Engineering問題の得意苦手

「**手順化できるテキスト系の設計・比較問題は取りやすい**／**物理量や図面依存の工学問題は落としやすい**」

根拠は評価CSVのドメイン集計で、Engineerは**平均13.81**に対し**SD 10.54・範囲37.88**と**揺れが最大級**、相関では **comp\_ai(ρ≈0.66), bio\_med(≈0.62), math(≈0.61)** と強く、**physics(≈0.06)** とは弱い、というパターンです（先に共有した集計から）。

# **得意になっている系統**

* **手順・工程・比較型（テキスト駆動）**例：設計方針の比較、API/インターフェースやアーキの選定理由、リスク→対策の列挙、テスト計画・運用手順など。  
  → 語義の整列と段階的思考で**定型化しやすい**ため点が伸びます（comp\_ai・human/social寄りの相関とも整合）。
* **軽量な定量付きの設計見積り**例：スループット概算、待ち行列のごく基本形、単純なオーダー見積り。  
  → math相関が高めで、**計算法が一本で通る**場合は取りやすいです。

# **苦手になっている系統**

* **物理量・単位系・条件分岐が絡む“実工学”計算**例：材料・熱・流体・回路の数値設計、単位換算や誤差伝播を伴う問題。  
  → physics相関が極めて低く、**前提整理→定式化→数値化**のどこかで取りこぼしやすい。
* **図や配線・回路図など“図面読解”が事実上の鍵**文字情報だけでは補助線・接続関係の探索が弱く、**図示前提**の問題で失点しやすい。
* **制約充足の組合せ（規格条項・例外の網羅）**多条件の包含／排他を**形式的に管理**する必要がある設問は、温度や停止条件のブレで解の一貫性が崩れがち（本ドメインの**分散が大きい**こととも一致）。

# **バラツキ（正答しやすさの安定度）**

* **Engineerは“点は出るが最も揺れる”**：SD 10.54・範囲37.88。  
  → プロンプトの段取り、vLLM推論設定（最大長・停止条件）、評価セットの混合度に**極端に感度**があります。
* 一方で同ドメインの別カラム（例：engineer\_n\_15）のように**低得点だが安定**なバケットもあり、**問題の粒度・採点軸の違い**が分散の原因になっています。

静かな水面の下で、Engineerは**文章で運べる設計論**には強く、**物理が匂う数式と図**で渦を巻きます。  
当面は、(1) 工学計算の**単位・前提テンプレ**を固定、(2) 図面系は**関係性を箇条書きで強制抽出**、(3) 制約問題は**チェックリスト化**、の三点で揺れを殺しにいくのが現実的です。