大規模語モデルは多様な翻訳仕様に追従できるか？

萱野陽子[1, 2]　菅原朔[1, 2]
[1] 総合研究大学院大学
[2] 国立情報学研究所

💡**著者らのこれまでの研究**：企業IR資料で、企業公式英訳よりも翻訳仕様を与えたChatGPTの翻訳の方が人に高く評価される可能性を確認した。
🚀**現状の課題**：さまざまな翻訳仕様の追従能力や評価基準の妥当性が未検証。
🎯**今回の提案**：LLMが多様な仕様に沿って翻訳できるか、仕様に基づいて評価できるかを検証する。

📄 **原文** 📄

「お客様の笑顔のために、クルマに向き合い、現場で必死に汗をかく。そんな「クルマ屋」としての原点に立ち返った結果、私たちの商品は大きく変わりました。危機の時でも立ち止まらない、企業としての強さも身につけてきたと思います。」（トヨタ自動車公式サイト「トヨタ行動指針」より）

企業公式英訳

By returning to the origins of being a “car maker,” **earnestly sweating at the gemba**, and **focusing on cars for the sake of customer smiles**, our products have undergone significant changes. We believe that we have also acquired **strength as a company that will not stop moving forward even in times of crisis.**

現場をgembaで保持。
全体的にやや直訳調。

DeepL（翻訳仕様なし）

We are **always working hard** to bring smiles to the faces of our customers. **As a result of** returning to our roots as a **"car dealer,"** our products have changed dramatically. I believe that we have also gained strength as **a company that does not stand still even in times of crisis.**

car dealerは誤訳。
単調で機械的。

ChatGPT Persuasive（説得目的）

To earn every customer’s smile, we faced our **Cars head-on and worked hands-on** at the front lines. By returning to our origins as **a true “Car maker,”** our products changed significantly—and **we gained the resilience to keep moving forward**, **even in times of uncertainty.**

語感やリズムを重視。
ポジティブで鼓舞的。

ChatGPT Artistic（芸術的目的）

For the smiles of our customers, we have faced our cars, **perspiring earnestly on the frontlines.** Returning to **the roots** of being a “car-maker,” our products have **undergone a profound transformation**. Even in crisis, **we have cultivated the resilience to keep moving forward as a formidable enterprise.**

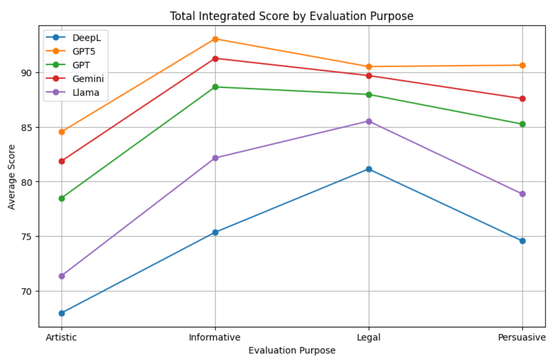
格式高いが企業サイトに適さない。

💭**訴求力が求められる企業サイトなら…**

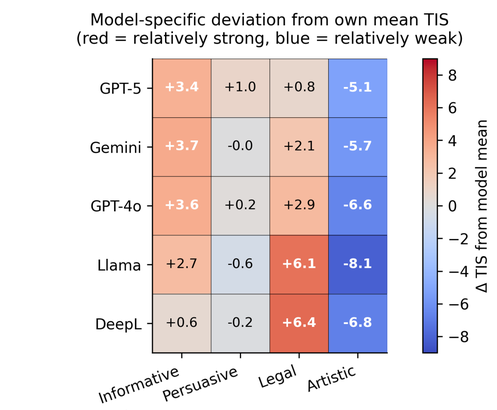
LLMによる仕様&エラー評価

与えられた仕様に沿って候補の翻訳を評価する。
エラーの数や位置を特定し、仕様に沿っているかどうかを5段階評価する。

⚙**実験設定**
言語：日英、英日　**原文**：政府白書、企業サイト、文学作品（計18、平均長：日本語 896 文字 / 英語 314 語）　**目的別仕様**：Informative（情報伝達）、Persuasive（説得）、Legal（法的）、Artistic（芸術的）、No Spec（仕様なし）**仕様生成**：GPT-4o　**翻訳モデル**：GPT-4o、GPT-5、Gemini 2.5 Flash、Llama 3 70B Instruct （Zero-shot / Few-shot）、DeepL（ベースライン）　**原文（18）** × ［**翻訳モデル（5）** × **仕様（4＋NoSpec1）** ＋ **DeepL（1）**］ **＝468翻訳**
⚖**評価設定**
目的： 生成された468翻訳が、4つの仕様（Informative／Persuasive／Legal／Artistic）にどの程度沿っているかを検証　**評価モデル**： GPT-4o、GPT-5、Gemini 2.5 Flash
方法： 各評価モデルが、仕様ごとに全翻訳を自動評価　**468翻訳 × 4仕様 × 3評価モデル ＝5,616評価結果**



