

Light over Heavy: Automated Performance Requirements Quantification with Linguistic Inducement

Shihai Wang, Tao Chen

ICSE 2026

背景 | 要件

要件定義



要件の種類や内容

①機能要件：システムを用いて直接行う振る舞いを定義

-検索機能,登録機能…

②非機能要件：機能以外の品質や制約事項を定義

-性能,セキュリティ,保守性…

性能要件—応答時間,処理能力,稼働状況など

背景|性能要件の数値化

目的

- ・テストでの「**合否判定**」の明確化
- ・関係者間の「**認識のずれ**」を無くするため

課題

- ・**性能の解釈が難しい**

15秒以内とする：1秒でも超えたらNG

15秒で返すとする：少し超えても許容

- ・**手作業での数値化には限界がある**

ーコスト,個人の解釈の違い

⇒アルゴリズムを利用して自動化

従来の手法

ドメイン固有言語 (DSL)

一文章を専用の型に当てはめる

Requirement [ID: R01] { Type: Performance, Action: Search, Trigger: ButtonClick, Metric: ResponseTime, Threshold: $\leq 15s$ }

< 課題 >

- ・ 人間による分析
- ・ 新しい言語としての学習コスト

機械学習/LLM

一要件を読み解かせる

< 課題 >

- ・ 計算コストが高い

軽が重を制す

提案手法 | LQPR

LQPR (Lightweight Quantification Performance Requi)

性能要件の数値化を「分類問題」として定義

- G(Greater) : 値が大きいほど嬉しいパターン
- S(Smaller) : 値が小さいほど嬉しいパターン
- E(Equal) : 区間内ならどれでも同じくらい満足なパターン

<前提>

- ・一つの要件に基準となる数値は基本的に一つ

例 : 「システムは2秒以内に応答すべきだ」

提案手法 | LQPR

LQPR (Lightweight Quantification Performance Requi)

