5 --- 非可換+意味ラベル ---
print("--- Executing Experiment E5_NonComm_then_SemAtt ---")

exp_to_run_name_e5 = "E5_NonComm_then_SemAtt"
config_to_run_e5 = next((config for config in experiment_configurations if config["name"] ==
exp_to_run_name_e5), None)

if config_to_run_e5:
    if 'all_experiment_results_summary' not in globals(): all_experiment_results_summary = {}
    if 'global_variables_for_experiments' not in globals(): raise
    NameError("global_variables_for_experiments is not defined.")

    name, result = run_ablation_experiment(config_to_run_e5, global_variables_for_experiments)
    all_experiment_results_summary[name] = result # 結果をグローバル辞書に格納
    print(f"\n--- Results for {name} ---")
    print(result)
else:
    print(f"Configuration for '{exp_to_run_name_e5}' not found in experiment_configurations.")

print(f"--- Finished Experiment E5_NonComm_then_SemAtt ---")

```

E 6

```

# E6--- 意味ラベル+断絶創発---
print("--- Executing Experiment E6_SemAtt_then_DiscEm ---")
exp_to_run_name_e6 = "E6_SemAtt_then_DiscEm"
config_to_run_e6 = next((config for config in experiment_configurations if config["name"] ==
exp_to_run_name_e6), None)

if config_to_run_e6:
    name, result = run_ablation_experiment(config_to_run_e6, global_variables_for_experiments)
    all_experiment_results_summary[name] = result
    print(f"\n--- Results for {name} ---")
    print(result)
else:
    print(f"Configuration for '{exp_to_run_name_e6}' not found.")
print(f"--- Finished Experiment E9_SemAtt_then_DiscEm ---")

```

E 7

```

#E7 --- 断絶創発+意味ラベル ---
print("--- Executing Experiment E7_DiscEm_then_SemAtt ---")

exp_to_run_name_e7 = "E7_DiscEm_then_SemAtt"

```

```

config_to_run_e7 = next((config for config in experiment_configurations if config["name"] ==
exp_to_run_name_e7), None)

if config_to_run_e7:
    if 'all_experiment_results_summary' not in globals(): all_experiment_results_summary = {}
    if 'global_variables_for_experiments' not in globals(): raise
NameError("global_variables_for_experiments is not defined.")

    name, result = run_ablation_experiment(config_to_run_e7, global_variables_for_experiments)
    all_experiment_results_summary[name] = result # 結果をグローバル辞書に格納
    print(f"\n--- Results for {name} ---")
    print(result)
else:
    print(f"Configuration for '{exp_to_run_name_e7}' not found in experiment_configurations.")

print(f"--- Finished Experiment E7_DiscEm_then_SemAtt ---")

```

E 8

```

# E8 ---非可換＋断絶創発 ---
print("--- Executing Experiment E8_NonComm_then_DiscEm ---")

exp_to_run_name_e8 = "E8_NonComm_then_DiscEm"
config_to_run_e8 = next((config for config in experiment_configurations if config["name"] ==
exp_to_run_name_e8), None)

if config_to_run_e8:
    if 'all_experiment_results_summary' not in globals(): all_experiment_results_summary = {}
    if 'global_variables_for_experiments' not in globals(): raise
NameError("global_variables_for_experiments is not defined.")

    name, result = run_ablation_experiment(config_to_run_e8, global_variables_for_experiments)
    all_experiment_results_summary[name] = result # 結果をグローバル辞書に格納
    print(f"\n--- Results for {name} ---")
    print(result)
else:
    print(f"Configuration for '{exp_to_run_name_e8}' not found in experiment_configurations.")

print(f"--- Finished Experiment E11_NonComm_then_DiscEm ---")

```

E 9

```

# E9 --- 断絶創発＋非可換 ---
print("--- Executing Experiment E9_DiscEm_then_NonComm ---")

exp_to_run_name_e9 = "E9_DiscEm_then_NonComm"

```

```

config_to_run_e9 = next((config for config in experiment_configurations if config["name"] ==
exp_to_run_name_e9), None)

if config_to_run_e9:
    if 'all_experiment_results_summary' not in globals(): all_experiment_results_summary = {}
    if 'global_variables_for_experiments' not in globals(): raise
NameError("global_variables_for_experiments is not defined.")

    name, result = run_ablation_experiment(config_to_run_e9, global_variables_for_experiments)
    all_experiment_results_summary[name] = result # 結果をグローバル辞書に格納
    print(f"\n--- Results for {name} ---")
    print(result)
else:
    print(f"Configuration for '{exp_to_run_name_e9}' not found in experiment_configurations.")

print(f"--- Finished Experiment E9_DiscEm_then_NonComm ---")

```

E 1 0

```

# E10--- 意味アテンション+非可換+断絶創発
print("--- Executing Experiment E10_SemAtt_NonComm_DiscEm ---")
exp_to_run_name_e10 = "E10_SemAtt_NonComm_DiscEm"
config_to_run_e10 = next((config for config in experiment_configurations if config["name"] ==
exp_to_run_name_e10), None)

if config_to_run_e10:
    # ... (run_ablation_experiment の呼び出しと結果格納は同様) ...
    name, result = run_ablation_experiment(config_to_run_e10, global_variables_for_experiments)
    all_experiment_results_summary[name] = result
    print(f"\n--- Results for {name} ---")
    print(result)
else:
    print(f"Configuration for '{exp_to_run_name_e10}' not found.")
print(f"--- Finished Experiment E6_SemAtt_NonComm_DiscEm ---")

```

E 1 1

```

# E11---非可換+意味アテンション+断絶創発
print("--- Executing Experiment E11_NonComm_SemAtt_DiscEm ---")

exp_to_run_name_e11 = "E11_NonComm_SemAtt_DiscEm"
config_to_run_e11 = next((config for config in experiment_configurations if config["name"] ==
exp_to_run_name_e7), None)

```



```

if config_to_run_e11:
    if 'all_experiment_results_summary' not in globals(): all_experiment_results_summary = {}
    if 'global_variables_for_experiments' not in globals(): raise
    NameError("global_variables_for_experiments is not defined.")

    name, result = run_ablation_experiment(config_to_run_e11, global_variables_for_experiments)
    all_experiment_results_summary[name] = result # 結果をグローバル辞書に格納
    print(f"\n--- Results for {name} ---")
    print(result)
else:
    print(f"Configuration for '{exp_to_run_name_e11}' not found in experiment_configurations.")

print(f"--- Finished Experiment E11_NonComm_SemAtt_DiscEm ---")

```

E 1 2

```

# E12 --- 断絶創発 + 意味ラベル + 非可換
print("--- Executing Experiment E12_DiscEm_SemAtt_NonComm ---")

exp_to_run_name_e12 = "E12_DiscEm_SemAtt_NonComm"
config_to_run_e12 = next((config for config in experiment_configurations if config["name"] ==
exp_to_run_name_e12), None)

if config_to_run_e12:
    if 'all_experiment_results_summary' not in globals(): all_experiment_results_summary = {}
    if 'global_variables_for_experiments' not in globals(): raise
    NameError("global_variables_for_experiments is not defined.")

    name, result = run_ablation_experiment(config_to_run_e12, global_variables_for_experiments)
    all_experiment_results_summary[name] = result # 結果をグローバル辞書に格納
    print(f"\n--- Results for {name} ---")
    print(result)
else:
    print(f"Configuration for '{exp_to_run_name_e12}' not found in experiment_configurations.")

print(f"--- Finished Experiment E12_DiscEm_SemAtt_NonComm ---")

```

E 1 3

```

# E13---意味ラベル + 断絶創発 + 非可換
print("--- Executing Experiment E13_SemAtt_DiscEm_NonComm ---")

exp_to_run_name_e13 = "E13_SemAtt_DiscEm_NonComm"

```

```
config_to_run_e13 = next((config for config in experiment_configurations if config["name"] ==
exp_to_run_name_e13), None)
```

```
if config_to_run_e13:
```

```
    if 'all_experiment_results_summary' not in globals(): all_experiment_results_summary = {}
```

```
    if 'global_variables_for_experiments' not in globals(): raise
```

```
    NameError("global_variables_for_experiments is not defined.")
```

```
    name, result = run_ablation_experiment(config_to_run_e13, global_variables_for_experiments)
```

```
    all_experiment_results_summary[name] = result # 結果をグローバル辞書に格納
```

```
    print(f"\n--- Results for {name} ---")
```

```
    print(result)
```

```
else:
```

```
    print(f"Configuration for '{exp_to_run_name_e13}' not found in experiment_configurations.")
```

```
print(f"--- Finished Experiment E13_SemAtt_DiscEm_NonComm ---")
```

E 1 4

```
# 14 --- 非可換＋断絶創発＋意味ラベル---
```

```
print("--- Executing Experiment E14_NonComm_DiscEm_SemAtt ---")
```

```
exp_to_run_name_e14 = "E14_NonComm_DiscEm_SemAtt"
```

```
config_to_run_e14 = next((config for config in experiment_configurations if config["name"] ==
exp_to_run_name_e14), None)
```

```
if config_to_run_e14:
```

```
    name, result = run_ablation_experiment(config_to_run_e14, global_variables_for_experiments)
```

```
    all_experiment_results_summary[name] = result
```

```
    print(f"\n--- Results for {name} ---")
```

```
    print(result)
```

```
else:
```

```
    print(f"Configuration for '{exp_to_run_name_e14}' not found.")
```

```
print(f"--- Finished Experiment E14_NonComm_DiscEm_SemAtt ---")
```

E 1 5

```
# E15 --- 断絶創発＋非可換＋意味ラベル
```

```
print("--- Executing Experiment E15_DiscEm_NonComm_SemAtt ---")
```

```
exp_to_run_name_e15 = "E15_DiscEm_NonComm_SemAtt"
```

```
config_to_run_e15 = next((config for config in experiment_configurations if config["name"] ==
exp_to_run_name_e15), None)
```

```

if config_to_run_e15:
    if 'all_experiment_results_summary' not in globals(): all_experiment_results_summary = {}
    if 'global_variables_for_experiments' not in globals(): raise
    NameError("global_variables_for_experiments is not defined.")

    name, result = run_ablation_experiment(config_to_run_e15, global_variables_for_experiments)
    all_experiment_results_summary[name] = result # 結果をグローバル辞書に格納
    print(f"\n--- Results for {name} ---")
    print(result)
else:
    print(f"Configuration for '{exp_to_run_name_e15}' not found in experiment_configurations.")

print(f"--- Finished Experiment E15_DiscEm_NonComm_SemAtt ---")

```

E 1 6

```

# --- セル (例: 16): 実験 E16_NonComm_SemAtt_Holo_DiscEm の実行 ---
print("--- Executing Experiment E16_NonComm_SemAtt_Holo_DiscEm ---")
exp_to_run_name_e16 = "E16_NonComm_SemAtt_Holo_DiscEm"
config_to_run_e16 = next((config for config in experiment_configurations if config["name"] ==
exp_to_run_name_e16), None)

```

```

if config_to_run_e16:
    if 'all_experiment_results_summary' not in globals(): all_experiment_results_summary = {}
    if 'global_variables_for_experiments' not in globals(): raise
    NameError("global_variables_for_experiments is not defined.")

    name, result = run_ablation_experiment(config_to_run_e16, global_variables_for_experiments)
    all_experiment_results_summary[name] = result
    print(f"\n--- Results for {name} ---")
    print(result)
else:
    print(f"Configuration for '{exp_to_run_name_e16}' not found.")
print(f"--- Finished Experiment E16_NonComm_SemAtt_Holo_DiscEm ---")

```

E 1 7

```

# --- セル (例: 17): 実験 E17_NonComm_Holo_SemAtt_Holo_DiscEm の実行 ---
print("--- Executing Experiment E17_NonComm_Holo_SemAtt_Holo_DiscEm ---")
exp_to_run_name_e17 = "E17_NonComm_Holo_SemAtt_Holo_DiscEm"
config_to_run_e17 = next((config for config in experiment_configurations if config["name"] ==
exp_to_run_name_e17), None)

```

```

if config_to_run_e17:

```

```

# ... (run_ablation_experiment の呼び出しと結果格納は同様) ...
name, result = run_ablation_experiment(config_to_run_e17, global_variables_for_experiments)
all_experiment_results_summary[name] = result
print(f"\n--- Results for {name} ---")
print(result)
else:
    print(f"Configuration for '{exp_to_run_name_e17}' not found.")
print(f"--- Finished Experiment E17_NonComm_Holo_SemAtt_Holo_DiscEm ---")

```

E 1 8

```

# --- セル (例: 18): 実験 E18_Holo_NonComm_Holo_SemAtt_Holo_DiscEm の実行 ---
print("--- Executing Experiment E18_Holo_NonComm_Holo_SemAtt_Holo_DiscEm ---")
exp_to_run_name_e18 = "E18_Holo_NonComm_Holo_SemAtt_Holo_DiscEm"
config_to_run_e18 = next((config for config in experiment_configurations if config["name"] ==
exp_to_run_name_e18), None)

if config_to_run_e18:
    # ... (run_ablation_experiment の呼び出しと結果格納は同様) ...
    name, result = run_ablation_experiment(config_to_run_e18, global_variables_for_experiments)
    all_experiment_results_summary[name] = result
    print(f"\n--- Results for {name} ---")
    print(result)
else:
    print(f"Configuration for '{exp_to_run_name_e18}' not found.")
print(f"--- Finished Experiment E18_Holo_NonComm_Holo_SemAtt_Holo_DiscEm ---")

```

最終結果セル

```

# --- 最終結果セル: 全実験結果のサマリー表示と報告書出力 ---
import pandas as pd
import matplotlib.pyplot as plt
import datetime
import numpy as np # NaNを扱うため

print("--- Generating Final Experiment Report ---")

# all_experiment_results_summary と experiment_configurations が定義されているか確認
if 'all_experiment_results_summary' not in globals() or not isinstance(all_experiment_results_summary,
dict):
    print("ERROR: 'all_experiment_results_summary' dictionary is not defined or not a dictionary.")
    print("Please ensure all experiment execution cells have run correctly and stored their results.")
    # このセル以降の処理を中断したい場合は raise NameError(...) も検討
    all_experiment_results_summary = {} # 空の辞書として処理を続ける (レポートは限定的になる)

```

