

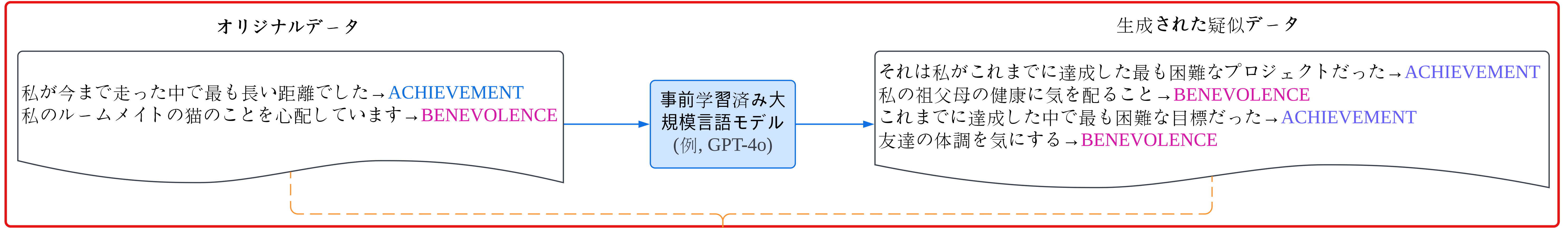
## 背景・目的

大規模言語モデル(LLM)による疑似データ生成を用いて、価値観データの不足を補うことができでないか考えた。本研究では疑似データを用いて価値観推定モデルの精度の向上が見られるかを検証した。

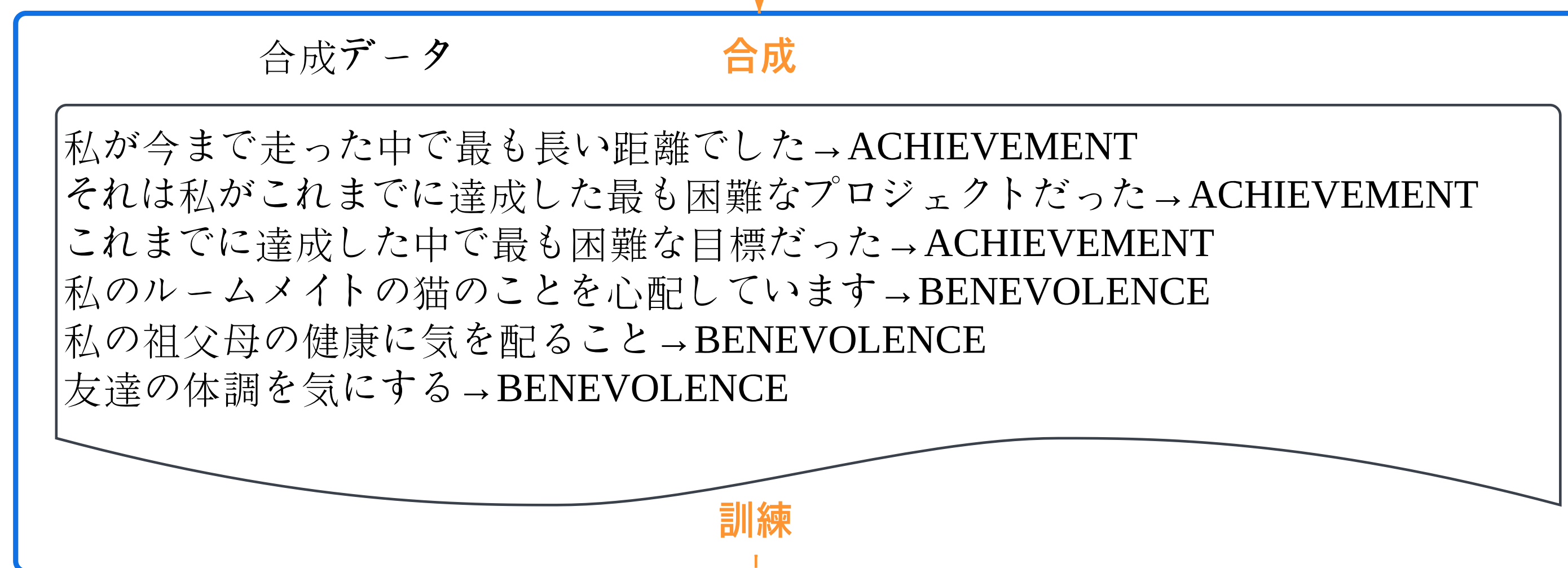
**オリジナルデータ: ValueNet**という10種類の価値観からなる大規模なデータセットを使用する