<特徴>

- ・どれだけ複雑な性能要件の文章が来ても,最終的に9パターンのどれかに分類するだけ

⇒シンプルで超軽量

例: 「システムは2秒以内に応答すべきだ」

基準の数値の左側と右側のパーツがG,E,Sどれか分類⇒ $3 \times 3 = 9$ パターン

2秒以内なら評価は同じ⇒E

2秒より後はなるべく早い方がいい⇒S

左はEで右がS⇒クラス〈E,S〉に分類

提案手法 | LQPR

処理手順

①パターン抽出

- ・過去のデータから「～以上」「～未満」といった言葉のテンプレートを集めラベル付け

②パターンと合うか検討

- ・自然言語処理ツール（spaCy）で数値と補助部分を抽出
- ・新しい性能要件の文章が入力されたらテンプレートと比較（文法と意味）

③数値化予測

- ・一番高得点だったテンプレートのクラス＜G,E,S＞を採用

④グラフ化

- ・グラフをパラメータの調整に使う

提案手法 | LQPR

処理手順

- ① パターン抽出
- ② パターンと合うか検討
- ③ 数値化予測
- ④ グラフ化

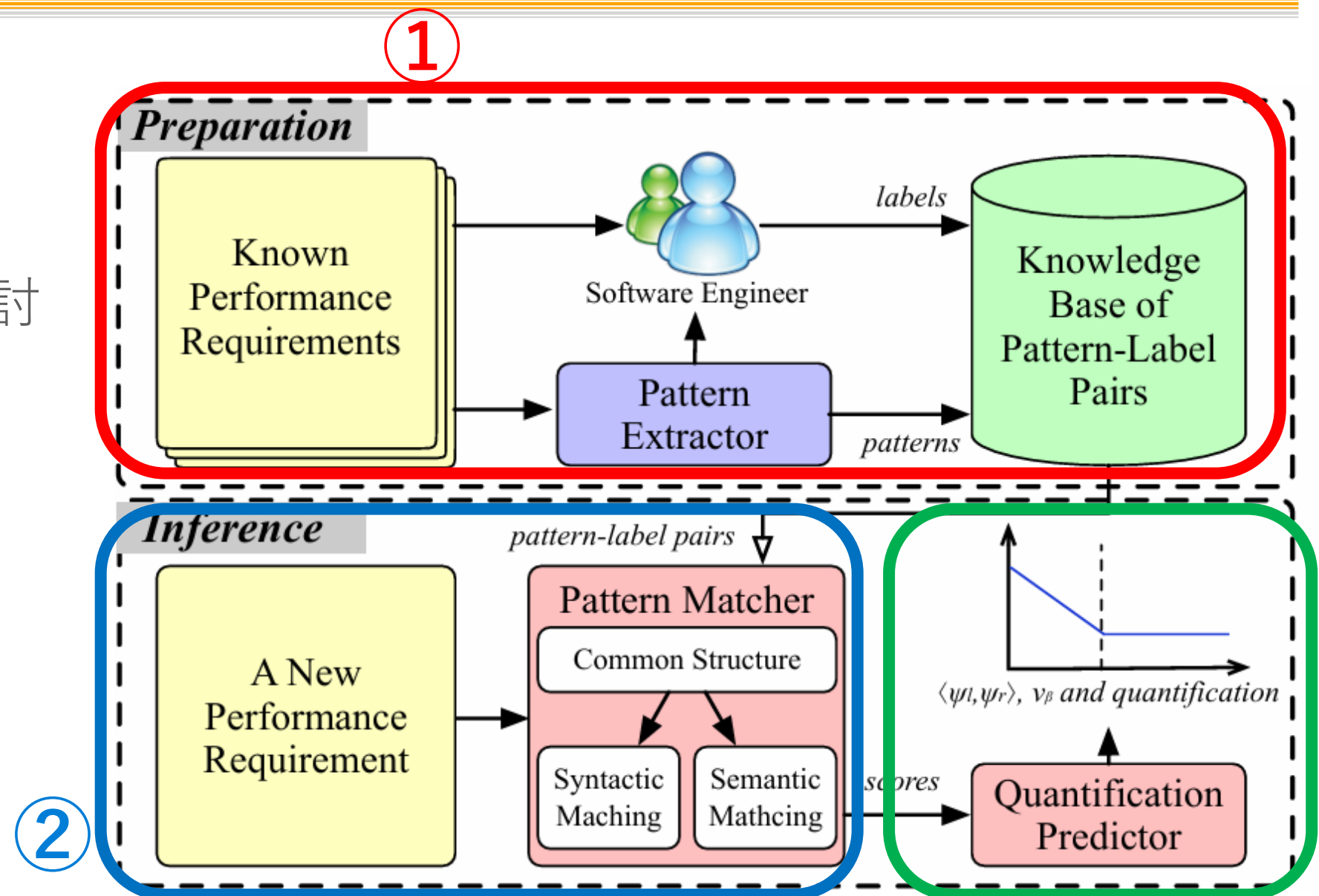


Figure 4: Workflow and architecture of LQPR.

提案手法の工夫点 | LQPR

LCS（最長共通部分系列）：構文スコア

一単語が連続してなくても,順番さえ合っていれば一致とみなせる

登録されている型： be capable of 数値

新しい要件の文： be capable of [handling the existing] 1000 users

⇒同じ型だと判定

ノイズを無視して比較できる：意味スコア

例：「USBを差し込んだ時,システムは100分未満でデプロイされ...」

ノイズ：「USB」「デプロイ」

- ・ AI：文章丸ごとをAIでベクトル化して意味を比べるとノイズに意味が引っ張られる

- ・ LQPR：核となるLCSだけをWord2Vecで比較

RQ

RQ1：データセット内においてLQPRの性能は最先端アプローチと比べてどうか？

RQ2：LQPRは、データセットをまたいだ性能要件に対して、どの程度汎化できるか？

RQ3：LQPRの各設計要素は、全体としてどのような貢献をしているか？

RQ4：LQPRの効率性はどうか？

データセット

選定条件

- ・ 正式な要件抽出プロセスが含まれていること
- ・ 性能関連かつ数値化可能であること
- ・ 要件が十分な数（10個以上）含まれていること

Table 1: Datasets and projects studied.

Dataset	# Projects	# Perf. Requirements	Source
PROMISE [1]	15	259	real-world
PURE [34]	79	23	real-world
Shaukat et al. [55]	4	15	real-world
FUNCTIONAL-QUALITY [11]	5	10	real-world
LLM-GEN	N/A	100	synthetic

比較対象（10個のライバル）

伝統的な機械学習

エンコーダー専用LLM（BERTなど）

デコーダー専用LLM（Gemma,Llama）

RQの答え

RQ1：データセット内においてLQPRの性能は最先端アプローチと比べてどうか？

⇒圧勝1位

RQ2：LQPRは,データセットをまたいだ性能要件に対して,どの程度汎化できるか？

⇒首位をキープ

RQ3：LQPRの各設計要素は,全体としてどのような貢献をしているか？

⇒すべての機能が必要

RQ4：LQPRの効率性はどうか？

⇒時間も目盛りも約100分の1

実験結果

RQ1

Table 2: Classifying and quantifying performance requirements from the same known dataset. We report on the mean and standard deviation (SD) over 30 runs. r denotes Scott-Knott rank; orange cells indicate the best on a metric. The mean results better than 0.8 are highlighted in bold.