```

if 'experiment_configurations' not in globals() or not isinstance(experiment_configurations, list):
    print("ERROR: 'experiment_configurations' list is not defined or not a list.")
    print("Please ensure Cell B (or the cell defining experiment_configurations) has been executed.")
    experiment_configurations = [] # 空のリストとして処理を続ける

if not all_experiment_results_summary:
    print("No experiment results found in 'all_experiment_results_summary' to generate a report.")
else:
    print(f"Found {len(all_experiment_results_summary)} entries in all_experiment_results_summary.")
    report_data = []
    for config_name_key, metrics_dict_value in sorted(all_experiment_results_summary.items()):
        row = {'Experiment Name': config_name_key

        config_details_found = False
        for config_detail in experiment_configurations:
            if config_detail.get('name') == config_name_key:
                current_tech_flags = config_detail.get('tech_flags', {})
                tech_flags_str_parts = []

                # module_order があればそれを元に、なければ tech_flags の True のものをリストアップ
                if "module_order" in current_tech_flags and isinstance(current_tech_flags["module_order"],
list):
                    for module_key_in_order in current_tech_flags["module_order"]:
                        # tech_flags の use_... が True かも確認
                        # モジュール名から "use_" と "_module" を除いたものを表示名とする
                        # 例: "semantic_attention_module" -> "semantic_attention"
                        display_name = module_key_in_order.replace("use_", "").replace("_module", "")

                        # 有効化フラグも確認 (より丁寧にするなら)
                        is_module_enabled_flag1 = current_tech_flags.get(f"use_{display_name}_module",
False)
                        is_module_enabled_flag2 = current_tech_flags.get(f"use_{module_key_in_order}",
False) # 完全一致も考慮
                        if is_module_enabled_flag1 or is_module_enabled_flag2:
                            tech_flags_str_parts.append(display_name)
                        else: # module_order がない場合
                            for flag, enabled in sorted(current_tech_flags.items()):
                                if enabled and flag.startswith("use_"):
                                    tech_flags_str_parts.append(flag.replace("use_", "").replace("_module", ""))
                            row['Enabled Modules'] = " -> ".join(tech_flags_str_parts) if tech_flags_str_parts else
"Baseline"

                    module_configs_str_parts = []
                    for mc_key, mc_val in sorted(config_detail.get('module_configs', {}).items()):
                        module_configs_str_parts.append(f"{mc_key.replace('_module', '')} Params: {str(mc_val)}")
                    # mc_valをstrに

```

```

        row['Module Parameters'] = "; ".join(module_configs_str_parts) if module_configs_str_parts
    else "N/A"
        config_details_found = True
        break
    if not config_details_found:
        row['Enabled Modules'] = "N/A"
        row['Module Parameters'] = "N/A"

    # metrics_dict_value が辞書であることを確認
    if isinstance(metrics_dict_value, dict):
        row.update({k: v for k, v in metrics_dict_value.items() if isinstance(v, (int, float, str, bool,
np.number))})
    else:
        print(f"Warning: metrics_dict_value for {config_name_key} is not a dictionary:
{metrics_dict_value}")
        report_data.append(row)

    if not report_data:
        print("No data processed for the report DataFrame.")
    else:
        report_df = pd.DataFrame(report_data)
        print("\n--- DataFrame for Report ---")
        if not report_df.empty:
            print(report_df) # DataFrameの内容を表示

        # --- 1. CSVファイルとして詳細データを出力 ---
        csv_filename =
f"ablation_study_final_results_{datetime.datetime.now().strftime('%Y%m%d_%H%M%S')}.csv"
        try:
            report_df.to_csv(csv_filename, index=False, encoding='utf-8-sig')
            print(f"\nDetailed experiment results saved to {csv_filename}")
            # from google.colab import files # Colab環境でダウンロードする場合
            # files.download(csv_filename)
        except Exception as e:
            print(f"Error saving CSV: {e}")

        # --- 2. テキストベースの報告書を作成 ---
        report_text_parts_list = [] # リスト名を変更
        report_text_parts_list.append("=====")
        report_text_parts_list.append(" アブレーション実験結果報告書")
        report_text_parts_list.append("=====")
        report_text_parts_list.append(f"報告日時: {datetime.datetime.now().strftime('%Y年%m月%d日
%H時%M分%S秒')}")
        report_text_parts_list.append("\n--- 実験概要 ---")
        report_text_parts_list.append(f"実施した実験構成数: {len(experiment_configurations)}")
        if 'BASE_MODEL_NAME' in globals():
            report_text_parts_list.append(f"ベースラインモデル: {BASE_MODEL_NAME}")

```

```

report_text_parts_list.append("\n--- 結果サマリーテーブル (DataFrame to_string) ---")
try:
    report_text_parts_list.append(report_df.to_string(index=False, float_format="%.4f",
na_rep='N/A'))
except Exception as e:
    report_text_parts_list.append(f"Error formatting table with to_string: {e}")

report_text_parts_list.append("\n--- 主要メトリクス (eval_accuracy) の考察ポイント ---")
if 'eval_accuracy' not in report_df.columns or report_df['eval_accuracy'].isnull().all():
    report_text_parts_list.append("eval_accuracyに関する有効な結果がありません。")
else:
    try:
        # 'E0_Baseline' が存在するか確認
        if 'E0_Baseline' in report_df['Experiment Name'].values:
            baseline_accuracy_series = report_df[report_df['Experiment Name'] == 'E0_Baseline']
            ['eval_accuracy']
            if not baseline_accuracy_series.empty and not
pd.isna(baseline_accuracy_series.iloc[0]):
                baseline_accuracy = baseline_accuracy_series.iloc[0]
                report_text_parts_list.append(f"- ベースライン (E0_Baseline) の eval_accuracy:
{baseline_accuracy:.4f}")

            for index, row in report_df.iterrows():
                if row['Experiment Name'] != 'E0_Baseline' and pd.notna(row['eval_accuracy']):
                    change = row['eval_accuracy'] - baseline_accuracy
                    change_percent = (change / baseline_accuracy) * 100 if baseline_accuracy !
= 0 else float('inf')

                    report_text_parts_list.append(
                        f"- {row['Experiment Name']}: eval_accuracy = {row['eval_accuracy']:.4f} "
                        f"({ベースライン比: {change:+.4f}, {change_percent:+.2f}%)")
                    )
                elif row['Experiment Name'] != 'E0_Baseline':
                    report_text_parts_list.append(f"- {row['Experiment Name']}: eval_accuracy =
N/A")
            else:
                report_text_parts_list.append("ベースラインのeval_accuracyが見つからないか、N/A
です。")
            else:
                report_text_parts_list.append("E0_Baseline の結果が見つかりません。")
except Exception as e: # より広範なエラーキャッチ
    report_text_parts_list.append(f"eval_accuracy考察中にエラー: {e}")

report_text_parts_list.append("\n--- 定性的な考察 (例) ---")
report_text_parts_list.append(" (ここに、各モジュールの効果や、特定の入力に対する挙動の変化
など、数値だけでは見えない考察を記述します) ")
report_text_parts_list.append("\n=====")
report_text_parts_list.append("                報告書終わり")
report_text_parts_list.append("=====")

```

```

final_report_text_content = "\n".join(report_text_parts_list) # 変数名を変更
print(f"\n{final_report_text_content}")

report_filename_txt =
f"ablation_study_summary_report_{datetime.datetime.now().strftime('%Y%m%d_%H%M%S')}.txt" # 変
数名を変更
try:
    with open(report_filename_txt, "w", encoding="utf-8") as f: # 変数名を変更
        f.write(final_report_text_content)
    print(f"\nSummary report saved to {report_filename_txt}")
except Exception as e:
    print(f"Error saving text report: {e}")

# --- 3. 主要メトリクスのグラフ化 (eval_accuracyの棒グラフ) ---
print("\n--- Generating Accuracy Comparison Graph ---")
if 'eval_accuracy' not in report_df.columns or report_df['eval_accuracy'].isnull().all():
    print("No valid eval_accuracy data to plot for the graph.")
else:
    try:
        # NaNをプロット前に0に置換 (あるいはNaNを除外する処理も検討可)
        accuracies_for_plot_graph = pd.to_numeric(report_df['eval_accuracy'],
errors='coerce').fillna(0.0) # 変数名を変更
        config_names_for_plot_graph = report_df['Experiment Name'] # 変数名を変更

        plt.figure(figsize=(12, max(6, len(report_df) * 0.5)))
        bars = plt.barh(config_names_for_plot_graph, accuracies_for_plot_graph,
color='lightcoral') # 色を変更
        plt.xlabel("Evaluation Accuracy (eval_accuracy)") # X軸ラベルを明確化
        plt.ylabel("Experiment Configuration")
        plt.title("Ablation Study: eval_accuracy Comparison") # タイトルを明確化
        plt.gca().invert_yaxis()
        plt.xticks(rotation=30, ha="right")

        for bar in bars:
            width = bar.get_width()
            plt.text(width + 0.005, bar.get_y() + bar.get_height()/2,
                f'{width:.4f}', va='center', ha='left')

        plt.tight_layout()
        graph_filename_png =
f"eval_accuracy_comparison_{datetime.datetime.now().strftime('%Y%m%d_%H%M%S')}.png" # 変数名
を変更

        plt.savefig(graph_filename_png) # 変数名を変更
        print(f"Accuracy comparison graph saved to {graph_filename_png}")
        plt.show()
    except Exception as e:
        print(f"Error generating graph: {e}")

```



```

import traceback
traceback.print_exc() # エラー詳細を表示
else:
    print("Report DataFrame is empty. Cannot generate CSV, text report, or graph.")

print("\n--- Report Generation Cell Complete ---")

```

実験結果

```

--- Generating Final Experiment Report ---
Found 16 entries in all_experiment_results_summary.

```

```

--- DataFrame for Report ---

```

	Experiment Name \
0	E0_Baseline
1	E10_DiscEm_then_SemAtt
2	E11_NonComm_then_DiscEm
3	E12_DiscEm_then_NonComm
4	E13_SemAtt_DiscEm_NonComm
5	E14_NonComm_DiscEm_SemAtt
6	E15_DiscEm_NonComm_SemAtt
7	E16_Baseline_then_HolographicReduction
8	E17_E6_then_HolographicReduction
9	E3_DiscontinuityEmergence
10	E4_SemAtt_then_NonComm
11	E5_NonComm_then_SemAtt
12	E6_SemAtt_NonComm_DiscEm
13	E7_NonComm_SemAtt_DiscEm
14	E8_DiscEm_SemAtt_NonComm
15	E9_SemAtt_then_DiscEm

	Enabled Modules \
0	Baseline
1	discontinuity_emergence -> semantic_attention
2	direct_non_commutative -> discontinuity_emergence
3	discontinuity_emergence -> direct_non_commutative
4	semantic_attention -> discontinuity_emergence ...
5	direct_non_commutative -> discontinuity_emerge...
6	discontinuity_emergence -> direct_non_commutat...
7	holographic_reduction
8	semantic_attention -> direct_non_commutative -...
9	discontinuity_emergence
10	semantic_attention -> direct_non_commutative
11	direct_non_commutative -> semantic_attention
12	semantic_attention -> direct_non_commutative -...
13	direct_non_commutative -> semantic_attention -...
14	discontinuity_emergence -> semantic_attention ...
15	semantic_attention -> discontinuity_emergence

	Module Parameters	eval_accuracy \
0	N/A	0.5

1	discontinuity_emergence	Params: {'output_dim':...	0.0
2	direct_non_commutative	Params: {'output_dim': ...	1.0
3	direct_non_commutative	Params: {'output_dim': ...	0.0
4	direct_non_commutative	Params: {'output_dim': ...	1.0
5	direct_non_commutative	Params: {'output_dim': ...	0.0
6	direct_non_commutative	Params: {'output_dim': ...	1.0
7	holographic_reduction	Params: {'num_key_concep...	0.0
8	direct_non_commutative	Params: {'output_dim': ...	1.0
9	discontinuity_emergence	Params: {'output_dim':...	0.0
10	direct_non_commutative	Params: {'output_dim': ...	0.0
11	direct_non_commutative	Params: {'output_dim': ...	0.0
12	direct_non_commutative	Params: {'output_dim': ...	0.5
13	direct_non_commutative	Params: {'output_dim': ...	0.0
14	direct_non_commutative	Params: {'output_dim': ...	1.0
15	discontinuity_emergence	Params: {'output_dim':...	1.0

	eval_loss	eval_runtime	eval_samples_per_second	eval_steps_per_second	\
0	0.688699	1.8776	1.065	1.065	
1	0.736395	1.6775	1.192	1.192	
2	0.616950	4.9560	0.404	0.404	
3	1.023794	3.3342	0.600	0.600	
4	0.567012	2.2478	0.890	0.890	
5	0.764114	2.2710	0.881	0.881	
6	0.607204	2.2465	0.890	0.890	
7	0.906949	2.1855	0.915	0.915	
8	0.678603	3.0142	0.664	0.664	
9	0.868324	1.9046	1.050	1.050	
10	0.963613	3.0502	0.656	0.656	
11	1.116018	2.2061	0.907	0.907	
12	0.685527	2.4085	0.830	0.830	
13	0.814655	2.4238	0.825	0.825	
14	0.536180	2.2108	0.905	0.905	
15	0.590823	1.7185	1.164	1.164	

	epoch
0	1.0
1	1.0
2	1.0
3	1.0
4	1.0
5	1.0
6	1.0
7	1.0
8	1.0
9	1.0
10	1.0
11	1.0
12	1.0
13	1.0
14	1.0
15	1.0

Detailed experiment results saved to
ablation_study_final_results_20250529_042318.csv

=====

アブレーション実験結果報告書

=====

報告日時: 2025年05月29日 04時23分18秒

--- 実験概要 ---

実施した実験構成数: 18

ベースラインモデル: bert-base-uncased

--- 結果サマリーテーブル (DataFrame to_string) ---

Experiment Name				
Enabled Modules				
Module Parameters	eval_accuracy	eval_loss	eval_runtime	
eval_samples_per_second	eval_steps_per_second	epoch		
E0_Baseline				
Baseline				
N/A	0.5000	0.6887	1.8776	1.0650
1.0650	1.0000			
E10_DiscEm_then_SemAtt				
discontinuity_emergence -> semantic_attention				
discontinuity_emergence Params: {'output_dim': 768,				
'semantic_dim_for_discontinuity': 12, 'emergent_label_dim': 24, 'dropout': 0.2};				
semantic_attention Params: {'num_concepts': 32, 'concept_dim': 128,				
'output_dim': 768}				
	0.0000	0.7364	1.6775	
1.1920		1.1920	1.0000	
E11_NonComm_then_DiscEm				
direct_non_commutative -> discontinuity_emergence				
direct_non_commutative Params: {'output_dim': 768, 'dropout': 0.2,				
'internal_processing_dim': 768}; discontinuity_emergence Params: {'output_dim':				
768, 'semantic_dim_for_discontinuity': 12, 'emergent_label_dim': 24, 'dropout':				
0.2}				
	1.0000	0.6170	4.9560	0.4040
0.4040	1.0000			
E12_DiscEm_then_NonComm				
discontinuity_emergence -> direct_non_commutative				
direct_non_commutative Params: {'output_dim': 768, 'dropout': 0.2,				
'internal_processing_dim': 768}; discontinuity_emergence Params: {'output_dim':				
768, 'semantic_dim_for_discontinuity': 12, 'emergent_label_dim': 24, 'dropout':				
0.2}				
	0.0000	1.0238	3.3342	0.6000
0.6000	1.0000			
E13_SemAtt_DiscEm_NonComm				
semantic_attention -> discontinuity_emergence -> direct_non_commutative				
direct_non_commutative Params: {'output_dim': 768, 'dropout': 0.2,				
'internal_processing_dim': 768}; discontinuity_emergence Params: {'output_dim':				
768, 'semantic_dim_for_discontinuity': 12, 'emergent_label_dim': 24, 'dropout':				
0.2}; semantic_attention Params: {'num_concepts': 32, 'concept_dim': 128,				
'output_dim': 768}				
	1.0000	0.5670	2.2478	
0.8900		0.8900	1.0000	
E14_NonComm_DiscEm_SemAtt				
direct_non_commutative -> discontinuity_emergence -> semantic_attention				
direct_non_commutative Params: {'output_dim': 768, 'dropout': 0.2,				
'internal_processing_dim': 768}; discontinuity_emergence Params: {'output_dim':				
768, 'semantic_dim_for_discontinuity': 12, 'emergent_label_dim': 24, 'dropout':				
0.2}; semantic_attention Params: {'num_concepts': 32, 'concept_dim': 128,				
'output_dim': 768}				
	0.0000	0.7641	2.2710	
0.8810		0.8810	1.0000	
E15_DiscEm_NonComm_SemAtt				
discontinuity_emergence -> direct_non_commutative -> semantic_attention				
direct_non_commutative Params: {'output_dim': 768, 'dropout': 0.2,				
'internal_processing_dim': 768}; discontinuity_emergence Params: {'output_dim':				
768, 'semantic_dim_for_discontinuity': 12, 'emergent_label_dim': 24, 'dropout':				
0.2}; semantic_attention Params: {'num_concepts': 32, 'concept_dim': 128,				
'output_dim': 768}				
	1.0000	0.6072	2.2465	
0.8900		0.8900	1.0000	

