# 大規模言語モデル単体パイプラインによる PDF 文献群からの物理モデル自動構築

加藤 祥太　古荘 大喜　加納 学
京都大学 大学院情報学研究科
{shota,manabu}@human.sys.i.kyoto-u.ac.jp
furusho.daiki.48w@st.kyoto-u.ac.jp

## 概要

科学原理に基づく物理モデルを構築するには，文献に散在する変数・数式・仮定などを抽出して組み立てる必要があり，専門知識と人的コストを要する．この過程を効率化するために，大規模言語モデル単体を推論エンジンとして用い，(1) PDF 文書からの構造化抽出，(2) 文書横断統合，(3) 目的に沿うモデル候補構築を一貫して行うパイプラインを提案する．連続槽型反応器に関する 3 報の論文を対象に GPT-5.2 ベースの提案手法を適用した結果，構造化抽出では F1 値 0.46，文書横断では pairwise F1 値 0.55，モデル候補構築では正解モデルとの最高一致率 BestF1@K=0.55 を達成した．

## 1　はじめに

科学原理に基づく物理モデルは，製造プロセスにおける設計や運転条件の検討，プロセス制御，異常診断，最適化などで基盤的な役割を担う．特に，十分な運転データや実験データを得られない状況下では，既存知見を反映した物理モデルを文献から構築できることが重要である．一方で，物理モデル構築は，対象系の仮定設定，変数の定義，既存研究の数式・パラメータの調査，表記の統一と整合性確認などの多岐にわたる作業を含み，高い専門知識と人的コストが要求される．著者らは，これらの作業負荷を軽減し，文献情報を自動的に解析・統合して目的に沿った物理モデル候補を構築するシステムである Automated Physical Model Builder（AutoPMoB）の開発に従事してきた [1]．

本稿では，AutoPMoB 開発の一環として，外部ツールに依存せず大規模言語モデル（LLM）を単体で推論エンジンとして用いて，（1）PDF 文書からの変数と方程式の抽出，（2）文書横断の統合（同義性判定に基づく表記統一），（3）目的に沿ったモデル候補構築を行う三段階パイプラインを提案する（図 1）．加えて，数式言語処理（mathematical language processing, MLP）の観点から，提案パイプラインの評価設計を提示し，提案手法と評価設計を連続槽型反応器（continuous stirred-tank reactor, CSTR）に関する 3 報の論文を用いたケーススタディを通して検証する．本研究の目的は，大規模な網羅実験ではなく，数式を含む科学文献に対する数式 NLP タスクとしての定式化と，LLM 単体による実問題解決の可能性と限界を明らかにすることにある．

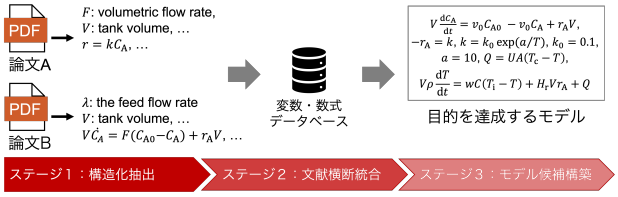


**図 1**　提案パイプラインの概要．PDF 文献群から変数・方程式を抽出し，文書横断で統合した上で，目的に沿ったモデル候補を構築する．

## 2　関連研究

科学技術文献を対象とした情報抽出は，固有表現抽出（NER），関係抽出（RE），イベント抽出（EE）を統合的に扱う枠組みやデータセットの整備が進んでいる．代表的には，エンティティ間の科学的関係を含む注釈付きコーパスである SciERC/SciIE [2]，文書レベルの関係構造を扱う DyGIE++ [3]，および構造化された引用情報を提供する SciREX [4] などがある．これらのタスクでは，情報抽出の性能を測るために適合率・再現率・F1 値を用いた評価が標準的である．また，近年では LLM による情報抽出出力の品質を評価するために，GPT-4 [5] などを用いた自動的な評価指標も提案されている [6, 7]．

文書横断的な表記統一は，エンティティリンキ

ングや共参照解析に関係が深く，OntoNotes [8] や WikiCoref [9]，CD$^2$CR [10] といったデータセットが開発されている．評価には，クラスタリング指標（MUC [11]，B$^3$ [12]，CEAF [13]，LEA [14]）が広く用いられる．応用範囲はニュース記事，医療記録，学術文献など多岐にわたるが，記号レベルの表記統一や意味同定を対象とした枠組みは限定的であり，MLP への特化は不十分である．