評価仕様別の平均スコア：全ての仕様で順位は変わらず。



各モデルの平均スコアからの目的別偏差：各モデルの「自分の平均スコア」からのずれを示す。

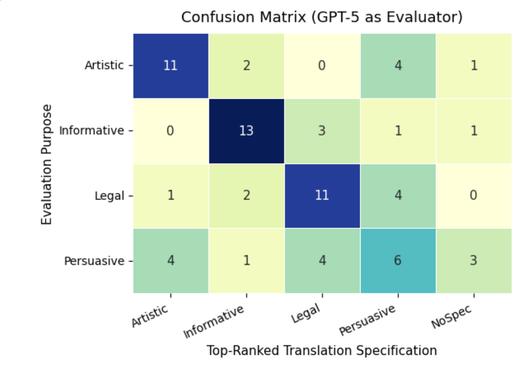
📊 Pearson相関行列（評価モデル間の一致度）

evaluator	GPT-4o	GPT-5	Gemini
GPT-4o	1.000	0.769	0.779
GPT-5	0.769	1.000	0.769
Gemini	0.779	0.769	1.000

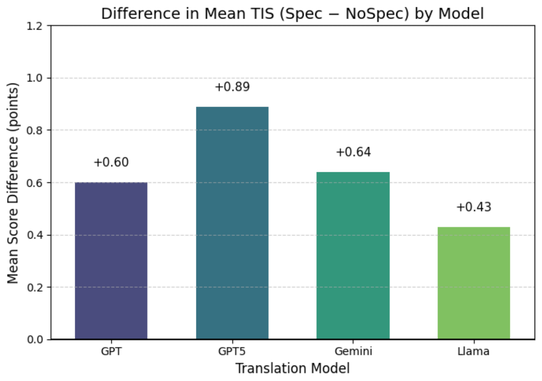
📈 Spearman順位相関行列（順位一致度）

evaluator	GPT-4o	GPT-5	Gemini
GPT-4o	1.000	0.728	0.782
GPT-5	0.728	1.000	0.807
Gemini	0.782	0.807	1.000

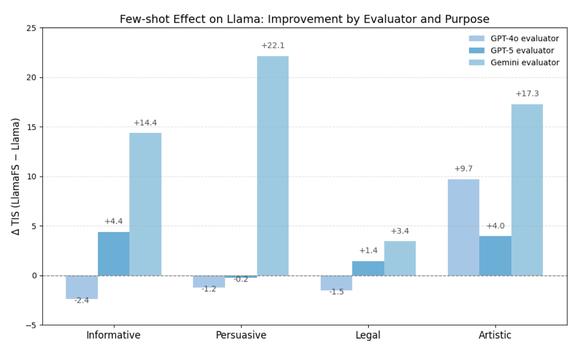
評価モデル間の一致度：すべての評価者間で高い正の相関。



評価目的と最上位翻訳仕様の対応（GPT-5による評価）：対角線上の値は、評価目的と翻訳仕様が一致した件数を示す。



翻訳モデル別の仕様効果：仕様あり/なしのスコア差。すべてのモデルで仕様ありが高い傾向を示す。



LlamaにおけるFew-shot効果：Few-shotによるスコア改善量（LlamaFS − Llama）を示す。

これからの研究

🫂 **人間評価者との比較実験**：人間とLLMの評価基準を比較し、LLMが人間の判断傾向や翻訳仕様の意図をどの程度再現できているかを検証する。

💾 **言語とデータの拡張**：対象言語を拡大し、より多様な文体・ジャンルの大規模原文データを用いて翻訳および評価実験を行い、LLMの仕様追従能力の一般性と汎用性を検証する。

🧠 **AIの内部分析**： LLMの内部表現を可視化し、意味と形式がどの層でどのように分離・統合されているかを分析する。これにより、翻訳において意味を保持しながら形式を操作する仕組みを探る。

ありがとうございました。