Approach	wP		wR		wF_1	
	r	Mean (SD)	r	Mean (SD)	r	Mean (SD)
BoW/NB	3	0.680 (0.063)	2	0.730 (0.048)	3	0.680 (0.058)
BoW/ k NN	3	0.690 (0.050)	3	0.680 (0.047)	3	0.650 (0.051)
TF-IDF/NB	4	0.610 (0.046)	2	0.720 (0.042)	3	0.650 (0.047)
TF-IDF/ k NN	3	0.680 (0.042)	3	0.690 (0.037)	3	0.670 (0.041)
BERT	2	0.840 (0.051)	1	0.850 (0.036)	2	0.830 (0.043)
PRCBERT	4	0.603 (0.051)	2	0.727 (0.042)	3	0.648 (0.091)
ZSL	4	0.590 (0.158)	5	0.390 (0.044)	5	0.400 (0.047)
Gemma-27B	4	0.600 (3.128)	4	0.590 (3.059)	4	0.590 (2.666)
Deepseek-67B	4	0.590 (2.660)	4	0.590 (2.983)	4	0.590 (3.115)
Llama-8B	4	0.600 (2.518)	4	0.590 (2.202)	4	0.600 (2.988)
LQPR	1	0.861 (0.024)	1	0.858 (0.026)	1	0.853 (0.025)

実験結果

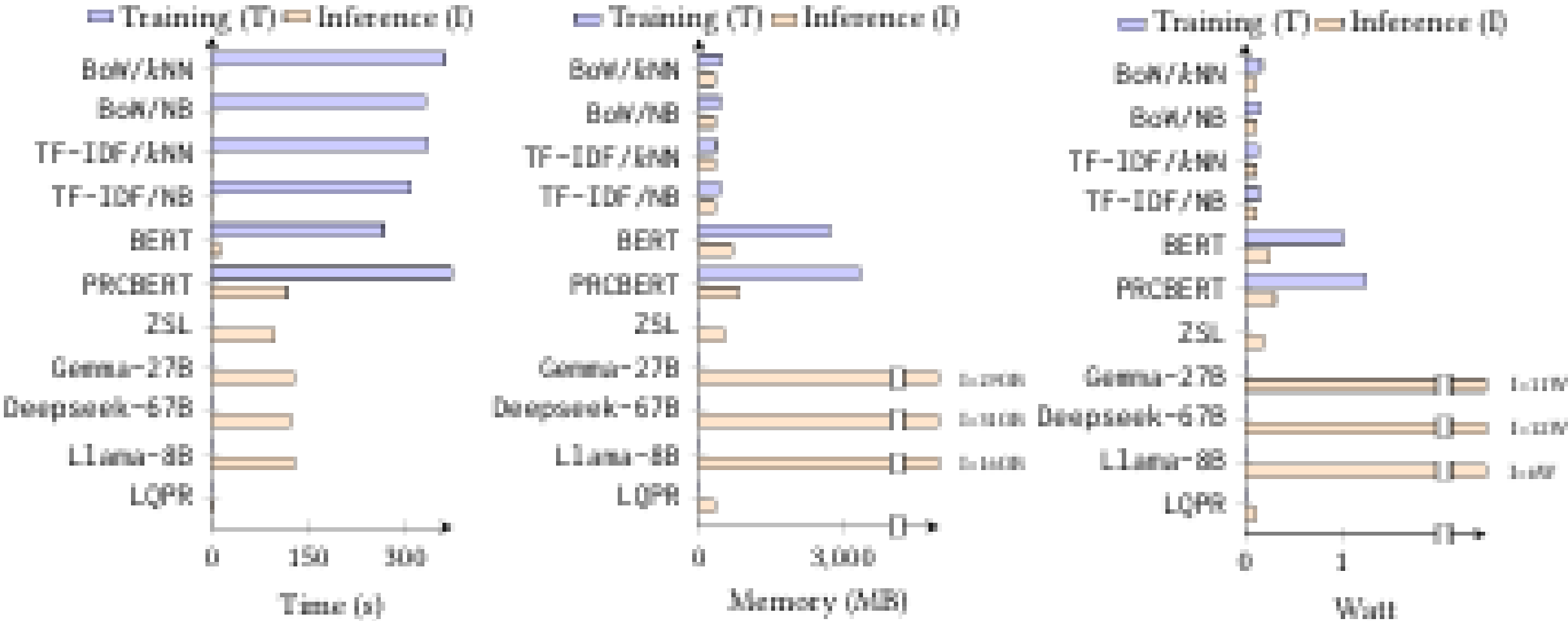
RQ2

Table 3: Classifying and quantifying performance requirements from unforeseen datasets. The format is as Table 2.