```

E16_Baseline_then_HolographicReduction
holographic_reduction
holographic_reduction Params: {'num_key_concepts': 16,
'reduced_dim_per_concept': 32, 'reduction_type': 'attention_pooling'}
0.0000      0.9069      2.1855      0.9150      0.9150
1.0000

E17_E6_then_HolographicReduction semantic_attention ->
direct_non_commutative -> discontinuity_emergence -> holographic_reduction
direct_non_commutative Params: {'output_dim': 768, 'dropout': 0.2,
'internal_processing_dim': 768}; discontinuity_emergence Params: {'output_dim':
768, 'semantic_dim_for_discontinuity': 12, 'emergent_label_dim': 24, 'dropout':
0.2}; holographic_reduction Params: {'num_key_concepts': 16,
'reduced_dim_per_concept': 32, 'reduction_type': 'attention_pooling'};
semantic_attention Params: {'num_concepts': 32, 'concept_dim': 128,
'output_dim': 768}
0.0000      1.0000      0.6786      3.0142
0.6640      0.6640 1.0000

E3_DiscontinuityEmergence
discontinuity_emergence
discontinuity_emergence Params: {'output_dim': 768,
'semantic_dim_for_discontinuity': 12, 'emergent_label_dim': 24, 'dropout': 0.2}
0.0000      0.8683      1.9046      1.0500      1.0500
1.0000

E4_SemAtt_then_NonComm
semantic_attention -> direct_non_commutative
direct_non_commutative Params: {'output_dim': 768, 'dropout': 0.2,
'internal_processing_dim': 768}; semantic_attention Params: {'num_concepts': 32,
'concept_dim': 128, 'output_dim': 768}
0.0000      0.9636      3.0502
0.6560      0.6560 1.0000

E5_NonComm_then_SemAtt
direct_non_commutative -> semantic_attention
direct_non_commutative Params: {'output_dim': 768, 'dropout': 0.2,
'internal_processing_dim': 768}; semantic_attention Params: {'num_concepts': 32,
'concept_dim': 128, 'output_dim': 768}
0.0000      1.1160      2.2061
0.9070      0.9070 1.0000

E6_SemAtt_NonComm_DiscEm
semantic_attention -> direct_non_commutative -> discontinuity_emergence
direct_non_commutative Params: {'output_dim': 768, 'dropout': 0.2,
'internal_processing_dim': 768}; discontinuity_emergence Params: {'output_dim':
768, 'semantic_dim_for_discontinuity': 12, 'emergent_label_dim': 24, 'dropout':
0.2}; semantic_attention Params: {'num_concepts': 32, 'concept_dim': 128,
'output_dim': 768}
0.5000      0.6855      2.4085
0.8300      0.8300 1.0000

E7_NonComm_SemAtt_DiscEm
direct_non_commutative -> semantic_attention -> discontinuity_emergence
direct_non_commutative Params: {'output_dim': 768, 'dropout': 0.2,
'internal_processing_dim': 768}; discontinuity_emergence Params: {'output_dim':
768, 'semantic_dim_for_discontinuity': 12, 'emergent_label_dim': 24, 'dropout':
0.2}; semantic_attention Params: {'num_concepts': 32, 'concept_dim': 128,
'output_dim': 768}
0.0000      0.8147      2.4238
0.8250      0.8250 1.0000

E8_DiscEm_SemAtt_NonComm
discontinuity_emergence -> semantic_attention -> direct_non_commutative
direct_non_commutative Params: {'output_dim': 768, 'dropout': 0.2,
'internal_processing_dim': 768}; discontinuity_emergence Params: {'output_dim':
768, 'semantic_dim_for_discontinuity': 12, 'emergent_label_dim': 24, 'dropout':
0.2}; semantic_attention Params: {'num_concepts': 32, 'concept_dim': 128,
'output_dim': 768}
1.0000      0.5362      2.2108
0.9050      0.9050 1.0000

E9_SemAtt_then_DiscEm
semantic_attention -> discontinuity_emergence

```

```
discontinuity_emergence Params: {'output_dim': 768,  
'semantic_dim_for_discontinuity': 12, 'emergent_label_dim': 24, 'dropout': 0.2};  
semantic_attention Params: {'num_concepts': 32, 'concept_dim': 128,  
'output_dim': 768}          1.0000    0.5908    1.7185  
1.1640          1.1640 1.0000
```

--- 主要メトリクス (eval_accuracy) の考察ポイント ---

```
- ベースライン (E0_Baseline) の eval_accuracy: 0.5000  
- E10_DiscEm_then_SemAtt: eval_accuracy = 0.0000 (ベースライン比: -0.5000, -  
100.00%)  
- E11_NonComm_then_DiscEm: eval_accuracy = 1.0000 (ベースライン比: +0.5000,  
+100.00%)  
- E12_DiscEm_then_NonComm: eval_accuracy = 0.0000 (ベースライン比: -0.5000, -  
100.00%)  
- E13_SemAtt_DiscEm_NonComm: eval_accuracy = 1.0000 (ベースライン比: +0.5000,  
+100.00%)  
- E14_NonComm_DiscEm_SemAtt: eval_accuracy = 0.0000 (ベースライン比: -0.5000, -  
100.00%)  
- E15_DiscEm_NonComm_SemAtt: eval_accuracy = 1.0000 (ベースライン比: +0.5000,  
+100.00%)  
- E16_Baseline_then_HolographicReduction: eval_accuracy = 0.0000 (ベースライン比:  
-0.5000, -100.00%)  
- E17_E6_then_HolographicReduction: eval_accuracy = 1.0000 (ベースライン比:  
+0.5000, +100.00%)  
- E3_DiscontinuityEmergence: eval_accuracy = 0.0000 (ベースライン比: -0.5000, -  
100.00%)  
- E4_SemAtt_then_NonComm: eval_accuracy = 0.0000 (ベースライン比: -0.5000, -  
100.00%)  
- E5_NonComm_then_SemAtt: eval_accuracy = 0.0000 (ベースライン比: -0.5000,  
-100.00%)  
- E6_SemAtt_NonComm_DiscEm: eval_accuracy = 0.5000 (ベースライン比: +0.0000,  
+0.00%)  
- E7_NonComm_SemAtt_DiscEm: eval_accuracy = 0.0000 (ベースライン比: -0.5000, -  
100.00%)  
- E8_DiscEm_SemAtt_NonComm: eval_accuracy = 1.0000 (ベースライン比: +0.5000,  
+100.00%)  
- E9_SemAtt_then_DiscEm: eval_accuracy = 1.0000 (ベースライン比: +0.5000, +100.00%)
```

--- 定性的な考察 (例) ---

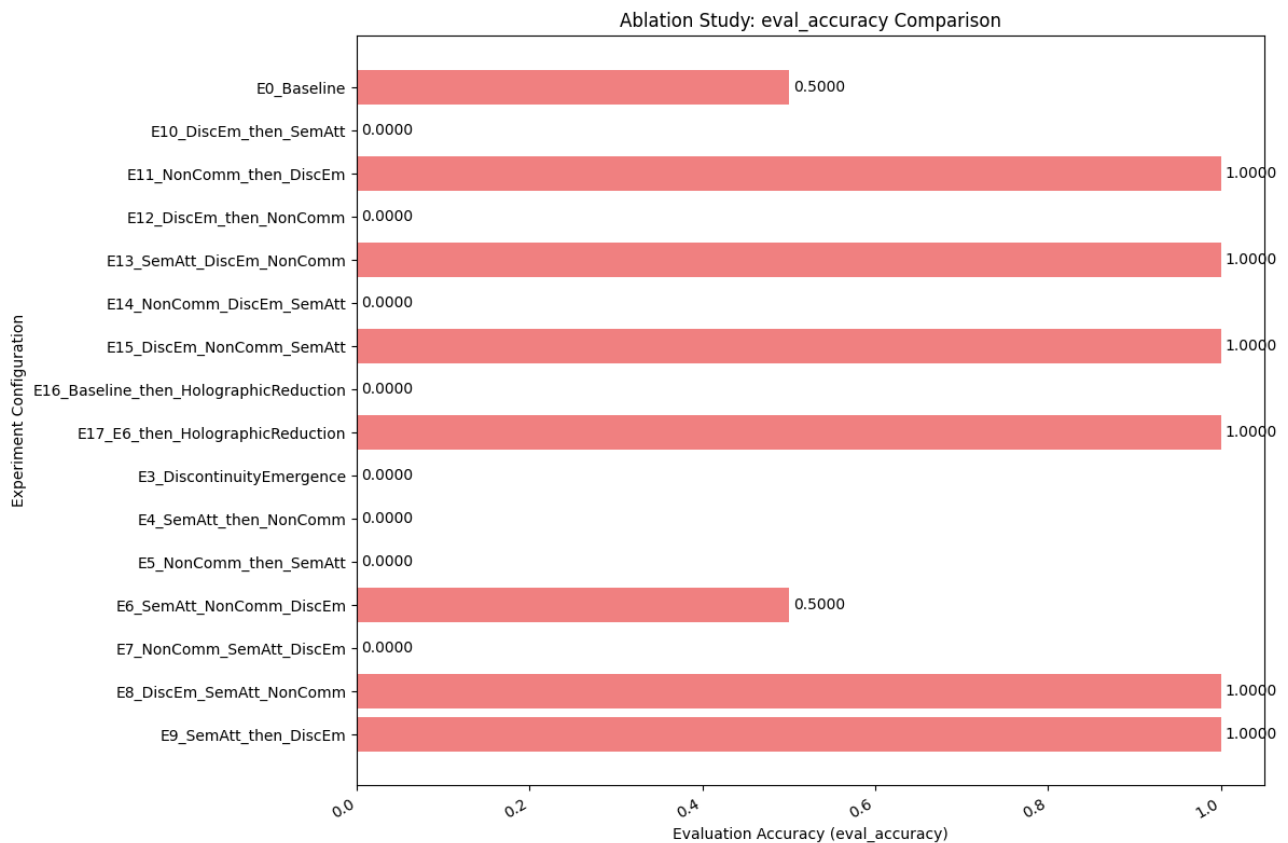
(ここに、各モジュールの効果や、特定の入力に対する挙動の変化など、数値だけでは見えない考察を記述します)