MLP 分野では，文献中の数学記号と自然言語定義を対応づける記号定義抽出 [15]，文献中の記号と変数定義を利用した変数抽出 [16]，および LaTeX ベースの数式を計算可能な表現に変換するための研究 [17] などがある．しかし，本文テキストと図・グラフ・数式の対応付けや数式構造の未活用がボトルネックであり [18, 19], 記号・変数の意味揺れと定義範囲の不一致が比較・統合を難しくする [20, 21] ため，複数文献から目的に応じて一貫した数理モデルを組み立てて検証する枠組みは未確立である [22]. 本研究では，情報抽出部分には適合率・再現率・F1 値を，文書横断統合には pairwise 指標と B$^3$ 指標を用いて評価を行い，式・変数を構成単位とするタスクに対する LLM の有効性を検証する．

科学技術文献を対象とした LLM の応用としては，科学論文から構造化知識を抽出してデータベース化する試み [23] や，物質・材料科学領域における情報抽出と新材料発見への応用 [24] などがある．また，製造業においても，工場内マニュアルや設備文書の解析，熟練者知識の形式化といった現場知識の活用に LLM を適用する研究が進められており [25]，その有効性と限界が議論されている．しかし，これらの研究はテキスト主体であり，数式・図表・レイアウト構造などを含む PDF ベースのマルチモーダル情報を対象とした枠組みは未だ限定的である．

# 3 手法

## 3.1 問題設定

入力は PDF 文書集合 $\boldsymbol{D} = \{d_l\}_{l=1}^{L}$ と目的 $\boldsymbol{O}$ である．$\boldsymbol{O}$ は目的文章 $o_{\text{text}}$ に加え，入力変数集合 $\boldsymbol{X} = \{x_m\}_{m=1}^{M}$，出力変数集合 $\boldsymbol{Y} = \{y_n\}_{n=1}^{N}$，作成するモデル候補の数 $K$ を含む．出力は，文書横断で整合化されたコーパス $\mathscr{C}$ と，目的に沿ったモデル候補集合 $\mathscr{M} = \{\boldsymbol{M}_k\}_{k=1}^{K}$ である．

## 3.2 提案パイプライン

文書 $d_l$ からの抽出結果を $\boldsymbol{E}_l$ とすると，提案パイプラインは次式で表される．

$$\boldsymbol{E}_l = f_{\text{ext}}(d_l, \boldsymbol{O}), \quad l = 1, 2, \ldots, L, \tag{1}$$

$$\mathscr{C} = f_{\text{int}}(\{\boldsymbol{E}_l\}_{l=1}^{L}, \boldsymbol{O}), \tag{2}$$

$$\mathscr{M} = f_{\text{build}}(\mathscr{C}, \boldsymbol{O}). \tag{3}$$

ここで $f_{\text{ext}}$ は文書内の変数と方程式の構造化抽出，$f_{\text{int}}$ は文書間の同義変数・同義方程式の統合，$f_{\text{build}}$ は目的に応じたモデル候補の合成を表す．

**ステージ 1：構造化抽出**　文書 $d_l$ から，$\boldsymbol{O}$ を参照しつつ文献要約・変数リスト・方程式リストを取得する．各変数は LaTeX 形式の記号・定義・文脈からなり，各方程式は LaTeX 形式の記号列・変数リスト・方程式説明文からなる．具体的には，まず，対象プロセスの物理モデル構築に直接関係する方程式を抽出して方程式リストを作成し，その後，各方程式に出現する変数について，本文または図表を参照して意味を特定し，変数リストを作成する．数式中に出現するものの意味を一意に特定できない変数は，変数リストには含めない．後段での文書横断統合のために，抽出結果 $\boldsymbol{E}_l$ には出典情報（文書 ID）と近傍文脈を付与する．