Approach	wP		wR		wF_1	
	r	Mean (SD)	r	Mean (SD)	r	Mean (SD)
<i>PURE Dataset</i>						
BoW/NB	2	0.830 (0.056)	3	0.810 (0.043)	2	0.820 (0.044)
BoW/ k NN	4	0.750 (0.068)	5	0.690 (0.073)	5	0.700 (0.066)
TF-IDF/NB	3	0.790 (0.035)	3	0.820 (0.033)	3	0.800 (0.032)
TF-IDF/ k NN	2	0.820 (0.042)	4	0.780 (0.047)	3	0.790 (0.041)
BERT	2	0.820 (0.033)	2	0.850 (0.039)	2	0.830 (0.042)
PROBERT	5	0.678 (0.053)	5	0.743 (0.082)	5	0.714 (0.039)
ZSL	5	0.249 (0.000)	7	0.197 (0.000)	6	0.220 (0.000)
Gemma-27B	3	0.802 (0.000)	3	0.808 (0.000)	4	0.775 (0.000)
Deepseek-67B	4	0.785 (0.000)	5	0.705 (0.000)	4	0.760 (0.000)
Llama-8B	3	0.792 (0.000)	6	0.634 (0.000)	5	0.649 (0.000)
LQPR	1	0.951 (0.000)	1	0.947 (0.000)	1	0.946 (0.000)
<i>Shaukat et al. Dataset</i>						
BoW/NB	3	0.800 (0.051)	2	0.820 (0.046)	2	0.800 (0.045)
BoW/ k NN	4	0.770 (0.042)	5	0.630 (0.103)	5	0.640 (0.097)
TF-IDF/NB	3	0.790 (0.043)	2	0.820 (0.034)	3	0.790 (0.034)
TF-IDF/ k NN	4	0.750 (0.059)	4	0.740 (0.056)	4	0.740 (0.051)
BERT	3	0.790 (0.041)	2	0.830 (0.034)	2	0.810 (0.036)
PROBERT	4	0.732 (0.061)	4	0.728 (0.040)	4	0.736 (0.059)
ZSL	5	0.210 (0.000)	6	0.168 (0.000)	6	0.187 (0.000)
Gemma-27B	2	0.846 (0.000)	3	0.807 (0.000)	2	0.823 (0.000)
Deepseek-67B	2	0.857 (0.000)	4	0.709 (0.000)	3	0.761 (0.000)
Llama-8B	3	0.806 (0.000)	5	0.672 (0.000)	5	0.674 (0.000)
LQPR	1	1.000 (0.000)	1	1.000 (0.000)	1	1.000 (0.000)
<i>FUNCTIONAL-QUALITY Dataset</i>						
BoW/NB	6	0.256 (0.276)	6	0.350 (0.086)	7	0.220 (0.135)
BoW/ k NN	5	0.559 (0.166)	4	0.580 (0.127)	6	0.513 (0.110)
TF-IDF/NB	2	0.918 (0.051)	1	0.877 (0.086)	2	0.880 (0.083)
TF-IDF/ k NN	2	0.871 (0.120)	2	0.803 (0.130)	3	0.816 (0.128)
BERT	4	0.681 (0.227)	2	0.793 (0.123)	4	0.717 (0.178)
PROBERT	3	0.772 (0.455)	2	0.825 (0.350)	3	0.784 (0.431)
ZSL	1	1.000 (0.000)	5	0.500 (0.000)	5	0.611 (0.000)
Gemma-27B	3	0.802 (0.000)	2	0.808 (0.000)	3	0.775 (0.000)
Deepseek-67B	3	0.785 (0.000)	3	0.705 (0.000)	3	0.760 (0.000)
Llama-8B	3	0.792 (0.000)	4	0.634 (0.000)	5	0.649 (0.000)
LQPR	1	1.000 (0.000)	1	0.900 (0.000)	1	0.946 (0.000)
<i>LLM-GEN Dataset</i>						
BoW/NB	4	0.780 (0.035)	4	0.680 (0.054)	4	0.720 (0.049)
BoW/ k NN	4	0.740 (0.048)	5	0.380 (0.051)	5	0.360 (0.073)
TF-IDF/NB	4	0.780 (0.026)	3	0.750 (0.075)	3	0.750 (0.068)
TF-IDF/ k NN	4	0.760 (0.046)	4	0.690 (0.088)	4	0.710 (0.076)
BERT	2	0.870 (0.022)	1	0.880 (0.016)	2	0.870 (0.018)
PROBERT	3	0.841 (0.091)	2	0.857 (0.035)	2	0.828 (0.077)
ZSL	5	0.215 (0.000)	6	0.190 (0.000)	6	0.202 (0.000)
Gemma-27B	2	0.901 (0.000)	2	0.840 (0.000)	2	0.864 (0.000)
Deepseek-67B	3	0.863 (0.000)	3	0.710 (0.000)	3	0.770 (0.000)
Llama-8B	2	0.879 (0.000)	4	0.700 (0.000)	4	0.685 (0.000)
LQPR	1	0.945 (0.000)	1	0.880 (0.000)	1	0.906 (0.000)

