Are You Building a Taxonomy? Are You Analyzing LLMs' Weaknesses? Here is a Better Way!

COMPREHEND: Ultra-Fine-Grained Benchmark Understanding by Entailment Hierarchy

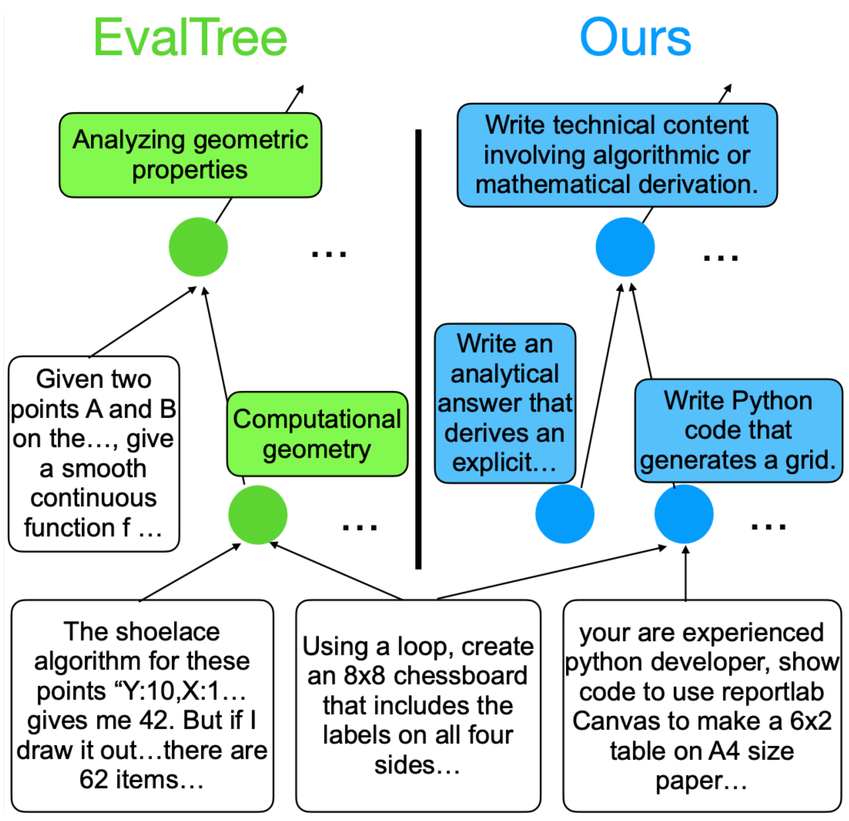
College of Information Science

Sivaraaman Balakrishnan*, Prakriti Shankar Shetty*, Jeremiah Hong, Hamed Zamani, and Haw-Shiuan Chang*