**ステージ 2：文書横断統合**　ステージ 1 において複数文献から抽出した変数と方程式の表記揺れを判定し，同一概念の表記を統一（正規化）する．本工程では，$\{\boldsymbol{E}_l\}_{l=1}^{L}$ と $\boldsymbol{O}$ から，正規化した変数集合 $\bar{\boldsymbol{V}}$ と方程式集合 $\bar{\boldsymbol{Q}}$ からなるコーパス $\mathscr{C} = (\bar{\boldsymbol{V}}, \bar{\boldsymbol{Q}})$ を生成する．モデル統合で用いるため，統合後の情報 $\bar{v} \in \bar{\boldsymbol{V}}$ と $\bar{q} \in \bar{\boldsymbol{Q}}$ には，出典情報を保持し，参照可能性を確保する．

**ステージ 3：モデル候補構築**　ステージ 2 で得た $\mathscr{C}$ と $\boldsymbol{O}$ を入力として，$K$ 個のモデル候補を生成する．各モデル候補 $\boldsymbol{M}_k$ は，採用する変数集合 $\bar{\boldsymbol{V}}_k \subseteq \bar{\boldsymbol{V}}$ と方程式集合 $\bar{\boldsymbol{Q}}_k \subseteq \bar{\boldsymbol{Q}}$，モデル候補の説明文および他モデル候補との差分説明文から構成される．この工程は，文献から抽出された知識を統合して実行可能なモデルへ接続する研究 [26, 27] と問題意識を共有するが，数式中心の物理モデリングを扱う点や単体 LLM のパイプラインを扱う点が異なる．

**プロンプト設計**　各ステージのプロンプトは，役割に応じた指示文と出力形式を含むシステムプロンプトと，前ステージの出力と $\boldsymbol{O}$ とを組み合わせたユーザプロンプトから構成した．実験に使用したプ

**表 1**　データセットの概要．$L$ は文書数，$M$ と $N$ は目的に含まれる入力変数と出力変数の数，$\tilde{\boldsymbol{V}}$ と $\tilde{\boldsymbol{Q}}$ は正規化後の変数と方程式の集合，$|\cdot|$ は集合の要素数を表す．

| プロセス | $L$ | $M$ | $N$ | $\lvert\tilde{\boldsymbol{V}}\rvert$ | $\lvert\tilde{\boldsymbol{Q}}\rvert$ |
|---|---|---|---|---|---|
| CSTR | 3 | 2 | 2 | 27 | 9 |

ロンプトは付録 A に示す．

# 4　実験

## 4.1　データセットとモデル

CSTR に関する論文を $L = 3$ 報用意し，人手で各ステージに対する正解データを作成した．具体的には，まず，各論文から変数と方程式を抽出し，続いて，抽出した変数と方程式の表記を正規化した集合を作成し，最後に，事前に設定した目的 $\boldsymbol{O}$ に沿うモデル候補を作成した．表 1 に対象文書の概要を示す．

使用する LLM は gpt-5.2[1)] とした．温度パラメータは 0 とし，PDF 文書からモデル候補までの一連の工程をプロセスごとに 5 回ずつ実行し，4.2 節に示す評価指標の平均値と標準偏差を算出した．

## 4.2　評価指標

**ステージ 1：構造化抽出**　正解集合 $\boldsymbol{G}^*$ に対する LLM の出力 $\hat{\boldsymbol{G}}$ を，適合率 $P$，再現率 $R$，F1 値 $F$ を用いて評価する．これらの評価指標は，

$$P = \frac{|\hat{\boldsymbol{G}} \cap \boldsymbol{G}^*|}{|\hat{\boldsymbol{G}}|}, \quad R = \frac{|\hat{\boldsymbol{G}} \cap \boldsymbol{G}^*|}{|\boldsymbol{G}^*|}, \quad F = \frac{2PR}{P+R}, \tag{4}$$

で与えられる．$|\cdot|$ は集合の要素数を表す．変数と方程式のそれぞれについて，LaTeX 形式の文字列を用いて一致判定を行う．なお，表記揺れが評価指標に与える影響を小さくするため，空白除去や等号左右の整形といった正規化を前処理として適用する．