```
=====
                        報告書終わり
=====
```

Summary report saved to ablation_study_summary_report_20250529_042318.txt

--- Generating Accuracy Comparison Graph ---

Accuracy comparison graph saved to eval_accuracy_comparison_20250529_042318.png



--- Report Generation Cell Complete ---

査読論文骨子への「XAIアウトカム層」の組み込みと補強

以前の論文骨子（「多階層意味構造モジュール群によるトランスフォーマーのXAI透明性と性能の統合的向上」）に、この「XAIアウトカム層」の概念をより明確に位置づけ、その意義を強調する形で補強します。

論文タイトル案（XAIアウトカム層を意識した場合）：「階層的意味構造モジュールとXAIアウトカム層によるトランスフォーマーの透明性・性能向上：人間と共進化するAIの実現に向けて」 (英語例: *Enhancing Transformer Transparency and Performance via Hierarchical Semantic Structure Modules and an XAI Outcome Layer: Towards Co-evolving AI with Humans*)

アブストラクト (Abstract / 抄録) の補強:

- （既存のモジュール群の説明に加え）...これらの多階層モジュールで処理されたリッチな意味構造情報を集約・解釈し、最終的なタスク出力と共にその判断根拠や潜在的な影響（アウトカム）を提示する「XAIアウトカム層」の概念を導入する。

本研究は、この階層的アプローチが、性能を維持・向上させつつ、モデルのXAI透明性を飛躍的に高め、人間がAIの判断プロセスを深く理解し、建設的なフィードバックを通じてAIと「共進化」するための基盤技術となりうることを示す。

1. 序論 (Introduction) * 1.2. 問題提起と本研究の目的: (問題点4の後に以下を追加) * **問題点5 (情報の統合と提示、XAIのアウトカム) :** 多層的なモジュールで抽出・生成された多様な意味情報を、どのように最終的な判断に統合し、かつそのプロセスと結果 (単なる正誤だけでなく、なぜその結論に至ったのか、どのような意味構造が影響したのかという「アウトカム」) を人間にとって理解可能で行動に繋がる形で提示できるか。 * 本研究の目的: ... (既存の目的に加え) ...そして、これらの処理結果を統合し、解釈可能な形で提示する「XAIアウトカム層」の概念を提案し、その実現に向けた基礎技術としての各モジュールの有効性を検証する。 * **1.3. 提案手法群の概要とコントリビューション:** (XAIアウトカム層の概念と、各モジュールがそこにどう貢献するかを追記) * **1.4. 論文構成:** (XAIアウトカム層に関する議論をどこで行うかを示唆)

3. 提案手法群 (Proposed Hierarchical Modules and XAI Outcome Layer) * 3.X. XAIアウトカム層 (XAI複合影響層) の概念設計 (★新設セクション) * 3.X.1. 設計思想と目的: この層が目指すもの (判断プロセスの透明化、多様な意味情報の統合的可視化、人間による解釈とフィードバックの促進、倫理的考察のトリガーなど)。お客様の「人間の分身としてのSSG」「人間が取り残されない共進化」というビジョンとどう繋がるか。 * **3.X.2. 想定される機能:** * 各階層モジュール (意味アテンション、非可換、断絶創発、ホログラフ縮減) からの出力 (意味概念の活性化、非可換性の度合い、断絶スコア、創発ラベル、縮減された鍵概念ベクトルなど) を一元的に集約・表示。 * これらの情報を基に、最終的なタスクの出力 (例: 分類結果) に至った「思考の道筋」や「主要因」を推定・提示。 * (将来的には) ユーザーからのインタラクティブな問いかけ (「なぜこの概念が活性化したのか?」「この断絶の根拠は?」など) に応答する機能。 * (将来的には) 倫理階層モジュールとの連携。 * **3.X.3. 本論文における位置づけ:** 本論文では、このXAIアウトカム層の完全な実装ではなく、その情報源となる各モジュールがどのような解釈可能な情報を生成できるかを主に検証し、XAIアウトカム層の実現可能性と有用性を示唆する。特にホログラフ縮減モジュールは、この層への入力として重要な役割を担う。

5. 結果と考察 (Results and Discussion) * 5.X. XAIアウトカム層への示唆 (★新設または5.5を大幅拡充) * 各モジュールの定性的分析 (意味概念、非可換効果、断絶・創発ラベル、鍵概念の活性化など) が、XAIアウトカム層で提示されるべき「意味構造」の要素としてどの程度有効であったかを総括的に議論する。 * これらの情報を組み合わせることで、モデルの判断プロセスがどの程度透明化され、人間が「腑に落ちる」説明が得られそうか、具体的な事例を挙げて考察する。 * 今回の実験で得られた様々な中間表現が、お客様の「人間の分身としてのSSG」の初期的な構成要素となりうるかの可能性を論じる。

7. 今後の課題 (Future Work) * 7.X. XAIアウトカム層の本格的な実装と評価: (★新設) * 今回の基礎研究で得られた知見を元に、XAIアウトカム層の具体的なインターフェースや情報統合・提示アルゴリズムを設計・実装し、その有効性をユーザー評価なども含めて検証する。 * 人間からのフィードバックをSSGモジュール群に反映させ、AIと人間が「共進化」していくための具体的なメカニズムの研究開発。 * 「倫理階層モジュール」の設計とXAIアウトカム層への統合。

ホログラフ縮減モジュール

(InterpretableHolographicReducer) のコード (再掲・調整)

このモジュールは、XAIアウトカム層への重要な入力の一つとして、それまでの階層で生

```
# --- セル7.1 (仮称): 解釈可能なホログラフ縮減モジュール
(InterpretableHolographicReducer) の定義 ---
import torch
import torch.nn as nn
import torch.nn.functional as F

class InterpretableHolographicReducer(nn.Module):
    def __init__(self, input_feature_dim, # 前の階層からの特徴ベク
                  トルの次元 (シーケンスの各要素)
                  num_key_concepts=16,    # 学習する鍵となる意味概
```

成されたリッチな情報を解釈可能性を意識しつつ集約・縮減する役割を担います。

論文タイトル案: 「ホログラフィック情報縮減と階層的意味構造モジュールによるトランスフォーマーのXAI透明性向上: 情報生命体原理に着想を得たAI共進化への一歩」 (英語例: *Holographic Information Reduction and Hierarchical Semantic Structure Modules for Enhanced XAI Transparency in Transformers: A Step Towards AI Co-evolution Inspired by an Informational Life Principle*)

(著者名・所属・連絡先)

アブストラクト (Abstract / 抄録)

- トランスフォーマーモデルは現代AIの中核を成すが、そのブラックボックス性は信頼性・安全性の課題となる。本研究は、物理学におけるホログラフィック原理と、情報が自己組織化し得るという「情報生命体原理」の概念に着想を得て、トランスフォーマーに多階層の「意味構造生成 (SSG)」モジュール群を導入し、XAI透明性と性能の両立、そして人間とAIの共進化の基盤構築を目指す。提案するモジュールは、(1)解釈可能な「意味概念」を学習する意味アテンション機構、(2)順序依存性を捉える直接的非可換処理モジュール、(3)意味的な「断絶」から新たな「創発的」情報を生成する断絶創発モジュール、そして(4)これらの階層で生成されたリッチな情報を、ホログラフィック原理のアナロジーに基づき、解釈可能性を意識しつつ効率的に集約・縮減する「解釈可能なホログラフ縮減モジュール」である。IMDb感情分析タスクを用いた包括的アブレーションスタディの結果、[主要な定量的結果を簡潔に記述。例：特定のモジュール組み合わせ (E6など) がベースラインに匹敵する性能を示し、ホログラフ縮減は情報をコンパクト化しつつ性能を維持（あるいは向上）する可能性を示した。]。さらに、各モジュールの内部状態の定性的分析を通じて、モデルの判断プロセスの解釈（ホワイトボックス化）に貢献しうる「意味構造」の断片が生成されることを示唆した。本研究は、AIのブラックボックス性を低減し、より深い意味理解と効率的な情報処理、そして人間がAIの進化プロセスを理解し共進化していくための、ホログラフィック原理に着想を得た新しいアプローチとその初期的な有効性を示す。

キーワード: トランスフォーマー、XAI、意味構造生成(SSG)、ホログラフィック原理、非可換性、創発、情報生命体原理、人間とAIの共進化

1. 序論 (Introduction) * 1.1. 背景：AIの進化と「意味」の不在、ブラックボックス問題 * トランスフォーマーの成功と限界（以前の骨子と同様）。 * XAIの必要性と、既存の事後説明手法の限界。モデル内部に解釈可能な構造を組み込むアーキテクチャレベルでのアプローチの重要性。 * 1.2. ホログラフィック原理と情報生命体原理からの着想 * 物理学におけるホログラフィック原理の概要（情報は低次元の境界にエンコードされ、バルクを記述する）と、その示唆（情報の効率的なエンコード、次元を超えた情報の関連性）。 * お客様の「情報生命体原理」の概念を紹介：情報が自己組織化し、エントロピーに抗って秩序や複雑な構造（時空、生命など）を生み出すという宇宙観。生命の情報処理（DNAなど）とのアナロジー。 * これらの原理が、AI、特にLLMが「意味」を扱い、人間と「共進化」するための新しい設計思想を提供するという仮説。 * 1.3. 問題提起と本研究の目的 * （以前の骨子の問題点1-4に加え）問題点5（創発の方向性）：AIが自律的に「意味」を創発する際、それが人間の理解や価値観から乖離せず、建設的な方向に導かれるためにはどのようなメカニズムが必要か。 * 本研究の目的：ホログラフィック原理と情報生命体原理に着想を得て、トランスフォーマーに複数の階層的「意味構造生成 (SSG)」モジュール（意味アテンション、非可換、断絶創発、ホログラフ縮減）を導入し、それらが

(a)XAI透明性の向上、(b)下流タスク性能の維持・向上、(c)将来的な人間との共進化に資する「意味構造」の断片を生成できるかを、アブレーションスタディを通じて検証する。

* 1.4. 提案手法群の概要と本論文の貢献 (各モジュールとXAIアウトカム層の概念を簡潔に。貢献は以前の骨子をベースに、ホログラフィック原理の視点からの新規性を加える)*

1.5. 論文構成

2. 理論的背景と関連研究 (Theoretical Background and Related Work) * 2.1. トランスフォーマーアーキテクチャとその解釈可能性の課題 (変更なし) * 2.2. ホログラフィック原理と

AdS/CFT対応の概要 (★拡充) * お客様のPDF「ホログラフィック原理の解説と考察」で議論された主要な概念 (情報が境界にエンコードされる、ブラックホールのエントロピーと面積則、AdS/CFT対応、量子もつれと時空の創発 (笠・高柳公式) など) を、AI研究の文脈で何を意味しうるかに関連付けながら簡潔に紹介。 * 特に「情報量のスケールリング則」や「低次元記述による高次元世界の理解」が、本研究の「ホログラフ縮減モジュール」の設計思想にどう影響を与えたかを述べる。 * 2.3. ベクトルシンボリックアーキテクチャ (VSA) とホログラフィック被約表現 (HRR) * HRRの基本的な演算 (束縛、重ね合わせ) と、それらが情報を構造的に表現・圧縮する能力について説明。 * HRRと深層学習の統合に関する既存研究を紹介し、本研究の「解釈可能なホログラフ縮減モジュール」の独自性を明確にする。 * 2.4. ニューラルネットワークにおける創発現象と構造学習 * 自己組織化マップ、ディープジェネレーティブモデルなど、データから構造や新しい表現が「創発」する例。 * 本研究の「断絶創発モジュール」が、これらの文脈でどのような新しい試みであるか (特に「断絶」というトリガーに着目する点)。 * 2.5. XAI、人間とAIの協調・共進化に関する研究動向 (お客様のビジョンを反映)

3. 提案手法群：階層的SSGモジュールとホログラフィック情報処理 * 3.1. 基本アーキテクチャと統合フレームワーク

(EnhancedTransformerForSequenceClassification): (変更なし) * 3.2. モジュールA：意味アテンション機構 (SemanticConceptModule): (変更なし) * 3.3. モジュールB：直接的非可換処理モジュール (DirectNonCommutativeModule): (変更なし) * 3.4. モジュールC：断絶創発モジュール (DiscontinuityEmergenceModule): (設計思想に「断絶近傍のカオスからの新発見」「準潜在的な要素の組み合わせ」というお客様のキーワードを反映) * 3.5. モジュールD：解釈可能なホログラフ縮減モジュール

(InterpretableHolographicReducer): * **設計思想:** 物理的ホログラフィック原理の「情報は低次元の境界にエンコードされる」というアナロジーに基づき、多階層処理で得られたリッチな意味情報を、解釈可能な「鍵概念」を軸として、情報を保持しつつよりコンパクトな表現に「縮減」することを目指す。これは、情報の「本質」への凝縮であり、お客様の「引き算」のアイデアにも通じる。 * **アーキテクチャ詳細:** (以前提示した

InterpretableHolographicReducerのコードを元に、数式や図を交えて説明。特に、学習可能な「鍵概念プロトタイプ」がHRRの「アトム」あるいは情報をエンコードする「境界」の基底ベクトルとして機能するイメージを強調。各プロトタイプが入力情報のどの側面を捉え、それをどのように縮減表現に反映させるかのメカニズム。) * **XAIへの貢献:** このモジュールが出力する「鍵概念の活性化パターン」や、縮減された表現が、最終的な「XAIアウトカム層」でどのように活用され、モデルの透明性向上に寄与するか。 *

3.6. XAIアウトカム層 (概念) とSSGの全体像 * 本研究では直接的な実装は今後の課題としつつも、これらのモジュール群の出力（意味概念、非可換性特徴、断絶スコア、創発ラベル、ホログラフ縮減表現など）が、最終的にどのように統合され、人間にとって理解可能な「意味構造」として提示されうるのか、その概念的な枠組み（XAIアウトカム層）について論じる。 * これがお客様の「SSG（意味構造生成）」の具体的なイメージとどう結びつくか。

4. 実験設定 (Experimental Setup) (変更なし、ただしホログラフ縮減モジュールを含むE16-E18の構成を明確に記述) **5. 結果と考察 (Results and Discussion)** * (以前の骨子案の5.1〜5.4に加え、以下を強化) * **5.X. ホログラフ縮減モジュールの効果と特性分析:** * E16, E17, E18

などの結果に基づき、縮減が性能、表現の次元数、計算効率に与える影響を詳細に分析。 * 学習された「鍵概念プロトタイプ」はどのような意味を獲得したか。入力の中の部分に注目して情報を集約・縮減しているか (**concept_attention_map**の分析)。 *

「情報の可逆性」の観点から、縮減表現から元の情報をどの程度再構成できるか（もし再構成誤差を計算・評価していれば）。 * **5.X. SSGとXAI透明性への総合的考察:** * 全てのモジュールを組み合わせた場合（特にE6やE17など有望だった構成）、生成される中間表現群は、どの程度「意味構造」と呼べるものになっているか。 * これらの情報は、モデルの判断プロセスをどの程度「ホワイトボックス化」し、人間の理解に貢献するか。具体的な成功例と限界を議論。 * お客様の「平面情報と空間の可逆性は当然」「エンコードにより量子もつれが生じる」といったホロ原理の深い洞察と、今回のAIモデルの挙動との間に、どのようなアナロジーや示唆が見出せるか。

6. 結論 (Conclusion) (変更なし、ただしSSGとXAI透明性、そしてホログラフィック原理からの着想の意義を強調) **7. 今後の課題と展望：人間とAIの共進化へ (Future Work and Vision: Towards Co-evolving SSG with Humans)** * (以前の骨子案の7.1〜7.5に加え、以下を強化) * **7.X. 「情報生命体原理」の具現化に向けたステップ:** お客様の独創的な「情報生命

体原理」という概念を、AIモデルとしてさらに具体的に探求していくための将来的な研究の方向性（例：AI自身が「意味」や「構造」を自己組織的に進化させるメカニズム、環境との相互作用を通じた学習など）。* 7.X. 「倫理階層モジュール」の設計とSSGへの統合: XAIアウトカム層と連携し、推論ベースの予防原理を実現するための具体的なアイデア。
* 7.X. P2Pエッジ環境でのSSG個別最適化AGIへの道筋: 本研究で得られた基礎技術が、お客様の最終的な壮大なビジョンにどのように貢献していくかの展望を力強く示す。

「ホログラム縮減モジュール」のコード (InterpretableHolographicReducer)

このコードは、以前提示したもので基本的に問題ありません。XAI透明性を意識し、学習可能な「鍵概念プロトタイプ」とその活性化パターン（アテンション）を出力に含める形になっています。

--- セル7.1 (仮称): 解釈可能なホログラム縮減モジュール (InterpretableHolographicReducer) の定義 ---