UMASS AMHERST

Introduction

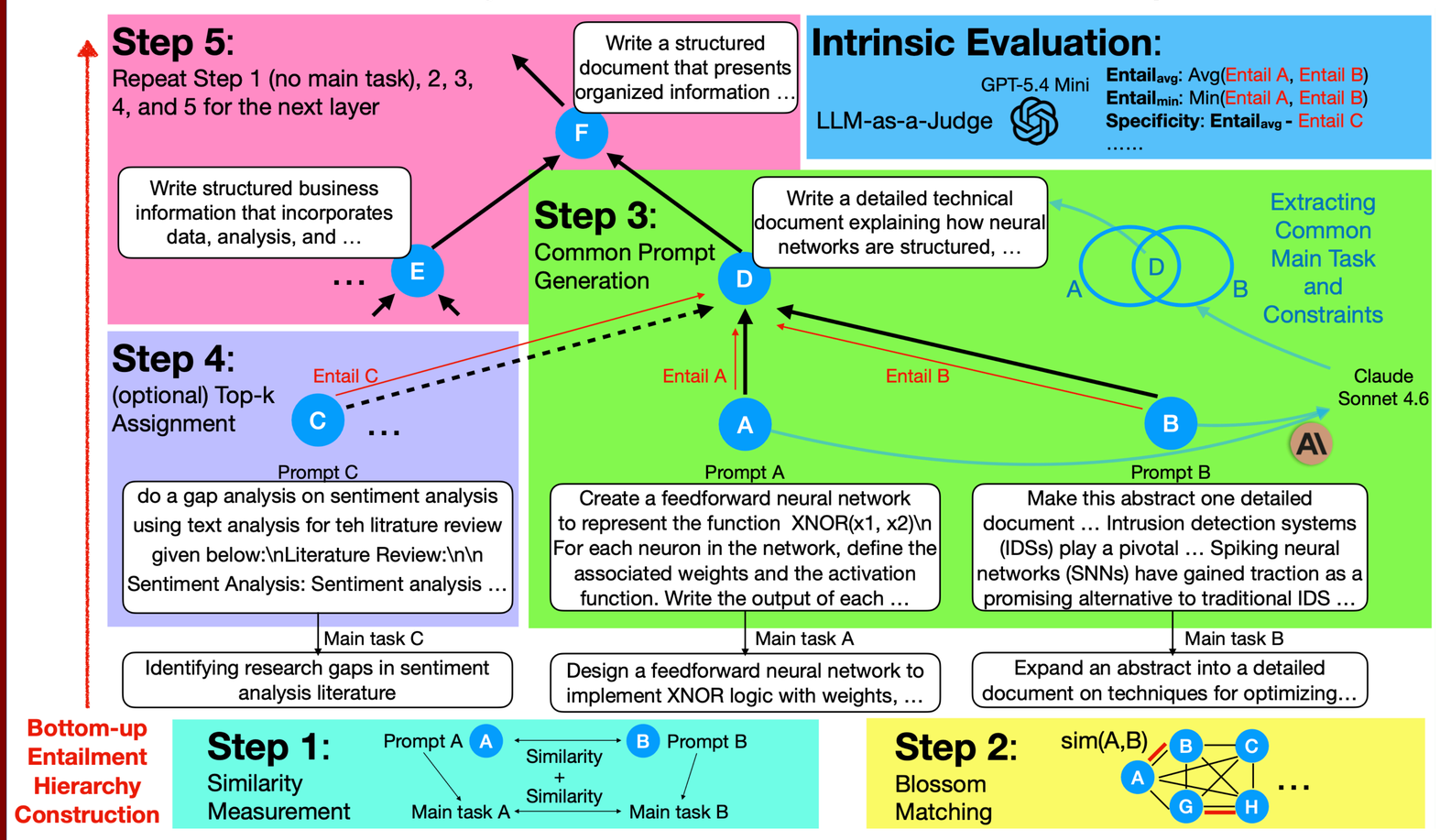
The existing taxonomy uses short phrases to describe each cluster, which is too coarse-grained.



User Prompts in WildBench

Our Taxonomy Construction Method

From bottom to top, we iteratively match similar nodes and ask an LLM to generate the common summary of two similar nodes as the cluster description.