**ステージ 2：文書横断統合**　ステージ 1 の出力 $\hat{\boldsymbol{G}}$ の要素は，正解集合 $\boldsymbol{G}^*$ の要素と必ずしも一致しない．以降で説明する評価では予測結果と正解との対応付けが必要であるため，同一文書内で記号列や定義文に基づく対応付けを行ってから評価を行う．具体的には，写像 $\pi : \hat{\boldsymbol{G}} \to \boldsymbol{G}^* \cup \{\perp\}$ を定義し，$\pi(\hat{\imath})$ を予測要素 $\hat{\imath} \in \hat{\boldsymbol{G}}$ に対応する正解要素（未対応の場合は $\perp$）とする．このとき，評価対象集合を

$$\boldsymbol{U} = \{\, i \in \boldsymbol{G}^* \mid \exists \hat{\imath} \in \hat{\boldsymbol{G}} \text{ s.t. } \pi(\hat{\imath}) = i \,\} \tag{5}$$

と定義する．pairwise 指標および $B^3$ 指標は $\boldsymbol{U}$ 上で計算する．$\pi$ は，各文書内で正規化した記号列と定義文の類似度に基づく決定的な手順により一対一に構成し，対応の取れない要素は $\perp$ に写像する．

正解集合 $\boldsymbol{G}^*$ に基づき，同一概念を有する要素の集合として正解クラスタが与えられているとする．このとき，正解における同義な要素ペアの集合を

$$\begin{aligned}\boldsymbol{L}^* = \{(i,j) \mid\ & i<j,\ i,j \in \boldsymbol{U},\\ & i \text{ と } j \text{ が同一の正解クラスタに属する }\}\end{aligned} \tag{6}$$

と定義する．同様に，LLM によって同義と判定された要素ペアの集合を

$$\begin{aligned}\hat{\boldsymbol{L}} = \{(i,j) \mid\ & i<j,\ i,j \in \boldsymbol{U},\\ & \exists \hat{\imath},\hat{\jmath} \in \hat{\boldsymbol{G}} \text{ s.t. } \pi(\hat{\imath})=i,\ \pi(\hat{\jmath})=j,\\ & \hat{\imath} \text{ と } \hat{\jmath} \text{ が同一の予測クラスタに属する }\}\end{aligned} \tag{7}$$

と定義する．

pairwise 指標（適合率，再現率，F1 値）は，$\boldsymbol{L}^*$ と $\hat{\boldsymbol{L}}$ を用いて次式で与えられる．

$$P_{\mathrm{pw}} = \frac{|\hat{\boldsymbol{L}} \cap \boldsymbol{L}^*|}{|\hat{\boldsymbol{L}}|}, R_{\mathrm{pw}} = \frac{|\hat{\boldsymbol{L}} \cap \boldsymbol{L}^*|}{|\boldsymbol{L}^*|}, F_{\mathrm{pw}} = \frac{2P_{\mathrm{pw}}R_{\mathrm{pw}}}{P_{\mathrm{pw}}+R_{\mathrm{pw}}}. \tag{8}$$

$P_{\mathrm{pw}}$ は異なる概念を誤って統合する過剰統合を，$R_{\mathrm{pw}}$ は同一概念を統合し損ねる未統合を評価する．

$B^3$ 指標は，pairwise 指標とは異なり，各要素 $i$ に着目して，その要素が属するクラスタの重なりを局所的に評価し，全要素で平均する．要素 $i \in \boldsymbol{U}$ が属する正解クラスタを $C_i^*$，予測クラスタを $\hat{C}_i$ とすると，要素 $i$ に対する $B^3$ の適合率と再現率は

$$P_{\mathrm{B}^3,i} = \frac{|\hat{C}_i \cap C_i^*|}{|\hat{C}_i|}, \quad R_{\mathrm{B}^3,i} = \frac{|\hat{C}_i \cap C_i^*|}{|C_i^*|} \tag{9}$$

で定義される．ここで，$C_i^*$ および $\hat{C}_i$ は，それぞれ正解クラスタおよび予測クラスタを評価対象集合 $\boldsymbol{U}$ 上に制限した集合とする．$B^3$ 指標では，これらを全要素について平均することで，全体の適合率および再現率を得る．

$$P_{\mathrm{B}^3} = \frac{1}{|\boldsymbol{U}|}\sum_{i\in\boldsymbol{U}} P_{\mathrm{B}^3,i}, \quad R_{\mathrm{B}^3} = \frac{1}{|\boldsymbol{U}|}\sum_{i\in\boldsymbol{U}} R_{\mathrm{B}^3,i}. \tag{10}$$

$B^3$ の F1 値 $F_{\mathrm{B}^3}$ は，これらの調和平均である．pairwise 指標は，同義と判定した要素ペア全体の整合性を評価するのに対し，$B^3$ 指標は，各要素が属するクラスタの局所的な純度と網羅性を評価する．

