Are Small Language Models Ready to Compete with Large Language Models for Practical Applications?

CONTACT

Neelabh Sinha, Vinija Jain, Aman Chadha

Georgia Tech · Meta · amazon

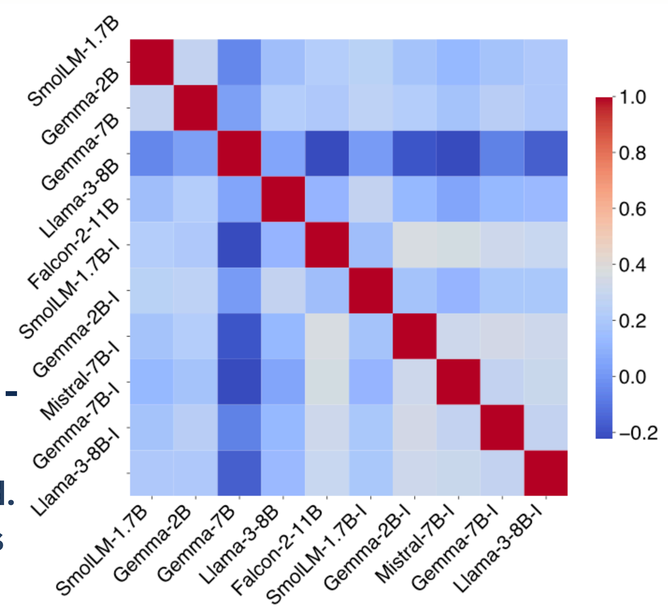
MOTIVATION

Issues with using Large Language Models -

- High costs of inference.
- Large compute requirements for hosting.
- Limited API Access.
- Concerns around data privacy.
- High variance in performance of LMs.

Challenges of replacing them by Small LMs -

- High variance in performance due to architecture, training data, and training method.
- May not perform well globally due to limitations of scale.
- Prompting them is tricky.



Correlation between performance of small LMs depicting high variability between their performance.

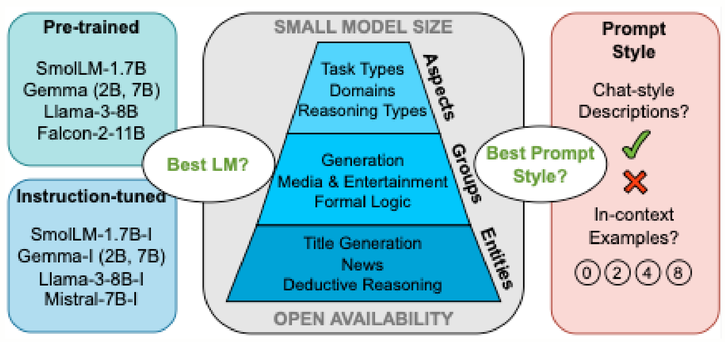
Can small, open LMs compete with large, proprietary LMs in practical usage?

What can be an exhaustive evaluation framework to conduct this analysis?

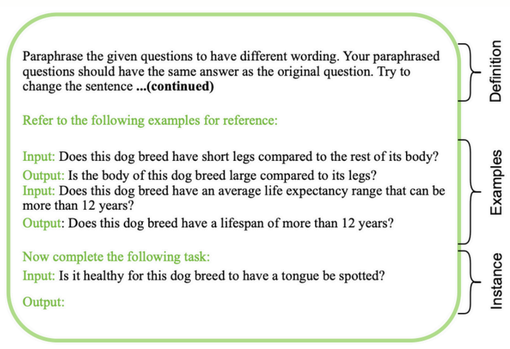
How do current best small, open LMs perform in comparison?

What type of prompt style should be used to extract best results from these LMs?