## 提案手法モデル

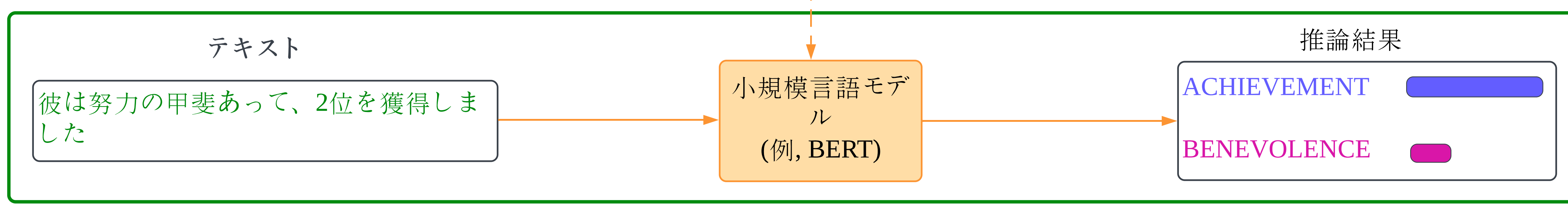
(1) 価値観推定モデルの作成



(2) 価値観推定モデルの作成



(3) 人間の価値観推定



### 価値観データセット

ID	価値観のラベル	データ数
0	ACHIEVEMENT	192
1	BENEVOLENCE	888
2	CONFORMITY	91
3	HEDONISM	819
4	POWER	438
5	SECURITY	637
6	SELF-DIRECTION	108
7	STIMULATION	305
8	TRADITION	98
9	UNIVERSALISM	294
	合計	3870

## 検証実験

### 検証項目

- オリジナル英語データを拡張し学習データ数を増やす
  - オリジナルの英語データを日本語翻訳した上で拡張し、学習データ数を増やす
- 翻訳に用いたツールはDeepLとGPT-3.5turboである

### 疑似データ生成のプロンプト

**System Prompt:**  
You are a sociologist who is knowledgeable about human values.

**User Prompt:**  
Your task is to generate sentences that reflect specific human values. The sentences should clearly convey the given human value based on its definition. The generated sentence should use words and structure similar to the example sentence.