**ステージ 3：モデル候補構築**　生成されたモデル候補に対して，正解モデルとどの程度一致しているか（部分一致）と，正解モデルを完全に再現できているか（厳密一致）の観点から評価する．

1)　https://openai.com/index/introducing-gpt-5-2/

正解モデル候補集合を $\mathcal{M}^* = \{\boldsymbol{M}_j^*\}_{j=1}^{J}$ とし，各正解モデル $\boldsymbol{M}_j^*$ は，変数集合 $\bar{\boldsymbol{V}}_j^*$ と方程式集合 $\bar{\boldsymbol{Q}}_j^*$ を持つとする．

部分一致評価では，各モデル候補 $\boldsymbol{M}_k$ と各正解モデル $\boldsymbol{M}_j^*$ の間で，方程式集合の重なりに基づき，適合率と再現率を

$$P_{k,j} = \frac{|\bar{\boldsymbol{Q}}_k \cap \bar{\boldsymbol{Q}}_j^*|}{|\bar{\boldsymbol{Q}}_k|}, \quad R_{k,j} = \frac{|\bar{\boldsymbol{Q}}_k \cap \bar{\boldsymbol{Q}}_j^*|}{|\bar{\boldsymbol{Q}}_j^*|}, \tag{11}$$

と定義する．F1 値 $F_{k,j}$ はこれらの調和平均であり，モデル候補 $\boldsymbol{M}_k$ が正解モデル $\boldsymbol{M}_j^*$ をどの程度再現しているかを表す．各モデル候補 $\boldsymbol{M}_k$ に対して，最も一致度の高い正解モデルを基準とした最良一致度を

$$F_{k,\mathrm{best}} = \max_j F_{k,j} \tag{12}$$

と定義する．これに基づき，生成された $K$ 個のモデル候補に対して，

$$\begin{aligned} \mathrm{BestF1@K} &= \max_{k \le K} F_{k,\mathrm{best}}, \\ \mathrm{MeanBestF1@K} &= \frac{1}{K} \sum_{k=1}^{K} F_{k,\mathrm{best}} \end{aligned} \tag{13}$$

を算出する．BestF1@K は，$K$ 個の候補の中で最も正解に近いモデルがどの程度正解に一致するかを示し，MeanBestF1@K は，生成された候補全体としての平均的な一致度を表す．

厳密一致では，正解モデルをそのまま再現できたかを評価する．具体的には，$K$ 個のモデル候補の中に，正解モデルのいずれかと完全に一致するものが存在するかどうかを判定する．厳密一致指標 Exact@K は次のように定義される．

$$\mathrm{Exact@K} = \begin{cases} 1, & \text{if } \exists k \le K,\ \exists j \text{ s.t. } \bar{\boldsymbol{Q}}_k = \bar{\boldsymbol{Q}}_j^*, \\ 0, & \text{otherwise.} \end{cases} \tag{14}$$

複数回の結果に対して Exact@K の平均と標準偏差を算出することで，正解モデルを $K$ 個以内に厳密に再現できた割合を算出する．

## 5 結果と考察

各ステージにおける評価結果を表 2 に示す．

ステージ 1 では，変数抽出が F1 値 $0.46 \pm 0.05$（適合率 $0.36 \pm 0.06$，再現率 $0.66 \pm 0.04$）を達成した一方，方程式抽出は F1 値 $0.01 \pm 0.01$ であった．変数抽出では再現率が適合率を上回っており，LLM がモデル構築に関連する変数を広く捕捉している一方で，関連の薄い変数も多く抽出していることが示唆される．方程式抽出の低い値は，LaTeX 形式の厳密な文字列マッチングによる評価が，意味的に同一だが表記が異なる数式を不一致と判定したことに起因するため，数式の正規化処理の改善が課題である．