EXPERIMENTAL DESIGN



Summary of Experimental Settings



Example Prompt

Task Types
- Semantic and Pragmatic Analysis
  - Textual Entailment
  - Coreference Resolution
  - Dialogue Act Recognition
  - Overlap Extraction
- Generation
  - Data to Text
  - Title Generation
  - Question Rewriting
- Classification and Recognition
  - Answerability Classification
  - Cause-effect Classification
  - Keyword Tagging
- Linguistic Relationship
  - Word Analogy
  - Grammar Error Correction

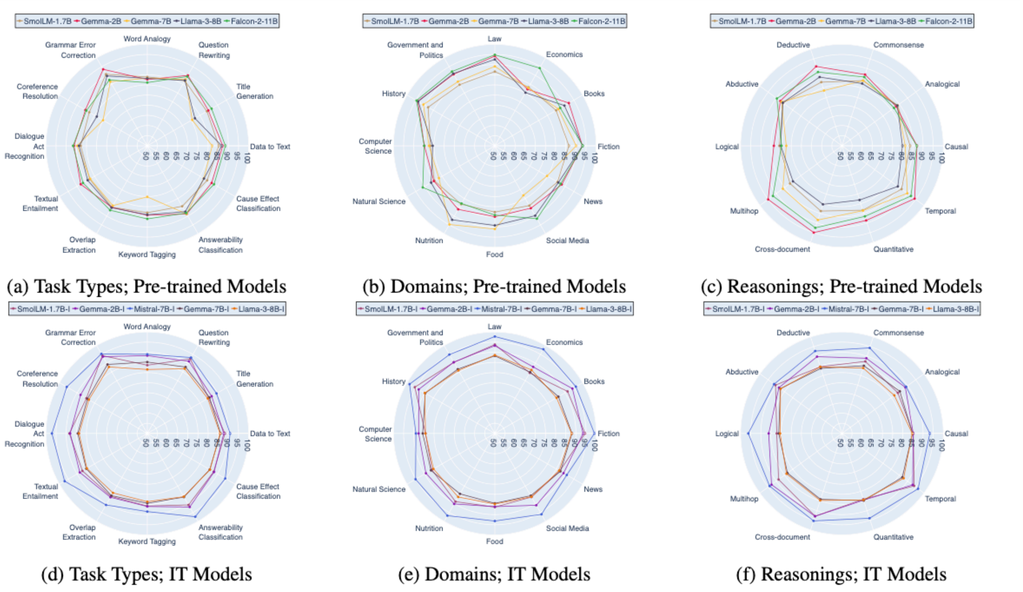
Domains
- Social Sciences and Humanities
  - Economics
  - Government and Politics
  - History
  - Law
- Science and Technology
  - Natural Science
  - Computer Science
- Health and Medical
  - Nutrition
  - Food
- Art and Literature
  - Fiction
  - Books
- Media and Entertainment
  - Social Media
  - News

Reasoning Types
- Formal Logic
  - Deductive Reasoning
  - Logical Reasoning
  - Abductive Reasoning
- Comparative and Relational
  - Analogical Reasoning
  - Causal Reasoning
  - Commonsense Reasoning
- Complex Inference and Analysis
  - Cross-document Reasoning
  - Multihop Reasoning
- Contextual Reasoning
  - Quantitative Reasoning
  - Temporal Reasoning

Categorization Taxonomy

- Created a three-tier taxonomy of applications for analyzing performance across hierarchies based on SuperNatural Instructions dataset.
- Used 5 pre-trained and 5 instruction-tuned SLMs across different families, range of parameters for an exhaustive comparison.
- Experimented with 8 prompt styles by varying chat style descriptions, in context examples to identify best ways to use an LM.
- Measured performance using BERTScore recall.