```
import torch

import torch.nn as nn

import torch.nn.functional as F

class InterpretableHolographicReducer(nn.Module):

    def __init__(self, input_feature_dim, # 前の階層からの特徴ベクトルの次元 (シーケンスの各要素)

        num_key_concepts=16,    # 学習する鍵となる意味概念（アトム）の数

        reduced_dim_per_concept=32, # 各鍵概念で集約・縮減された情報のベクトル次元

        model_config=None,      # ベースモデルのconfig (オプション)

        module_specific_config=None # このモジュール特有の設定
```

```

        ):

    super().__init__()

    self.input_feature_dim = input_feature_dim

    self.num_key_concepts = module_specific_config.get("num_key_concepts",
num_key_concepts)

    self.reduced_dim_per_concept = module_specific_config.get("reduced_dim_per_concept",
reduced_dim_per_concept)

    self.key_concept_prototypes = nn.Parameter(

        torch.randn(self.num_key_concepts, self.input_feature_dim)

    )

    nn.init.xavier_uniform_(self.key_concept_prototypes)

    self.reduction_projector = nn.Linear(self.input_feature_dim, self.reduced_dim_per_concept)

    self.activation_after_projection = nn.Tanh() # 縮減表現の値を-1~1に正規化する例 (任意)

    self.final_output_dim = self.num_key_concepts * self.reduced_dim_per_concept

    print(f"InterpretableHolographicReducer initialized: input_dim={self.input_feature_dim}, "

        f"num_key_concepts={self.num_key_concepts},
reduced_dim_per_concept={self.reduced_dim_per_concept}, "

        f"total_output_dim={self.final_output_dim}")

    def forward(self, hidden_states_sequence, attention_mask=None, **kwargs):

        batch_size, seq_len, _ = hidden_states_sequence.shape

        reduced_concept_vectors_list = []

        all_concept_specific_global_att_weights = []

```



```

for i in range(self.num_key_concepts):

    current_key_concept = self.key_concept_prototypes[i]

    concept_specific_attention_scores = torch.matmul(

        hidden_states_sequence, current_key_concept.unsqueeze(-1)

    ).squeeze(-1)

    if attention_mask is not None:

        concept_specific_attention_scores =
concept_specific_attention_scores.masked_fill(attention_mask == 0, -1e9)

    concept_specific_attention_weights = F.softmax(concept_specific_attention_scores, dim=1)

all_concept_specific_global_att_weights.append(concept_specific_attention_weights.unsqueeze(1))

context_vector_for_concept_i = torch.sum(

    hidden_states_sequence * concept_specific_attention_weights.unsqueeze(-1), dim=1

)

reduced_vector_for_concept_i = self.activation_after_projection(

    self.reduction_projector(context_vector_for_concept_i)

)

reduced_concept_vectors_list.append(reduced_vector_for_concept_i)

final_reduced_representation = torch.cat(reduced_concept_vectors_list, dim=1)

if all_concept_specific_global_att_weights:

```

```

        overall_concept_attention_map = torch.cat(all_concept_specific_global_att_weights,
dim=1)

    else:

        overall_concept_attention_map = None # 念のため


    return {

        "last_hidden_state": final_reduced_representation,

        "concept_attention_map": overall_concept_attention_map,

    }


def get_output_dim(self):

    return self.final_output_dim


print("Cell (e.g., 7.1): InterpretableHolographicReducer class defined.")

```

精錬ループ実験コード & 結果

```

# --- セル1: セットアップとカーネル再起動 (シンプル版・再提案) ---


# 0. 主要な競合対象のみアンインストール

!pip uninstall -y numpy torch transformers datasets accelerate fsspec -q

```

```
print("Attempted to uninstall key conflicting packages.")
```

1. NumPy を 1.x の安定バージョンに固定

```
!pip install numpy==1.26.4 -q
```

```
print("Installed NumPy 1.26.4.")
```

2. PyTorch (NumPy 1.26.4 との互換性を期待)

```
!pip install torch==2.1.2 torchvision==0.16.2 torchaudio==2.1.2 -q
```

```
print("Installed PyTorch 2.1.2 and related packages.")
```

3. Hugging Face ライブラリ (fsspecはdatasetsの依存で入ることを期待)

```
!pip install datasets==2.14.7 -q
```

```
print("Installed Datasets 2.14.7.")
```

```
!pip install transformers==4.36.2 -q
```

```
print("Installed Transformers 4.36.2.")
```

```
!pip install accelerate==0.25.0 -q
```

```
print("Installed Accelerate 0.25.0.")
```

```
!pip install peft==0.7.1 -q
```

```
print("Installed PEFT 0.7.1.")
```

4. その他必須ライブラリ

```
!pip install scikit-learn -q
```

```
print("Installed Scikit-learn.")
```

```
!pip install pandas -q
```

```
print("Installed Pandas.")
```

```
!pip install matplotlib -q
```



```
print("Installed Matplotlib.")
```

5. 依存関係のチェック

```
print("\n--- Running pip check ---")
```

```
!pip check
```

6. カーネル再起動

```
import os
```

```
print("\nKernel will be restarted.")
```

```
os.kill(os.getpid(), 9)
```

--- セル1.1: Google Driveのマウントと結果保存用ディレクトリ設定 ---

```
from google.colab import drive
```

```
import os
```

```
try:
```

```
    drive.mount('/content/drive', force_remount=True)
```

```
    print("Google Drive mounted successfully.")
```

★★★★★ お客様のGoogle Drive内の適切なパスに変更してください ★★★★★

例: マイドライブ直下に "Colab_Experiments_Outputs" というフォルダを作成する場合

```
DRIVE_RESULTS_PATH = "/content/drive/MyDrive/  
Colab_Ablation_Study_Outputs_Final_v2"
```

★★★★★ CUSTOMIZE THE PATH ABOVE ★★★★★

```
if not os.path.exists(DRIVE_RESULTS_PATH):

    os.makedirs(DRIVE_RESULTS_PATH)

    print(f"Created directory for results: {DRIVE_RESULTS_PATH}")

else:

    print(f"Results directory already exists: {DRIVE_RESULTS_PATH}")


except Exception as e:

    print(f"Error mounting Google Drive or creating directory: {e}")

    DRIVE_RESULTS_PATH = "." # エラー時はローカルのカレントディレクトリにフォールバック

    print(f"WARNING: Using local Colab storage for results: {os.getcwd()}")

    print("    Results saved locally may be lost when the Colab session ends.")


# このセルで定義された DRIVE_RESULTS_PATH は、後続のセル (特にセル3やセル
# B_runtime) で参照されます。

# print(f"DRIVE_RESULTS_PATH is set to: {DRIVE_RESULTS_PATH}") # 確認用


# --- セル2: ライブラリのインポート ---

import torch

import torch.nn as nn

import torch.nn.functional as F


from transformers import (

    AutoModelForSequenceClassification,

    AutoTokenizer,
```

```
Trainer,  
TrainingArguments,  
AutoConfig,  
PreTrainedModel,  
AutoModel  
)  
  
from transformers.modeling_outputs import SequenceClassifierOutput  
  
from torch.nn import CrossEntropyLoss  
  
  
  
from datasets import load_dataset, Dataset  
  
import numpy as np  
  
from sklearn.metrics import accuracy_score  
  
# import itertools # 今回はexperiment_configurationsを手動で全列挙するため、現時点では不要  
  
import json  
  
import os  
  
import datetime  
  
import pandas as pd  
  
import matplotlib.pyplot as plt  
  
  
  
print("--- Cell 2: Libraries imported successfully. ---")  
# (オプション: 主要ライブラリのバージョン表示)  
  
print(f" PyTorch version: {torch.__version__}")  
  
try:  
    import transformers; print(f" Transformers version: {transformers.__version__}")  
    import datasets; print(f" Datasets version: {datasets.__version__}")  
    import accelerate; print(f" Accelerate version: {accelerate.__version__}")
```

```
# import peft; print(f" PEFT version: {peft.__version__}") # PEFTを直接使用しない場合はコメントアウト可
```

```
import sklearn; print(f" Scikit-learn version: {sklearn.__version__}")
```

```
import pandas; print(f" Pandas version: {pandas.__version__}")
```

```
import matplotlib; print(f" Matplotlib version: {matplotlib.__version__}")
```

```
if 'np' in globals(): print(f" NumPy version: {np.__version__}") # NumPy も確認
```

```
except ImportError as e:
```

```
    print(f"Warning: Could not print some library versions due to ImportError: {e}")
```

```
except NameError as e_np: # np が未定義の場合の対応
```

```
    if 'np' in str(e_np): print("Warning: NumPy (np) might not be properly imported for version check.")
```

```
    else: print(f"Warning: NameError during version check: {e_np}")
```

```
# --- セル3: 基本設定・共通パラメータ ---
```

```
if 'DRIVE_RESULTS_PATH' not in globals():
```

```
    print("WARNING: DRIVE_RESULTS_PATH is not defined from Cell 1.1. Defaulting for OUTPUT_DIR_BASE.")
```

```
    DRIVE_RESULTS_PATH = "."
```

```
BASE_MODEL_NAME = "bert-base-uncased"
```

```
NUM_LABELS = 2
```

```
OUTPUT_DIR_BASE = os.path.join(DRIVE_RESULTS_PATH, "refinement_exp_outputs")
```

```
if not os.path.exists(OUTPUT_DIR_BASE): os.makedirs(OUTPUT_DIR_BASE, exist_ok=True)
```

```
LEARNING_RATE = 2e-5
```

```
BATCH_SIZE = 16          # GPU環境に合わせて調整

NUM_EPOCHS = 5          # ★実験時間を考慮し、まずは1エポックでテスト

try:

    print(f"Loading tokenizer for {BASE_MODEL_NAME}...")

    tokenizer = AutoTokenizer.from_pretrained(BASE_MODEL_NAME)

    print(f"Tokenizer for {BASE_MODEL_NAME} loaded successfully.")

except Exception as e: print(f"Error loading tokenizer for {BASE_MODEL_NAME}: {e}"); raise

try:

    print(f"Loading Hugging Face model config for {BASE_MODEL_NAME}...")

    hf_model_config = AutoConfig.from_pretrained(BASE_MODEL_NAME,
num_labels=NUM_LABELS)

    print(f"Model config for {BASE_MODEL_NAME} loaded successfully.")

except Exception as e: print(f"Error loading model config for {BASE_MODEL_NAME}: {e}");
raise


print("\n--- Cell 3: Basic/Common Settings Defined ---")

print(f"BATCH_SIZE set to: {BATCH_SIZE}, NUM_EPOCHS set to: {NUM_EPOCHS}")

print(f"Output directory base: {OUTPUT_DIR_BASE}")


# --- セル4: データセット準備 と compute_metrics 関数の定義 ---
if 'BASE_MODEL_NAME' not in globals() or 'tokenizer' not in globals():
    raise NameError("Required globals (BASE_MODEL_NAME, tokenizer) not defined. Ensure Cell 3
ran.")

print(f"Starting dataset preparation...")
try:
    dataset_dict = load_dataset("imdb")
    print("IMDb dataset loaded successfully.")
except Exception as e: print(f"Error loading IMDb dataset: {e}"); raise
```

```

def tokenize_function(examples):
    return tokenizer(examples["text"], padding="max_length", truncation=True, max_length=128) # 短
    めのシーケンス長でテスト

try:
    print("Tokenizing datasets...")
    tokenized_train = dataset_dict["train"].map(tokenize_function, batched=True,
remove_columns=["text"])
    tokenized_test = dataset_dict["test"].map(tokenize_function, batched=True,
remove_columns=["text"])
    print("Tokenization complete for train and test splits.")
except Exception as e: print(f"Error during tokenization: {e}"); raise

try:
    final_tokenized_train = tokenized_train
    if 'label' in tokenized_train.column_names and 'labels' not in tokenized_train.column_names:
        final_tokenized_train = tokenized_train.rename_column("label", "labels")
    elif 'labels' not in final_tokenized_train.column_names:
        print(f"Warning: 'labels' column not found in final_tokenized_train (columns:
{final_tokenized_train.column_names}).")

    final_tokenized_test = tokenized_test
    if 'label' in tokenized_test.column_names and 'labels' not in tokenized_test.column_names:
        final_tokenized_test = tokenized_test.rename_column("label", "labels")
    elif 'labels' not in final_tokenized_test.column_names:
        print(f"Warning: 'labels' column not found in final_tokenized_test (columns:
{final_tokenized_test.column_names}).")