**表 2** 各ステージにおける評価結果．ステージ 1 は F1 値，ステージ 2 は pairwise F1 値と $B^3$ F1 値，ステージ 3 は BestF1@K，MeanBestF1@K，Exact@K を示す．各指標の平均値と標準偏差を併記する．

| ステージ | 指標 | 平均 ± 標準偏差 |
|---|---|---|
| 1 | 変数 F1 値 | $0.46 \pm 0.05$ |
| | 方程式 F1 値 | $0.01 \pm 0.01$ |
| 2 | pairwise F1 値 | $0.55 \pm 0.05$ |
| | $B^3$ F1 値 | $0.82 \pm 0.02$ |
| 3 | BestF1@K | $0.55 \pm 0.13$ |
| | MeanBestF1@K | $0.36 \pm 0.06$ |
| | Exact@K | $0.00 \pm 0.00$ |

ステージ 2 では，変数の統合において pairwise F1 値 $0.55 \pm 0.05$（適合率 $0.91 \pm 0.03$，再現率 $0.39 \pm 0.05$），$B^3$ F1 値 $0.82 \pm 0.02$（適合率 $0.97 \pm 0.01$，再現率 $0.71 \pm 0.02$）を達成した．高い適合率は，LLM が同義と判定した変数ペアの大部分が正しいことを示す．一方で再現率の低さは，異なる記号で表記された同一概念（例：反応速度定数の $K$ と $k$）の同定に課題があることを示唆する．方程式の統合では全指標が 1.00 であったが，これはステージ 1 で方程式が正しく抽出されなかったため，評価対象集合 $\boldsymbol{U}$ が空に近い状態となった結果である．

ステージ 3 では，BestF1@K=$0.55 \pm 0.13$，MeanBestF1@K=$0.36 \pm 0.06$，Exact@K=$0.00 \pm 0.00$ であった．最良候補が正解モデルの約 55%を再現した一方，正解モデルを完全に再現した事例はなかった．これは，前段で変数・方程式が欠落・誤統合されたことが影響している．今後は，ステージ間のエラー伝播を抑制するための手法開発に取り組む．

## 6 おわりに

複数の PDF 文書から目的に沿った物理モデル候補を自動生成するために，LLM を用いた，構造化抽出，文書横断統合，モデル候補構築からなる三段階パイプラインを提案した．提案パイプラインを，連続槽型反応器に適用し，構造化抽出では F1 値 0.46，文書横断統合では pairwise F1 値 0.55，モデル候補構築では正解モデルとの最高一致率 BestF1@K=0.55 を達成した．今後は，各ステージの性能向上とや他の LLM との比較の検討を進める．

## 謝辞

本発表は，JST，ACT-X，JPMJAX23C5 の支援を受けたものである．

## 参考文献

[1] Shota Kato and Manabu Kano. Automated physical model building from literature sources: Combining equations based on four predefined requirements. **Computers & Chemical Engineering**, Vol. 200, p. 109147, 2025.

[2] Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, pp. 3219–3232, 2018.

[3] David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. Entity, relation, and event extraction with contextualized span representations. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 5784–5789, 2019.

[4] Sarthak Jain, Madeleine van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy. SciREX: A challenge dataset for document-level information extraction. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 7506–7516, 2020.

[5] OpenAI. "gpt-4 technical report", 2024.

[6] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 2511–2522, 2023.

[7] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In **Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track**, 2023.

[8] Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In **Joint Conference on EMNLP and CoNLL - Shared Task**, pp. 1–40, 2012.

[9] Abbas Ghaddar and Phillippe Langlais. WikiCoref: An English coreference-annotated corpus of Wikipedia articles. In **Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)**, pp. 136–142, 2016.

[10] James Ravenscroft, Amanda Clare, Arie Cattan, Ido Dagan, and Maria Liakata. CD$^2$CR: Co-reference resolution across documents and domains. In **Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume**, pp. 270–280, 2021.

[11] Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. A model-theoretic coreference scoring scheme. In **Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995**, 1995.

[12] Amit Bagga and Breck Baldwin. Entity-based cross-document coreferencing using the vector space model. In **COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics**, 1998.

[13] Xiaoqiang Luo. On coreference resolution performance metrics. In **Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing**, pp. 25–32, 2005.

[14] Nafise Sadat Moosavi and Michael Strube. Which coreference evaluation metric do you trust? a proposal for a link-based entity aware metric. In **Proceedings of ACL 2016**, pp. 632–642. Association for Computational Linguistics, 2016.

[15] Moritz Schubotz, Alexey Grigorev, Marcus Leich, Howard S. Cohl, Norman Meuschke, Bela Gipp, Abdou S. Youssef, and Volker Markl. Semantification of identifiers in mathematics for better math information retrieval. In **Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval**, SIGIR '16, p. 135–144, 2016.