実験結果

RQ3



(a) Clock time

(b) Memory

(c) Power

実験結果

RQ4

Table 4: Ablation analysis of LQPR by excluding different designs one at a time. Other formats are the same as Table 2.

Approach	wP		wR		wF_1	
	r	Mean (SD)	r	Mean (SD)	r	Mean (SD)
<i>Promise Dataset</i>						
LQPR-L	3	0.787 (0.032)	2	0.752 (0.040)	2	0.747 (0.042)
LQPR-se	2	0.827 (0.033)	3	0.460 (0.033)	3	0.563 (0.033)
LQPR-sy	1	0.875 (0.021)	1	0.862 (0.024)	1	0.865 (0.023)
LQPR	1	0.861 (0.024)	1	0.858 (0.026)	1	0.853 (0.025)
<i>PURE Dataset</i>						
LQPR-L	2	0.895 (0.000)	2	0.895 (0.000)	2	0.895 (0.000)
LQPR-se	1	0.952 (0.000)	3	0.473 (0.000)	3	0.588 (0.000)
LQPR-sy	3	0.857 (0.000)	2	0.895 (0.000)	2	0.870 (0.000)
LQPR	1	0.951 (0.000)	1	0.947 (0.000)	1	0.946 (0.000)
<i>Shaukat et al. Dataset</i>						
LQPR-L	2	0.950 (0.000)	2	0.933 (0.000)	2	0.937 (0.000)
LQPR-se	2	0.942 (0.000)	3	0.600 (0.000)	4	0.694 (0.000)
LQPR-sy	3	0.872 (0.000)	2	0.933 (0.000)	3	0.901 (0.000)
LQPR	1	1.000 (0.000)	1	1.000 (0.000)	1	1.000 (0.000)
<i>FUNCTIONAL-QUALITY Dataset</i>						
LQPR-L	2	0.925 (0.000)	2	0.800 (0.000)	2	0.840 (0.000)
LQPR-se	3	0.700 (0.000)	3	0.300 (0.000)	3	0.420 (0.000)
LQPR-sy	2	0.925 (0.000)	1	0.900 (0.000)	1	0.903 (0.000)
LQPR	1	1.000 (0.000)	1	0.900 (0.000)	1	0.946 (0.000)
<i>LLM-Gen Dataset</i>						
LQPR-L	2	0.761 (0.000)	3	0.580 (0.000)	3	0.617 (0.000)
LQPR-se	1	0.952 (0.000)	2	0.750 (0.000)	2	0.827 (0.000)
LQPR-sy	1	0.942 (0.000)	1	0.870 (0.000)	1	0.896 (0.000)
LQPR	1	0.945 (0.000)	1	0.880 (0.000)	1	0.906 (0.000)

考察

なぜLQPRは統計的学習を凌駕するのか？

- ・ 性能要件の文章は短すぎ、文脈を読み取れない
- ・ 9マス分類がドメイン知識すぎる
- ・ 期待値の記述を持たない要件に弱い

LQPRの弱点

- ・ 未知の型に対応できない⇒辞書をパワーアップで対応

結論

- ・ なんでもAIが最適とは限らない、場合によっては軽量な手法が勝る
- ・ 高精度と低コストの両立