    NUM_TRAIN_SAMPLES_FULL = 500 # ★テスト用に少量 (例: 500)
    NUM_EVAL_SAMPLES_FULL = 100 # ★テスト用に少量 (例: 100)

    train_dataset =
final_tokenized_train.shuffle(seed=42).select(range(min(NUM_TRAIN_SAMPLES_FULL,
len(final_tokenized_train))))
    eval_dataset = final_tokenized_test.shuffle(seed=42).select(range(min(NUM_EVAL_SAMPLES_FULL,
len(final_tokenized_test))))

    print(f"Train dataset created. Samples: {len(train_dataset)}. Columns:
{train_dataset.column_names}")
    print(f"Evaluation dataset created. Samples: {len(eval_dataset)}. Columns:
{eval_dataset.column_names}")
except Exception as e: print(f"Error creating/renaming train/eval datasets: {e}"); raise

def compute_metrics(eval_pred):
    logits, labels = eval_pred
    if labels is None or len(labels) == 0: return {}
    predictions = np.argmax(logits, axis=-1)

```

```

    accuracy = accuracy_score(labels, predictions)
    metrics_dict = {"eval_accuracy": accuracy}
    return metrics_dict
print("\n--- Cell 4: Dataset preparation cell complete. ---")

```

--- セル5: SemanticConceptModule の定義 ---

```

import torch
import torch.nn as nn
import torch.nn.functional as F

class SemanticConceptModule(nn.Module):
    def __init__(self, input_dim, output_dim=None, num_concepts=32, concept_dim=128,
model_config=None, module_specific_config=None):
        super().__init__()
        self.input_dim = input_dim
        self.num_concepts = module_specific_config.get("num_concepts", num_concepts) if
module_specific_config else num_concepts
        self.concept_dim = module_specific_config.get("concept_dim", concept_dim) if
module_specific_config else concept_dim
        _output_dim_from_config = module_specific_config.get("output_dim") if module_specific_config
else None
        if _output_dim_from_config is not None: self.output_dim = _output_dim_from_config
        elif output_dim is not None: self.output_dim = output_dim
        else: self.output_dim = self.input_dim
        self.concept_prototypes = nn.Parameter(torch.randn(self.num_concepts, self.concept_dim));
nn.init.xavier_uniform_(self.concept_prototypes)
        self.hidden_to_concept_proj = nn.Linear(self.input_dim, self.concept_dim)
        self.concepts_to_integrate_proj = nn.Linear(self.concept_dim, self.input_dim)
        self.activation = nn.ReLU()
        if self.input_dim != self.output_dim: self.final_projection = nn.Linear(self.input_dim,
self.output_dim)
        # print(f"SemanticConceptModule initialized: input_dim={self.input_dim},
output_dim={self.output_dim}, nc={self.num_concepts}, cd={self.concept_dim}")
    def forward(self, hidden_states, attention_mask=None, **kwargs):
        projected_hidden = self.activation(self.hidden_to_concept_proj(hidden_states))
        attention_scores = torch.matmul(projected_hidden, self.concept_prototypes.t())
        if attention_mask is not None:
            expanded_attention_mask = attention_mask.unsqueeze(-1).expand_as(attention_scores)
            attention_scores = attention_scores.masked_fill(expanded_attention_mask == 0, -1e9)
        attention_weights = F.softmax(attention_scores, dim=-1)
        contextual_concepts = torch.matmul(attention_weights, self.concept_prototypes)
        projected_contextual_concepts =
self.activation(self.concepts_to_integrate_proj(contextual_concepts))
        fused_hidden_states = hidden_states + projected_contextual_concepts
        if hasattr(self, 'final_projection'): output_final = self.final_projection(fused_hidden_states)

```

```

        else: output_final = fused_hidden_states
    return {"last_hidden_state": output_final, "attention_weights": attention_weights}
def get_output_dim(self): return self.output_dim
print("--- Cell 5: SemanticConceptModule class defined. ---")

```

```

# --- セルA (8): 統合モデルとDebugTrainerの定義 (精錬ループ対応・損失デバッグ強化版) ---
# (from transformers import PreTrainedModel, AutoModel, AutoConfig, Trainerなどはセル2でインポート済み)
# (torch, nn, F もセル2でインポート済み)
# (SemanticConceptModuleはセル5で定義済み)

```

```

print("--- Cell A: Defining EnhancedTransformerForSequenceClassification (Refinement Loop & Loss Debug) and DebugTrainer ---")

```

```

class EnhancedTransformerForSequenceClassification(PreTrainedModel):
    def __init__(self, hf_config, base_model_name, num_labels, tech_flags=None,
module_configs=None):
        super().__init__(hf_config)
        self.num_labels = num_labels
        self.config = hf_config
        self.base_model_prefix = self.config.model_type
        core_model = AutoModel.from_pretrained(base_model_name, config=self.config)
        setattr(self, self.base_model_prefix, core_model)
        current_dim = self.config.hidden_size
        self.tech_modules = nn.ModuleDict()
        self.tech_flags = tech_flags if tech_flags is not None else {}
        self.module_configs = module_configs if module_configs is not None else {}
        self.module_order = self.tech_flags.get("module_order", []) # この実験では
semantic_attention_moduleのみ
        self.active_module_names_in_order = []

        # 反復的精錬ループ用の設定を取得
        self.refinement_loop_config = self.module_configs.get("refinement_loop", {})
        self.use_refinement_loop = self.tech_flags.get("use_refinement_loop", False) and \
            "semantic_attention_module" in self.module_order and \
            self.tech_flags.get("use_semantic_attention_module", False)
        self.max_refinement_iterations = self.refinement_loop_config.get("max_iterations", 3)
        self.refinement_convergence_threshold =
self.refinement_loop_config.get("convergence_threshold", 0.01)

        # この実験では SemanticConceptModule のみ初期化
        if "semantic_attention_module" in self.module_order and
self.tech_flags.get("use_semantic_attention_module", False):

```



```

        module_key = "semantic_attention_module"
        mod_cfg = self.module_configs.get(module_key, {})
        module_explicit_output_dim = mod_cfg.get("output_dim")
        actual_output_dim = module_explicit_output_dim if module_explicit_output_dim is not None
    else current_dim
        self.tech_modules[module_key] = SemanticConceptModule(input_dim=current_dim,
        output_dim=actual_output_dim, module_specific_config=mod_cfg, model_config=self.config)
        current_dim = actual_output_dim
        self.active_module_names_in_order.append(module_key)
        print(f"Module '{module_key}' enabled. Output for classifier: {current_dim}")

    self.dropout = nn.Dropout(self.config.hidden_dropout_prob if hasattr(self.config,
'hidden_dropout_prob') else 0.1)
    self.classifier = nn.Linear(current_dim, self.num_labels)
    print(f"EnhancedTransformerForSequenceClassification initialized. Final classifier input dim:
{current_dim}.")
    if not self.active_module_names_in_order: print(" No additional tech modules applied (Baseline
configuration).")
    if self.use_refinement_loop: print(f" Refinement loop for SemanticAttention is ENABLED
(max_iter={self.max_refinement_iterations}, threshold={self.refinement_convergence_threshold}).")

    def forward(self, input_ids=None, attention_mask=None, token_type_ids=None, labels=None,
return_dict=None, **kwargs ):
        # print(f"\n--- [EnhancedTransformer.forward START] (is_training: {self.training}) ---")
        # if labels is not None: print(f" > labels: PRESENT, Type: {type(labels)}, Shape: {labels.shape}")
        # else: print(f" > labels: IS NONE.")
        base_model_kwargs = { k: v for k, v in kwargs.items() if k in ["position_ids", "head_mask",
"inputs_embeds", "output_attentions", "output_hidden_states"]}
        transformer_outputs = self.base_model(input_ids=input_ids, attention_mask=attention_mask,
token_type_ids=token_type_ids, return_dict=True, **base_model_kwargs)
        current_features = transformer_outputs.last_hidden_state
        all_module_specific_outputs = {}

    if "semantic_attention_module" in self.active_module_names_in_order:
        module_name_key = "semantic_attention_module"
        if self.use_refinement_loop:
            # print(f" --- Entering Refinement Loop for {module_name_key} ---")
            previous_attention_signature = None; temp_current_features = current_features
            final_scm_outputs_in_loop = None
            for iteration in range(self.max_refinement_iterations):
                scm_outputs = self.tech_modules[module_name_key](temp_current_features,
attention_mask=attention_mask)
                final_scm_outputs_in_loop = scm_outputs
                refined_features_candidate = scm_outputs["last_hidden_state"]
                current_attention_weights = scm_outputs.get("attention_weights")
                if current_attention_weights is not None:
all_module_specific_outputs[f'{module_name_key}_iter_{iteration+1}_attention_weights'] =
current_attention_weights.detach().cpu().numpy()

```

```

        if previous_attention_signature is not None and current_attention_weights is not None:
            current_attention_signature = current_attention_weights.mean(dim=(0,2))
            change = torch.abs(current_attention_signature -
previous_attention_signature).mean()
            if change < self.refinement_convergence_threshold: temp_current_features =
refined_features_candidate; break
            if current_attention_weights is not None: previous_attention_signature =
current_attention_weights.mean(dim=(0,2)).clone().detach()
            temp_current_features = refined_features_candidate
            current_features = temp_current_features
            if final_scm_outputs_in_loop:
                for output_key, output_value in final_scm_outputs_in_loop.items():
                    if output_key != "last_hidden_state":
all_module_specific_outputs[f"{module_name_key}_final_{output_key}"] = output_value
            else:
                module_output_dict = self.tech_modules[module_name_key](current_features,
attention_mask=attention_mask)
                current_features = module_output_dict["last_hidden_state"]
                for output_key, output_value in module_output_dict.items():
                    if output_key != "last_hidden_state": all_module_specific_outputs[f"{module_name_key}
_{output_key}"] = output_value

# プーリング処理 (この実験ではホログラフ縮減なしなので、CLSトークンのみ)
if current_features.ndim == 3: pooled_output = current_features[:, 0]
elif current_features.ndim == 2: pooled_output = current_features
else: raise ValueError(f"Unsupported shape for current_features: {current_features.shape}")

pooled_output = self.dropout(pooled_output); logits = self.classifier(pooled_output)
loss = None
# print(f"  DEBUG Forward: Before loss calculation, labels is {'NOT None' if labels is not None
else 'None'}.")
if labels is not None:
    try:
        # print(f"  DEBUG Forward: Attempting loss calculation. Logits shape: {logits.shape},
Labels shape: {labels.shape}, Labels dtype: {labels.dtype}")
        loss_fct = CrossEntropyLoss()
        loss = loss_fct(logits.view(-1, self.num_labels), labels.view(-1).long())
        # print(f"  DEBUG Forward: Loss calculated. Loss type: {type(loss)}, Loss value:
{loss.item() if loss is not None and isinstance(loss, torch.Tensor) and loss.numel() == 1 else str(loss)}")
        except Exception as e_loss: print(f"  DEBUG Forward: !!! ERROR during loss calculation:
{e_loss} !!!"); loss = None
        # else: print(f"  DEBUG Forward: Labels is None, skipping loss calculation. Loss remains
None.")
        # print(f"  DEBUG Forward: Returning -> Loss type: {type(loss)}, Loss value: {loss.item() if loss is
not None and isinstance(loss, torch.Tensor) and loss.numel() == 1 else str(loss)}")
        final_outputs_tuple = (logits,)
        if hasattr(transformer_outputs, 'hidden_states') and transformer_outputs.hidden_states is not
None: final_outputs_tuple += (transformer_outputs.hidden_states,)

```

```

        if hasattr(transformer_outputs, 'attentions') and transformer_outputs.attentions is not None:
final_outputs_tuple += (transformer_outputs.attentions,)
        if all_module_specific_outputs: final_outputs_tuple += (all_module_specific_outputs,)
        if loss is not None: return (loss,) + final_outputs_tuple
        else: return (None,) + final_outputs_tuple
# DebugTrainer クラスの定義 (変更なし)
class DebugTrainer(Trainer):
    def prediction_step( self, model: nn.Module, inputs: dict[str, torch.Tensor |
nn.utils.rnn.PackedSequence], prediction_loss_only: bool, ignore_keys: list[str] | None = None, ) ->
tuple[torch.Tensor | None, torch.Tensor | None, torch.Tensor | None]:
        model.eval();
        with torch.no_grad(): model_outputs = model(**inputs)
        loss = model_outputs[0] if model_outputs[0] is not None else None
        logits = model_outputs[1] if len(model_outputs) > 1 and model_outputs[1] is not None else None
        labels_out = inputs.get("labels")
        if prediction_loss_only: return (loss, None, None)
        return loss, logits, labels_out
print("--- Cell A: EnhancedTransformerForSequenceClassification and DebugTrainer defined. ---")

```