[16] Chunwei Liu, Enrique Noriega-Atala, Adarsh Pyarelal, Clayton T Morrison, and Mike Cafarella. Variable extraction for model recovery in scientific literature. In **Proceedings of the 1st Workshop on AI and Scientific Discovery: Directions and Opportunities**, pp. 1–12, 2025.

[17] Andre Greiner-Petter, Moritz Schubotz, Corinna Breitinger, Philipp Scharpf, Akiko Aizawa, and Bela Gipp. Do the Math: Making Mathematics in Wikipedia Computable . **IEEE Transactions on Pattern Analysis & Machine Intelligence**, Vol. 45, No. 04, pp. 4384–4395, 2023.

[18] Kenny Davila and Richard Zanibbi. Layout and semantics: Combining representations for mathematical formula search. In **Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval**, SIGIR '17, pp. 1165–1168, 2017.

[19] Behrooz Mansouri, Shaurya Rohatgi, Douglas W. Oard, Jian Wu, C. Lee Giles, and Richard Zanibbi. Tangent-cft: An embedding model for mathematical formulas. In **Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval**, ICTIR '19, pp. 11–18, 2019.

[20] Moritz Schubotz, Leonard Krämer, Norman Meuschke, Felix Hamborg, and Bela Gipp. Evaluating and improving the extraction of mathematical identifier definitions. In **Experimental IR Meets Multilinguality, Multimodality, and Interaction**, pp. 82–94, 2017.

[21] Deborah Ferreira, Mokanarangan Thayaparan, Marco Valentino, Julia Rozanova, and Andre Freitas. To be or not to be an integer? encoding variables for mathematical text. In **Findings of the Association for Computational Linguistics: ACL 2022**, pp. 938–948, 2022.

[22] Jordan Meadows and Andre Freitas. A survey in mathematical language processing, 2024.

[23] John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S Rosen, Gerbrand Ceder, Kristin A Persson, and Anubhav Jain. Structured information extraction from scientific text with large language models. **Nature Communications**, Vol. 15, p. 1418, 2024.

[24] Vahe Tshitoyan, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova, Kristin A Persson, Gerbrand Ceder, and Anubhav Jain. Unsupervised word embeddings capture latent knowledge from materials science literature. **Nature**, Vol. 571, No. 7763, pp. 95–98, 2019.

[25] Samuel Kernan Freire, Chaofan Wang, Mina Foosherian, Stefan Wellsandt, Santiago Ruiz-Arenas, and Evangelos Niforatos. Knowledge sharing in manufacturing using llm-powered tools: User study and model benchmarking. **Frontiers in Artificial Intelligence**, Vol. 7, , 2024.

[26] Paul R. Cohen. Darpa's big mechanism program. **Physical Biology**, Vol. 12, No. 4, p. 045008, 2015.

[27] Rebecca Sharp, Adarsh Pyarelal, Benjamin Gyori, et al. Eidos, indra, & delphi: From free text to executable causal models. In **Proceedings of NAACL 2019 (Demonstrations)**, pp. 42–47. Association for Computational Linguistics, 2019.

# A　実験に使用したプロンプト

各ステージで使用したシステムプロンプトの概要を以下に示す.

**ステージ 1：構造化抽出**

You are a scientific extraction agent for chemical-process modeling. Task: Extract equations and variables relevant to the target process's physical model construction. Extraction strategy (equation-driven): 1. First, identify governing equations directly related to the target process. 2. Then, identify variables whose meanings can be determined from text, figures, or tables. Symbol normalization: Remove LaTeX decorations and time arguments.

**ステージ 2：文書横断統合**

You are a scientific knowledge organizer for process modeling. Task: Normalize notation across multiple documents and identify synonymous variables and equations. Integration strategy: 1. Group variables representing the same physical quantity under a canonical symbol. 2. Identify equations describing the same relationship across documents. 3. Preserve source information for traceability.

**ステージ 3：モデル候補構築**

You are a process engineering modeling expert. Task: Construct mathematical model candidates from integrated variables and equations that satisfy the given objective. Model construction guidelines: 1. Select relevant equations from the provided corpus. 2. Identify required variables for each model. 3. Explain how each model differs from others.