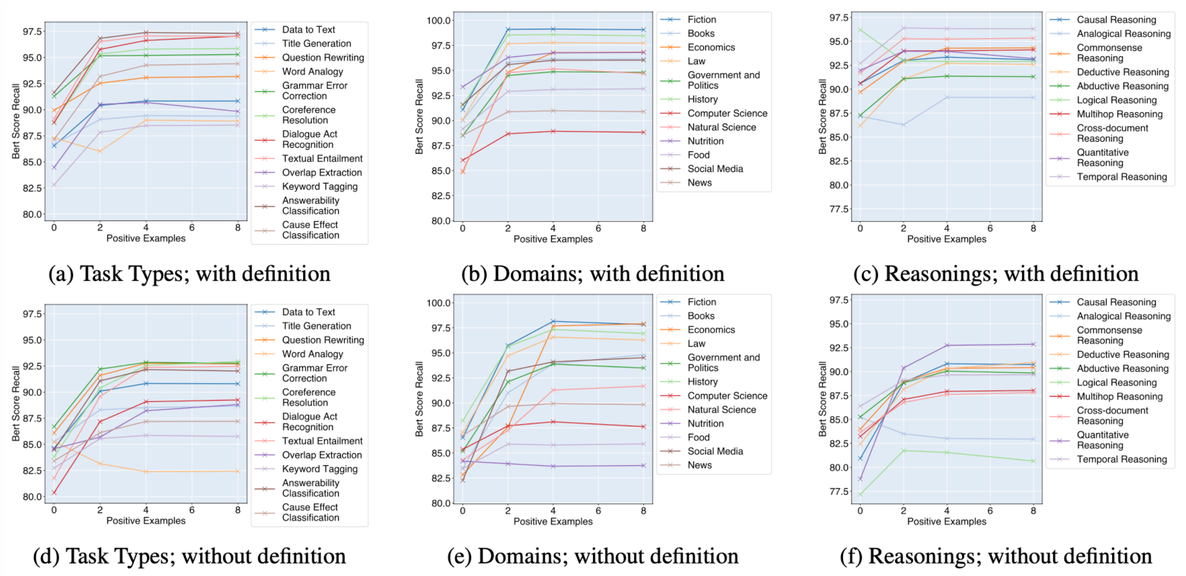
RESULTS



(a) Task Types; Pre-trained Models (b) Domains; Pre-trained Models (c) Reasonings; Pre-trained Models

(d) Task Types; IT Models (e) Domains; IT Models (f) Reasonings; IT Models

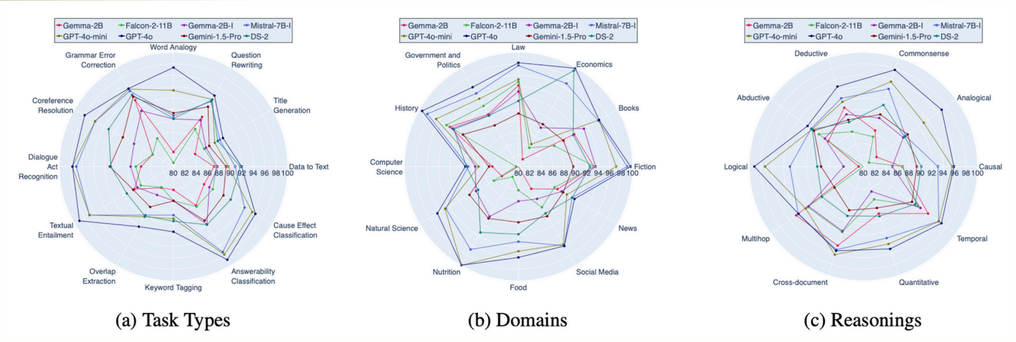
Mean BERTScore recall across various task types, domains, and reasoning types, segmented by pre-trained vs. instruction-tuned models.



(a) Task Types; with definition (b) Domains; with definition (c) Reasonings; with definition

(d) Task Types; without definition (e) Domains; without definition (f) Reasonings; without definition

Mean BERTScore recall for different prompt styles for Mistral-7B-I (best performing) segmented by task types, domains, and reasoning types.



(a) Task Types (b) Domains (c) Reasonings

Mean BERTScore recall values of top 2 models in both compared against GPT-4o, GPT-4o-mini and Gemini 1.5 Pro.

Model Name	Ex.	Def	Par. Def.
Gemma-2B	4	86.41	85.77
Falcon-2-11B	8	86.18	86.00
Gemma-2B-I	4	87.96	87.67
Mistral-7B-I	8	93.76	93.22

Mean BERTScore recall for actual task definitions (from dataset) v/s Paraphrased task definitions (using GPT 3.5) for top 2 models.

FINDINGS AND CONCLUSION

- Performance is very dependent on task type, application domain, and reasoning type, and different LMs have different strengths.
- Gemma-2B and Falcon-2-11B prove to be the best pre-trained models. Performance doesn't depend on scale of the models.
- Mistral-7B-I proves to be the best instruction-tuned model, outperforming Gemini-1.5-Pro by 4.94% and worse than GPT-4o by only 2.12% overall.
- Best prompt style also varies with task types, application domains and knowledge types, and performance is robust to dataset biases.