

ChatGPT application in Systematic Literature Reviews in Software Engineering: an evaluation of its accuracy to support the selection activity

ESEM '24, October 24–25, 2024, Barcelona, Spain

背景

SLR：システマティック・リテラチャー・レビュー

- 特定の研究トピックに関する関連文献を検索、選択、統合するプロセス。

項目	通常の文献レビュー	SLR（システマティック・レビュー）
文献選定の基準	明示しないことが多い	明確な基準に基づいて選定
手続き	任意で柔軟	明示された手順に従い再現可能
客観性	主観的	客観的・体系的
再現性	低い	高い

SLRを実施する傾向が強まっている

問題

1. 時間と労力がかかる
2. SLRプロセスのが不十分
3. 質の評価が不十分
4. 初心者研究者にとってSLRプロセスの習得は困難

目的・プローチ

「ChatGPTはSLRプロセス、特に選択活動に関してどの程度サポートできるか」

研究選択フェーズ

- SLRプロセスの最初のステップであり、最も時間がかかる。

目標を達成するために

1. 研究に含まれるデータと除外されるデータのリストから抽出。
2. 選択基準に関する情報を含むプロンプトを用意。
3. もとの研究との評価。

関連研究とその相違

1. Linked Data

- テキストマイニングを半自動化

SluRp

- SLR全体をサポートし，研究グループ内でタスクを共有できる.

ChatGPTには限界がある

- 情報をでっちあげてしまう等...
⇒どの程度サポートできるかに焦点を当てている.

サンプルSLR

SLR1 「人間-コンピュータ相互作用と人工知能の収束」

- 536の研究候補から134の論文をランダムに選択.
- 4人の専門家が3つのスコアを用いて, ベンチマークを作成.
 - 0 – 除外, 1 – 不明, 2 – 該当

SLR2 「ゲームやゲーミフィケーション環境におけるユーザープロファイルの特定と評価」

- 448件の既知の研究コレクションからキーワードやタイトル, 要約を抽出し, 検索文字列を作成し, それが含まれているかどうか.
 - 0 – 除外, 1 – 含む

精度評価のステップ

ステージ 1 入力

- 初期選択のすべての記事のタイトル、要約、およびキーワードを取得.

ステージ 2 一致率を7段階の尺度で評価

- 7段階の尺度
 - 1 – 強く同意しない、2 – そう思わない、3 – ややそう思わない、4 – どちらとも思わない、5 – やや同意、6 – そう思う、7 – 強く同意

ステージ 3 分析(出力を人間の分析と比較)

- 一致はTP, TN, 不一致をFP, FN
- $\text{精度} = \text{TP} + \text{TN} / \text{TP} + \text{TN} + \text{FP} + \text{FN}$