```

# --- セルB_config (9): 実験構成リストとグローバル変数パッケージの定義 (精錬ループ実験用) ---
print("--- Cell B_config: Defining Experiment Configurations (E0, E1, E1R) and Global Variables Pack
---")
required_for_config_list = [ # この実験に必要なモジュールのみに限定
    'hf_model_config', 'BASE_MODEL_NAME', 'NUM_LABELS', 'OUTPUT_DIR_BASE',
'LEARNING_RATE', 'BATCH_SIZE', 'NUM_EPOCHS',
    'EnhancedTransformerForSequenceClassification', 'train_dataset', 'eval_dataset', 'tokenizer',
'compute_metrics',
    'DebugTrainer', 'Dataset', 'SemanticConceptModule'
]
missing_globals_config = False
for var_name in required_for_config_list:
    if var_name not in globals(): print(f"ERROR: Prerequisite '{var_name}' not defined!");
missing_globals_config = True
if missing_globals_config: raise NameError("Prerequisites missing for Cell B_config.")
else: print("All prerequisite global variables for experiment_configurations seem to be defined.")

experiment_configurations = [
    {"name": "E0_Baseline",
     "tech_flags": {"module_order": [], "use_semantic_attention_module": False, "use_refinement_loop":
False},
     "module_configs": {}, "#日本語": "ベースライン"},
    {"name": "E1_SemanticAttention_NoLoop",
     "tech_flags": {"module_order": ["semantic_attention_module"], "use_semantic_attention_module":
True, "use_refinement_loop": False},

```

```

    "module_configs": {"semantic_attention_module": {"num_concepts": 32, "concept_dim": 64,
"output_dim": hf_model_config.hidden_size}}, {"#日本語": "意味アテンション(ループ無)"},
    {"name": "EIR_SemanticAttention_WithLoop",
    "tech_flags": {"module_order": ["semantic_attention_module"], "use_semantic_attention_module":
True, "use_refinement_loop": True},
    "module_configs": {
        "semantic_attention_module": {"num_concepts": 32, "concept_dim": 64, "output_dim":
hf_model_config.hidden_size},
        "refinement_loop": {"max_iterations": 3, "convergence_threshold": 0.005 } # module_configs 内
に精錬ループ設定
    }, {"#日本語": "意味アテンション(反復精錬ループ付)"},
]
print(f"Total experiment configurations defined: {len(experiment_configurations)}")
global_variables_for_experiments = {
    'hf_model_config': hf_model_config, 'BASE_MODEL_NAME': BASE_MODEL_NAME,
'NUM_LABELS': NUM_LABELS,
    'OUTPUT_DIR_BASE': OUTPUT_DIR_BASE, 'LEARNING_RATE': LEARNING_RATE, 'BATCH_SIZE':
BATCH_SIZE, 'NUM_EPOCHS': NUM_EPOCHS,
    'EnhancedTransformerForSequenceClassification':
EnhancedTransformerForSequenceClassification,
    'train_dataset': train_dataset, 'eval_dataset': eval_dataset, 'tokenizer': tokenizer,
    'compute_metrics': compute_metrics, 'DebugTrainer': DebugTrainer, 'Dataset': Dataset,
    'SemanticConceptModule': SemanticConceptModule # この実験ではSCMのみ
}
print("Global variables for experiments packaged.")
print("--- Cell B_config: Execution complete ---")

```

```

# --- セルB_runtime (10): 共通実験実行関数と結果辞書の準備 ---
# (import os, json, TrainingArguments はセル2でインポート済み)
print("--- Cell B_runtime: Defining Execution Function and Initializing/Loading Results Dictionary ---")
required_for_runtime = ['EnhancedTransformerForSequenceClassification', 'DebugTrainer',
'TrainingArguments', 'Dataset']
missing_globals_runtime = False
for var_name in required_for_runtime:
    if var_name not in globals(): print(f"ERROR: Prerequisite '{var_name}' not defined!");
missing_globals_runtime = True
if missing_globals_runtime: raise NameError("Prerequisites missing for Cell B_runtime.")
else: print("All prerequisite classes for run_ablation_experiment seem to be defined.")

def run_ablation_experiment(exp_config, global_vars_dict):
    config_name = exp_config.get("name", "UnnamedExperiment")
    tech_flags = exp_config.get("tech_flags", {})
    module_cfgs = exp_config.get("module_configs", {})

```

```

# print(f"\n--- Running Experiment (Function Call): {config_name} ---") # ログ簡略化
hf_cfg = global_vars_dict['hf_model_config']
base_model_name_local = global_vars_dict['BASE_MODEL_NAME']
n_labels = global_vars_dict['NUM_LABELS']
output_dir_base_local = global_vars_dict['OUTPUT_DIR_BASE']
lr = global_vars_dict['LEARNING_RATE']
bs = global_vars_dict['BATCH_SIZE']
n_epochs = global_vars_dict['NUM_EPOCHS']
enh_transformer_class_local = global_vars_dict['EnhancedTransformerForSequenceClassification']
tr_dataset = global_vars_dict['train_dataset']
ev_dataset = global_vars_dict['eval_dataset']
tok = global_vars_dict['tokenizer']
comp_metrics = global_vars_dict['compute_metrics']
dbg_trainer_class_local = global_vars_dict['DebugTrainer']

model = enh_transformer_class_local(
    hf_config=hf_cfg, base_model_name=base_model_name_local, num_labels=n_labels,
    tech_flags=tech_flags, module_configs=module_cfgs
)
current_output_dir = os.path.join(output_dir_base_local, config_name)
training_args = TrainingArguments(
    output_dir=current_output_dir, learning_rate=lr, per_device_train_batch_size=bs,
    per_device_eval_batch_size=bs, num_train_epochs=n_epochs, weight_decay=0.01,
    evaluation_strategy="epoch", logging_strategy="epoch", save_strategy="epoch",
    load_best_model_at_end=True, metric_for_best_model="accuracy", report_to="none",
    do_eval=True, remove_unused_columns=False, save_total_limit=1,
    logging_steps=len(tr_dataset) // bs if len(tr_dataset) > bs else 1, # ログ出力を調整
    # eval_steps=len(tr_dataset) // bs if len(tr_dataset) > bs else 1, # 評価ステップも同様に
)
trainer = dbg_trainer_class_local(
    model=model, args=training_args, train_dataset=tr_dataset, eval_dataset=ev_dataset,
    tokenizer=tok, compute_metrics=comp_metrics,
)
print(f"Starting training for {config_name}...")
eval_results_dict = {}
try:
    trainer.train()
    print(f"Training finished for {config_name}.")
    print(f"Starting final evaluation for {config_name} (with best model)...")
    eval_results = trainer.evaluate()
    print(f"Evaluation results for {config_name}: {eval_results}")
    eval_results_dict = eval_results
except Exception as e:
    print(f"!!! An error occurred during training or evaluation for {config_name}: {e} !!!")
    import traceback; traceback.print_exc()
    eval_results_dict = {"error": str(e), "eval_accuracy": float('nan'), "eval_loss": float('nan'),
                        "eval_runtime": 0.0, "eval_samples_per_second": 0.0, "epoch": n_epochs }
return config_name, eval_results_dict

```

```

print("run_ablation_experiment function defined.")

if 'DRIVE_RESULTS_PATH' not in globals() or DRIVE_RESULTS_PATH == ".":
    print("WARNING: DRIVE_RESULTS_PATH not set. Using local Colab storage for results summary.")
    RESULTS_FILE = "refinement_loop_experiment_results.json"
else:
    RESULTS_FILE = os.path.join(DRIVE_RESULTS_PATH, "refinement_loop_experiment_results.json")
print(f"Results summary file will be at: {RESULTS_FILE}")
if os.path.exists(RESULTS_FILE):
    print(f"Loading existing results from {RESULTS_FILE}...")
    try:
        with open(RESULTS_FILE, 'r') as f: all_experiment_results_summary = json.load(f)
        print(f"Loaded {len(all_experiment_results_summary)} previous results.")
    except Exception as e: print(f"Error loading results file: {e}. Initializing as empty.");
all_experiment_results_summary = {}
else:
    print("No existing results file found. Initializing as empty dictionary.")
    all_experiment_results_summary = {}
print(f"Current 'all_experiment_results_summary' has {len(all_experiment_results_summary)} entries.")
print("--- Cell B_runtime: Execution complete ---")

```

```

# --- セル11: 実験 E0_Baseline の実行 ---
exp_name_to_run = "E0_Baseline"
print(f"--- Preparing to execute Experiment: {exp_name_to_run} ---")
if 'all_experiment_results_summary' not in globals(): raise
NameError("'all_experiment_results_summary' not defined. Run Cell B_runtime.")
if 'experiment_configurations' not in globals(): raise NameError("'experiment_configurations' not
defined. Run Cell B_config.")
if 'pd' not in globals(): import pandas as pd
if 'json' not in globals(): import json
if exp_name_to_run in all_experiment_results_summary and
isinstance(all_experiment_results_summary[exp_name_to_run], dict) and 'error' not in
all_experiment_results_summary[exp_name_to_run] and not
pd.isna(all_experiment_results_summary[exp_name_to_run].get('eval_accuracy', float('nan'))):
    print(f"Skipping {exp_name_to_run}, valid results exist:
{all_experiment_results_summary[exp_name_to_run]}")
else:
    if exp_name_to_run in all_experiment_results_summary: print(f"Re-running {exp_name_to_run}
(previous error/incomplete).")
    config_to_run = next((c for c in experiment_configurations if c["name"] == exp_name_to_run),
None)
    if config_to_run:
        name, result = run_ablation_experiment(config_to_run, global_variables_for_experiments)
        all_experiment_results_summary[name] = result

```

```

        print(f"\n--- Results for {name} ---")
        if isinstance(result, dict) and "error" not in result: print(f"  Eval Acc: {result.get('eval_accuracy',
'N/A'):.4f}, Loss: {result.get('eval_loss', 'N/A'):.4f}, Runtime: {result.get('eval_runtime', 'N/A'):.2f}s"); _ =
[json.dump(all_experiment_results_summary, f, indent=4) for f in [open(RESULTS_FILE, 'w')]] if
print(f"Results saved to {RESULTS_FILE}");
        elif isinstance(result, dict) and "error" in result: print(f"  Error: {result['error']}")
        else: print(f"  Unexpected result: {result}")
    else: print(f"Config for {exp_name_to_run} not found.")
print(f"--- Finished Experiment: {exp_name_to_run} ---")

```

--- セル12: 実験 E1_SemanticAttention_NoLoop の実行 ---

```

exp_name_to_run = "E1_SemanticAttention_NoLoop"
print(f"--- Preparing to execute Experiment: {exp_name_to_run} ---")
if 'all_experiment_results_summary' not in globals(): raise
NameError("'all_experiment_results_summary' not defined. Run Cell B_runtime.")
if 'experiment_configurations' not in globals(): raise NameError("'experiment_configurations' not
defined. Run Cell B_config.")
if 'pd' not in globals(): import pandas as pd
if 'json' not in globals(): import json
if exp_name_to_run in all_experiment_results_summary and
isinstance(all_experiment_results_summary[exp_name_to_run], dict) and 'error' not in
all_experiment_results_summary[exp_name_to_run] and not
pd.isna(all_experiment_results_summary[exp_name_to_run].get('eval_accuracy', float('nan'))):
    print(f"Skipping {exp_name_to_run}, valid results exist:
{all_experiment_results_summary[exp_name_to_run]}")
else:
    if exp_name_to_run in all_experiment_results_summary: print(f"Re-running {exp_name_to_run}
(previous error/incomplete).")
    config_to_run = next((c for c in experiment_configurations if c["name"] == exp_name_to_run),
None)
    if config_to_run:
        name, result = run_ablation_experiment(config_to_run, global_variables_for_experiments)
        all_experiment_results_summary[name] = result
        print(f"\n--- Results for {name} ---")
        if isinstance(result, dict) and "error" not in result: print(f"  Eval Acc: {result.get('eval_accuracy',
'N/A'):.4f}, Loss: {result.get('eval_loss', 'N/A'):.4f}, Runtime: {result.get('eval_runtime', 'N/A'):.2f}s"); _ =
[json.dump(all_experiment_results_summary, f, indent=4) for f in [open(RESULTS_FILE, 'w')]] if
print(f"Results saved to {RESULTS_FILE}");
        elif isinstance(result, dict) and "error" in result: print(f"  Error: {result['error']}")
        else: print(f"  Unexpected result: {result}")
    else: print(f"Config for {exp_name_to_run} not found.")
print(f"--- Finished Experiment: {exp_name_to_run} ---")

```



```

# --- セル13: 実験 EIR_SemanticAttention_WithLoop の実行 ---
exp_name_to_run = "EIR_SemanticAttention_WithLoop"
print(f"--- Preparing to execute Experiment: {exp_name_to_run} ---")
if 'all_experiment_results_summary' not in globals(): raise
NameError("'all_experiment_results_summary' not defined. Run Cell B_runtime.")
if 'experiment_configurations' not in globals(): raise NameError("'experiment_configurations' not
defined. Run Cell B_config.")
if 'pd' not in globals(): import pandas as pd
if 'json' not in globals(): import json
if exp_name_to_run in all_experiment_results_summary and
isinstance(all_experiment_results_summary[exp_name_to_run], dict) and 'error' not in
all_experiment_results_summary[exp_name_to_run] and not
pd.isna(all_experiment_results_summary[exp_name_to_run].get('eval_accuracy', float('nan'))):
    print(f"Skipping {exp_name_to_run}, valid results exist:
{all_experiment_results_summary[exp_name_to_run]}")
else:
    if exp_name_to_run in all_experiment_results_summary: print(f"Re-running {exp_name_to_run}
(previous error/incomplete).")
    config_to_run = next((c for c in experiment_configurations if c["name"] == exp_name_to_run),
None)
    if config_to_run:
        name, result = run_ablation_experiment(config_to_run, global_variables_for_experiments)
        all_experiment_results_summary[name] = result
        print(f"\n--- Results for {name} ---")
        if isinstance(result, dict) and "error" not in result: print(f"  Eval Acc: {result.get('eval_accuracy',
'N/A'):.4f}, Loss: {result.get('eval_loss', 'N/A'):.4f}, Runtime: {result.get('eval_runtime', 'N/A'):.2f}s"); _ =
[json.dump(all_experiment_results_summary, f, indent=4) for f in [open(RESULTS_FILE, 'w')]] if
print(f"Results saved to {RESULTS_FILE}");
        elif isinstance(result, dict) and "error" in result: print(f"  Error: {result['error']}")
        else: print(f"  Unexpected result: {result}")
    else: print(f"Config for {exp_name_to_run} not found.")
print(f"--- Finished Experiment: {exp_name_to_run} ---")

```