Here is the definition of the value {VALUE}:  
{DEFINITION}

Here are example sentences that reflect this value {VALUE}:  
{SENTENCE}

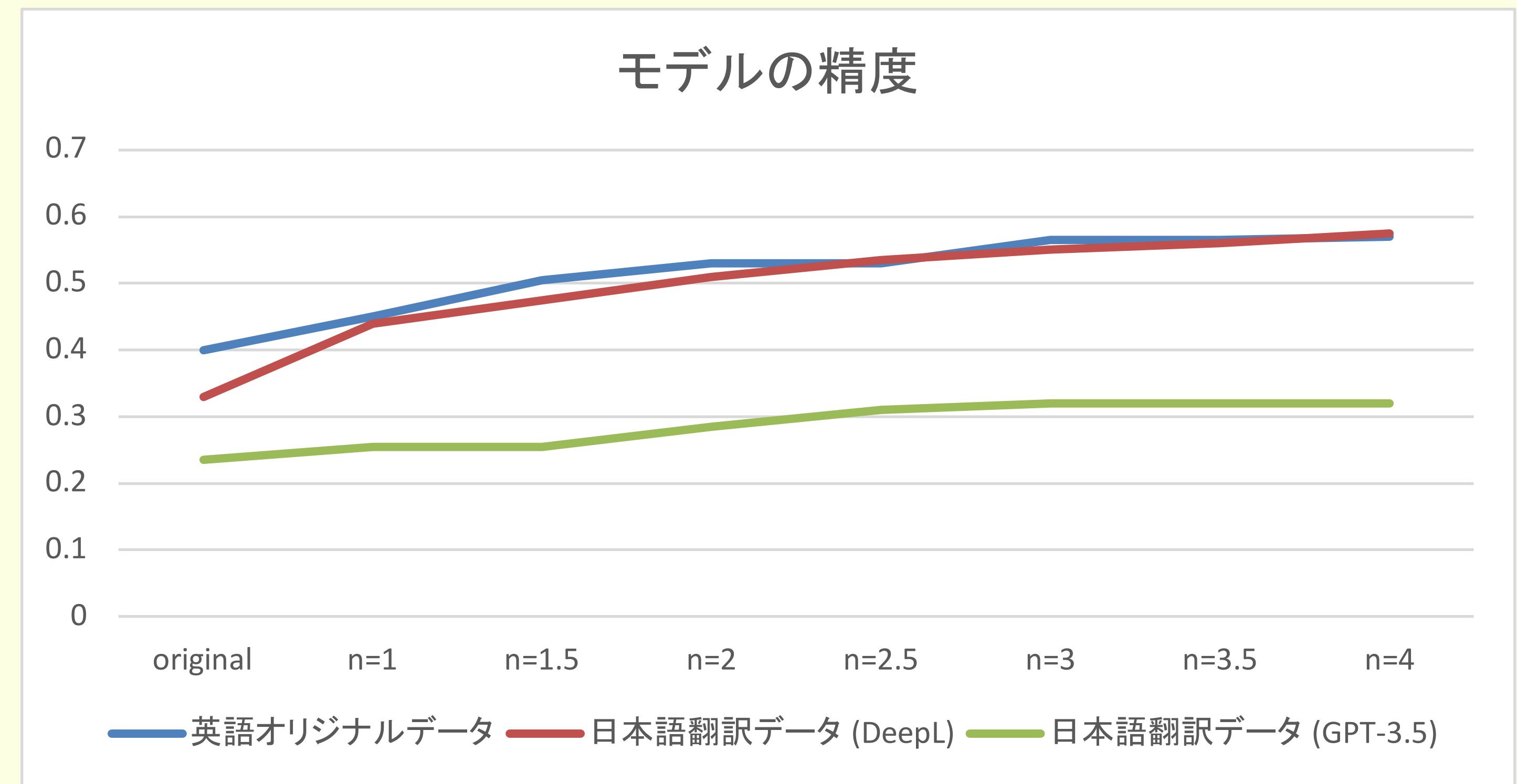
Generate only one sentence with words similar to the example sentence. The length of generated sentence should be similar to the sample sentence. Please generate the sentence in Japanese.

### BERTのファインチューニングの設定

パラメータ	値
学習率 (Learning rate)	1e-5
トークナイザーの最大長 (Max length of tokenizer)	64
データセットの分割比率 (Training:Validation:Testing)	8:1:1
エポック数 (Epochs)	25

## 実験結果

- 学習データとしての疑似データの数を増やすと精度が上がる
- データの偏りをなくすことで精度が上がる
- どのモデルも一定の値まで疑似データ数を増やすと精度が上昇しなくなる
- 英語のオリジナルデータのほうが翻訳した日本語データよりも精度が高い
- DeepLで翻訳したほうがgpt-3.5turboで翻訳するよりも精度が高い



## まとめと今後の課題

- LLMを用いた疑似データを使ってデータ不足を補うことにより、人間の価値観推定モデルの精度が向上した
- 英語の価値観データを日本語に翻訳して価値観モデルを構築する場合、翻訳の精度によって価値観分類モデルの精度に影響を及ぼす

今後の課題としては、LLMによる生成文は類似性が高いものが多いため、より多様性に富んだテキストを生成できるようにすることで価値観推定モデルの精度向上に寄与できないかを検討する