一連の流れ

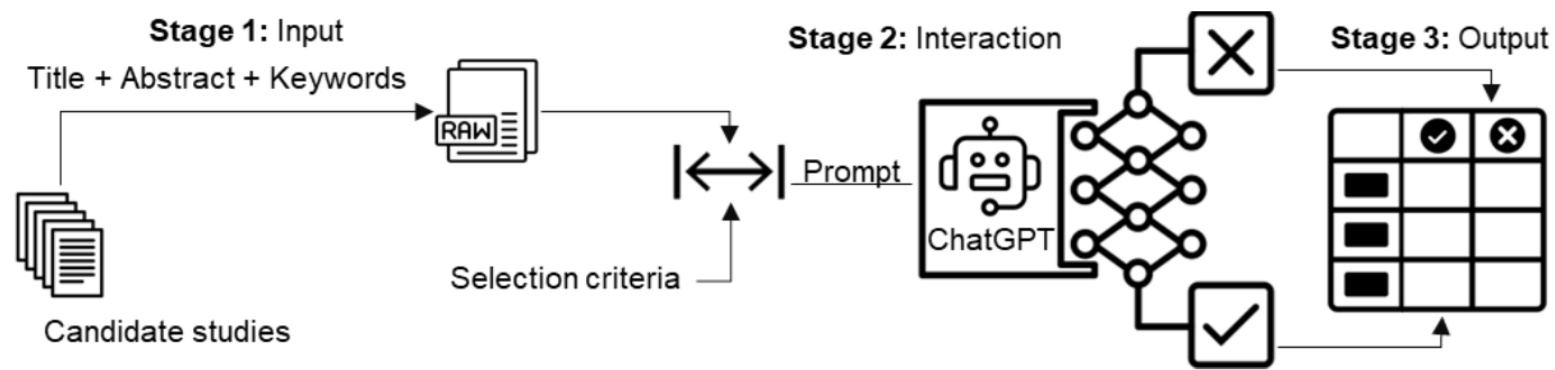
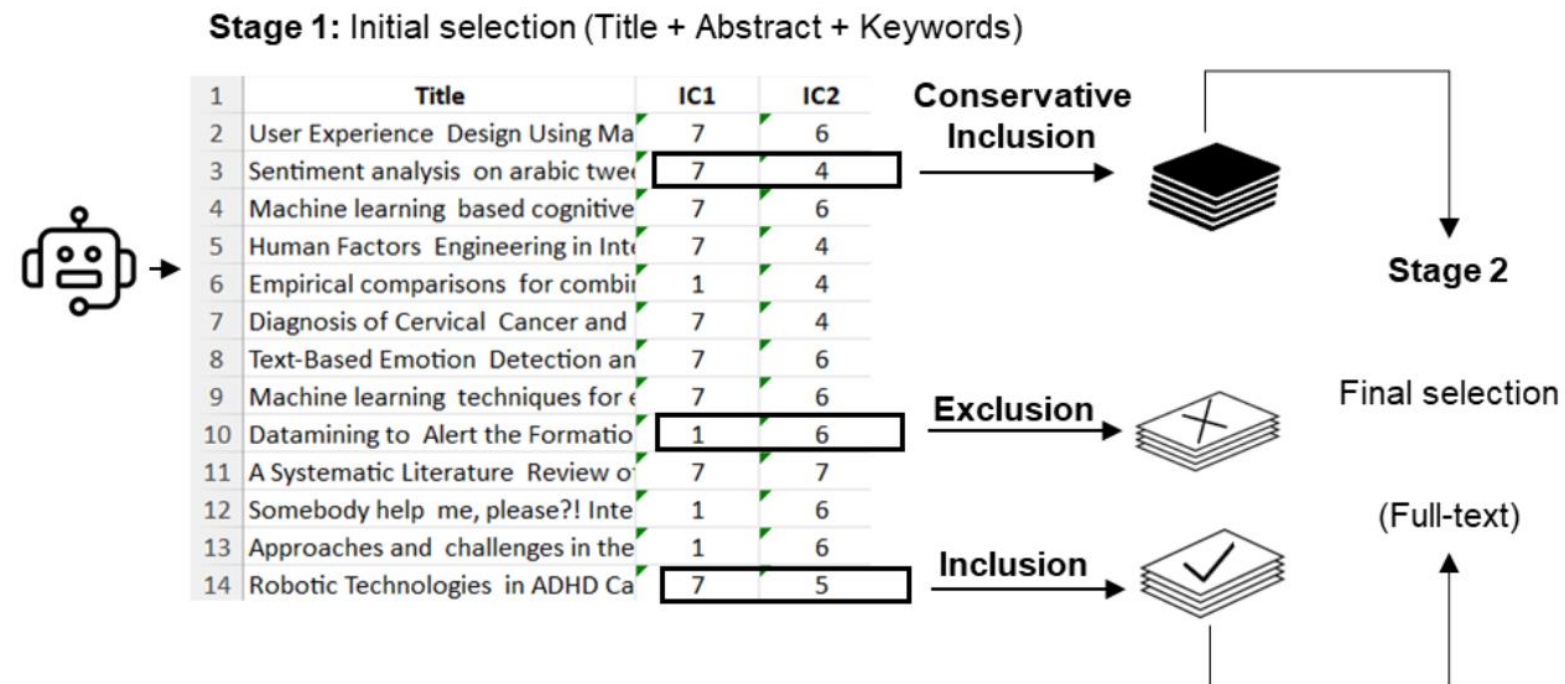


Figure 1: Method research: assessing the accuracy of adopting ChatGPT in selection studies.



結果

Table 4: Comparative of ChatGPT classifications regarding our benchmarks.

Description	Likert ≥ 4 – “Conservative” classification		Likert ≥ 5	
	SLR1	SLR2	SLR1	SLR2
Correct inclusions (TP)	50	128	48	113
Correct exclusions (TN)	35	232	53	273
Incorrect inclusions (FP) – Overwork	35	68	17	27
Incorrect exclusions (FN)	14	20	16	35
Accuracy	63.4%	80.3%	75.3%	86.1%

SLR1

- 正101, 誤33 $k = 0.59$

SLR2

- 正386, 誤62 $k = 0.68$

FNはFPよりもSLRの結果に悪影響

- 二次選択段階での読み込み負担を増加させる。

今後の取り組み

プロンプトに影響を与える特徴の特定
(順序感度、コンテキスト長)

SLRプロセス全体を支えるためのLLMの応答の精度評価

現時点では、75%以上の精度でフィルタリングが可能であるが、AIによる懸念も残されている