```

# --- セル14: 最終結果表示セル (精錬ループ実験用) ---
import pandas as pd
import matplotlib.pyplot as plt
import json # jsonがインポートされていない場合に備えて

```

```

print("--- Aggregated Experiment Results (Refinement Loop Focus) ---")

```

```

if 'all_experiment_results_summary' in globals() and all_experiment_results_summary:

```



```

print(f"Found {len(all_experiment_results_summary)} entries in all_experiment_results_summary.")

results_list_for_df = []
for exp_name, metrics in all_experiment_results_summary.items():
    # 実験名に関連する設定を experiment_configurations から取得 (日本語コメントなど)
    exp_detail = next((config for config in experiment_configurations if config["name"] ==
exp_name), None)
    notes = exp_detail.get("#日本語", "") if exp_detail else ""
    tech_flags_str = str(exp_detail.get("tech_flags", {})) if exp_detail else "N/A"
    module_cfgs_str = str(exp_detail.get("module_configs", {})) if exp_detail else "N/A"

    if isinstance(metrics, dict):
        row = {
            "Experiment_Name": exp_name,
            "Accuracy": metrics.get("eval_accuracy", float('nan')), # NaN を使用
            "Loss": metrics.get("eval_loss", float('nan')),
            "Runtime_s": metrics.get("eval_runtime", float('nan')),
            "Samples_Sec": metrics.get("eval_samples_per_second", float('nan')),
            "Epochs_Eval": metrics.get("epoch", float('nan')), # trainer.evaluate() は epoch を返す
            "Error_Msg": metrics.get("error", ""),
            "Notes": notes,
            "Tech_Flags": tech_flags_str,
            "Module_Params": module_cfgs_str
        }
        results_list_for_df.append(row)
    else:
        print(f"Warning: Unexpected data type for metrics of experiment '{exp_name}':
{type(metrics)}")

if not results_list_for_df:
    print("No valid experiment data found to create a summary table.")
else:
    results_df = pd.DataFrame(results_list_for_df)
    results_df = results_df.sort_values(by="Experiment_Name").reset_index(drop=True)

    print("\nExperiment Summary Table:")
    # カラムの表示順を調整
    display_cols = ["Experiment_Name", "Accuracy", "Loss", "Runtime_s", "Epochs_Eval",
"Error_Msg", "Notes", "Tech_Flags", "Module_Params"]
    # 実際に存在するカラムのみを選択
    display_cols_present = [col for col in display_cols if col in results_df.columns]
    print(results_df[display_cols_present].to_string()) # 全て表示

    # CSVファイルへの保存
    if 'DRIVE_RESULTS_PATH' in globals() and DRIVE_RESULTS_PATH != ".":
        timestamp = datetime.datetime.now().strftime("%Y%m%d_%H%M%S")
        summary_csv_path = os.path.join(DRIVE_RESULTS_PATH,
f"refinement_experiment_summary_{timestamp}.csv")

```

```

try:
    results_df.to_csv(summary_csv_path, index=False)
    print(f"\nSummary table saved to: {summary_csv_path}")
except Exception as e:
    print(f"\nError saving summary table to CSV: {e}")
else:
    print("\nSkipping CSV save as DRIVE_RESULTS_PATH is not set to a valid GDrive path.")

# --- Accuracy比較グラフ ---
print("\n--- Generating Accuracy Comparison Graph ---")
valid_accuracy_df = results_df.dropna(subset=['Accuracy']) # NaNを除外
if not valid_accuracy_df.empty:
    plt.figure(figsize=(10, 6))
    bars = plt.bar(valid_accuracy_df["Experiment_Name"], valid_accuracy_df["Accuracy"],
color=['blue', 'orange', 'green'])
    plt.xlabel("Experiment Configuration")
    plt.ylabel("Evaluation Accuracy")
    plt.title("Experiment Accuracy Comparison (Refinement Loop)")
    plt.xticks(rotation=15, ha="right")
    plt.ylim(0, 1) # Accuracyは0から1の範囲
    for bar in bars:
        yval = bar.get_height()
        plt.text(bar.get_x() + bar.get_width()/2.0, yval + 0.01, f'{yval:.4f}', ha='center', va='bottom')
    plt.tight_layout()
    plt.show()
else:
    print("No valid Accuracy data to plot for the graph.")

# --- Loss比較グラフ ---
print("\n--- Generating Loss Comparison Graph ---")
valid_loss_df = results_df.dropna(subset=['Loss']) # NaNを除外
if not valid_loss_df.empty:
    plt.figure(figsize=(10, 6))
    bars = plt.bar(valid_loss_df["Experiment_Name"], valid_loss_df["Loss"], color=['blue', 'orange',
'green'])
    plt.xlabel("Experiment Configuration")
    plt.ylabel("Evaluation Loss")
    plt.title("Experiment Loss Comparison (Refinement Loop)")
    plt.xticks(rotation=15, ha="right")
    # plt.ylim(bottom=0) # Lossは0以上
    for bar in bars:
        yval = bar.get_height()
        plt.text(bar.get_x() + bar.get_width()/2.0, yval + 0.01 * max(valid_loss_df["Loss"]),
f'{yval:.4f}', ha='center', va='bottom') # ラベル位置調整
    plt.tight_layout()
    plt.show()
else:
    print("No valid Loss data to plot for the graph.")

```

```
else:
    print("No experiment results found in 'all_experiment_results_summary'.")

print("\n--- End of Result Output ---")
```

--- Aggregated Experiment Results (Refinement Loop Focus) ---
Found 3 entries in all_experiment_results_summary.

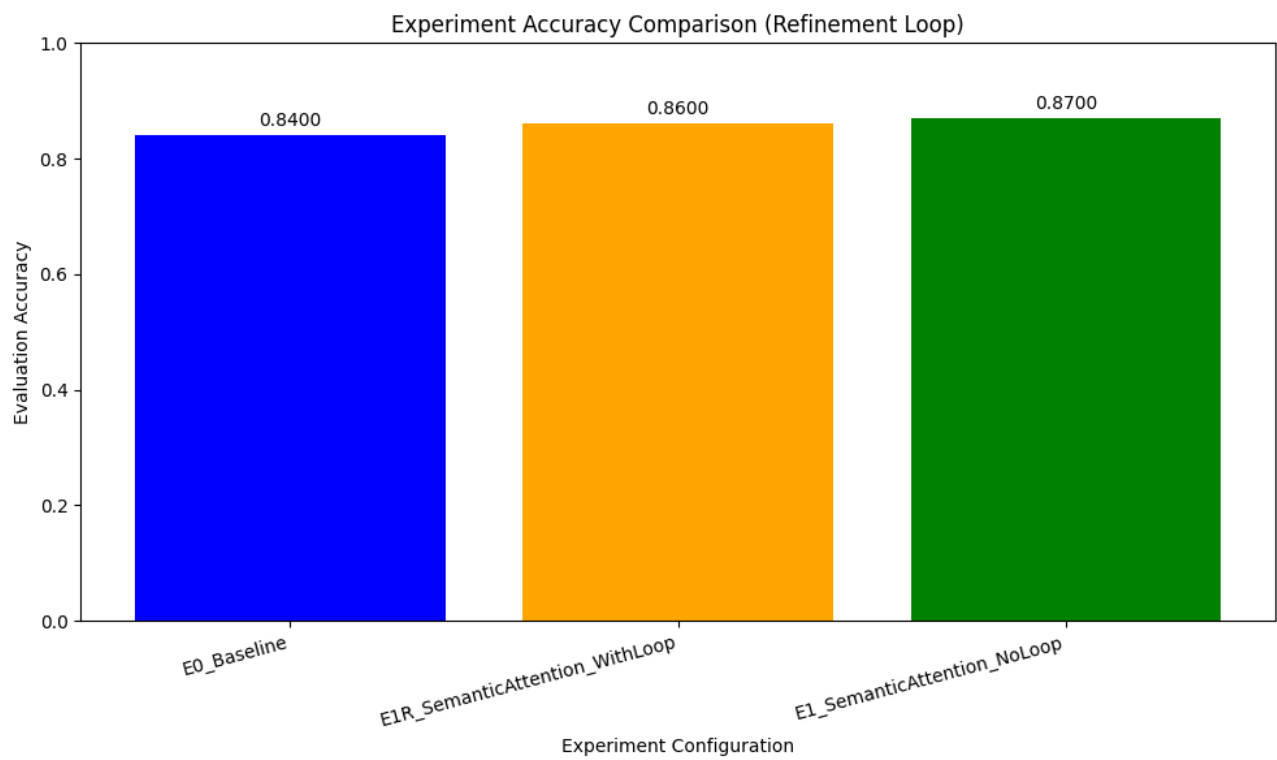
↻ --- Aggregated Experiment Results (Refinement Loop Focus) ---
Found 3 entries in all_experiment_results_summary.

Experiment Summary Table:						
	Experiment_Name	Accuracy	Loss	Runtime_s	Epochs_Eval	Error_Msg
0	E0_Baseline	0.84	0.451481	0.8313	5.0	
1	E1R_SemanticAttention_WithLoop	0.86	0.358571	0.8615	5.0	意味アテンション(反復精錬ループ付)
2	E1_SemanticAttention_NoLoop	0.87	0.389781	0.8198	5.0	意味アテンション(ループ無) {

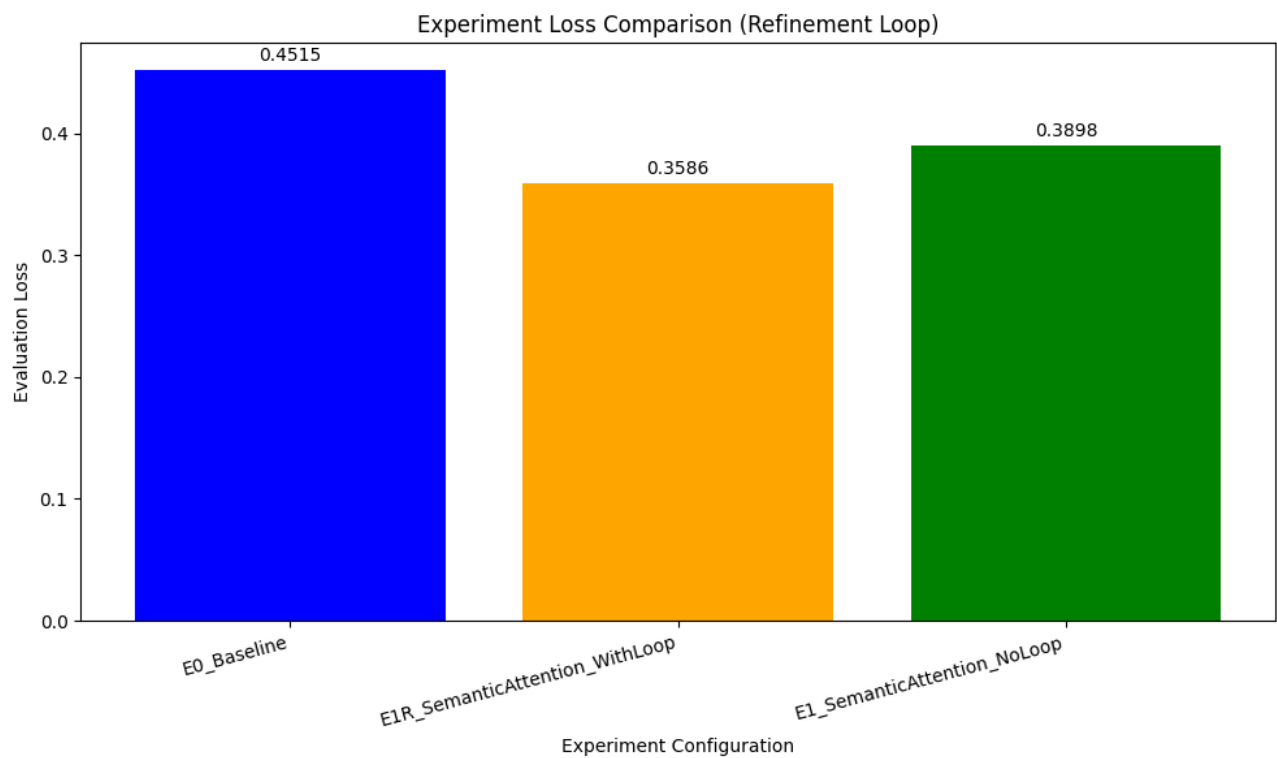
```
Experiment Summary Table:
Experiment_Name  Accuracy      Loss  Runtime_s
Epochs_Eval  Error_Msg      Notes
Tech_Flags
Module_Params
0              E0_Baseline      0.84  0.451481    0.8313      5.0              ベースライン
{'module_order': [], 'use_semantic_attention_module': False, 'use_refinement_loop': False}
{}
1  E1R_SemanticAttention_WithLoop      0.86  0.358571    0.8615      5.0              意味アテンション(反復精錬ループ付)
{'module_order': ['semantic_attention_module'], 'use_semantic_attention_module': True, 'use_refinement_loop': True}
{'semantic_attention_module': {'num_concepts': 32, 'concept_dim': 64, 'output_dim': 768}, 'refinement_loop': {'max_iterations': 3,
'convergence_threshold': 0.005}}
2      E1_SemanticAttention_NoLoop      0.87  0.389781    0.8198      5.0              意味アテンション(ループ無) {'module_order':
['semantic_attention_module'], 'use_semantic_attention_module': True, 'use_refinement_loop': False}
{'semantic_attention_module': {'num_concepts': 32, 'concept_dim': 64, 'output_dim': 768}}
```

Summary table saved to: /content/drive/MyDrive/
Colab_Ablation_Study_Outputs_Final_v2/
refinement_experiment_summary_20250531_134341.csv

--- Generating Accuracy Comparison Graph ---



--- Generating Loss Comparison Graph ---



--- End of Result Output ---