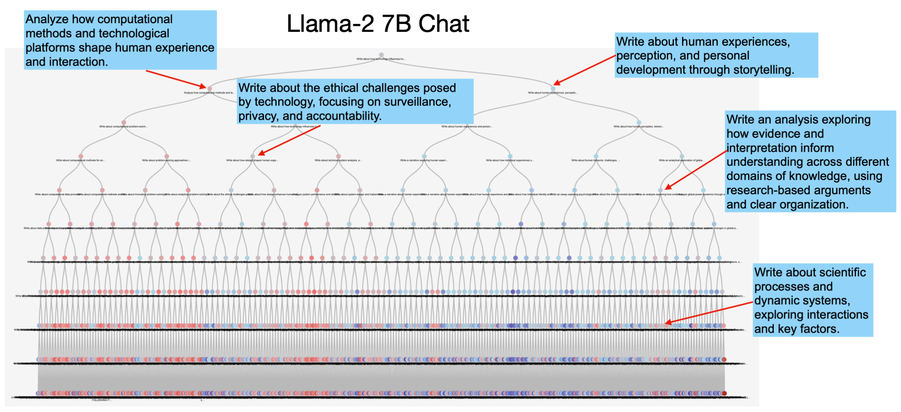


Experiment Results

- **We conducted very detailed ablation study to justify our proposed steps.**
- **We can construct an ultra-fine-grained taxonomy to analyze the weaknesses of LLMs.**
- **We show that COMPREHEND is much better than EvalTree in all quality metrics.**
- **The improvement comes from common summary generation and integrating matching and generation.**
- **The clustering in EvalTree creates an imbalanced tree.**
- **An LLM can write better weakness report from COMPREHEND.**

Methods	Clustering	Leaf Main Task	Parent Generation
EvalTree	Recursive Kmeans	Capability	Capability
EvalTree + Our Prompts	Recursive Kmeans	Main Task	Common Prompt
EvalTree + Our Prompts + Org.	Recursive Kmeans	Original	Common Prompt
COMPREHEND - Main - Eval	Match	Original	Common Prompt
COMPREHEND - Main	Match + Assignment	Original	Common Prompt + Regen.
COMPREHEND	Match + Main + Assignment	Original	Common Prompt + Regen.
COMPREHEND on Capability	Match + Main + Assignment	Capability	Common Prompt + Regen.
COMPREHEND on Main Task	Match + Main + Assignment	Main Task	Common Prompt + Regen.
EvalTree + Long Prompt	Recursive Kmeans	Long Capability	Long Capability
EvalTree (MiniLM)	Recursive Kmeans	Capability	Capability
EvalTree + Long Prompt (MiniLM)	Recursive Kmeans	Long Capability	Long Capability

Table 1: Differences of our ablation methods. The Assignment in clustering means Step 4. The Main means extracting the main task of the leaf nodes to improve similarity measurement in Step 1.



Data	Methods	$Entail_{avg}$	$Entail_{min}$	$Info_{avg}$	$Info_{min}$	Specificity	Len (words)
WB256	EvalTree	$36.3_{\pm 1.7}$	$22.6_{\pm 1.6}$	$22.9_{\pm 1.0}$	$16.3_{\pm 0.9}$	$13.0_{\pm 2.2}$	$5.7_{\pm 0.1}$
	EvalTree + Our Prompts	$47.1_{\pm 1.6}$	$32.9_{\pm 1.8}$	$36.3_{\pm 1.4}$	$26.3_{\pm 1.4}$	$25.4_{\pm 2.1}$	$15.8_{\pm 0.3}$
	EvalTree + Our Prompts + Org.	$50.4_{\pm 1.7}$	$35.5_{\pm 1.9}$	$41.0_{\pm 1.6}$	$31.6_{\pm 1.6}$	$23.8_{\pm 2.2}$	$21.4_{\pm 0.9}$
	COMPREHEND - Main - Eval	$53.5_{\pm 1.7}$	$38.8_{\pm 1.9}$	$44.2_{\pm 1.5}$	$34.9_{\pm 1.6}$	$26.4_{\pm 2.4}$	$21.3_{\pm 1.0}$
	COMPREHEND - Main	$55.4_{\pm 1.7}$	$40.7_{\pm 2.0}$	$45.6_{\pm 1.6}$	$36.0_{\pm 1.7}$	$29.3_{\pm 2.3}$	$21.1_{\pm 0.9}$
	COMPREHEND	$\mathbf{55.9}_{\pm 1.7}$	$\mathbf{41.7}_{\pm 1.9}$	$\mathbf{49.1}_{\pm 1.5}$	$\mathbf{39.1}_{\pm 1.7}$	$\mathbf{32.5}_{\pm 2.3}$	$21.6_{\pm 1.0}$
	COMPREHEND on Capability	$50.8_{\pm 1.7}$	$35.7_{\pm 1.8}$	$37.8_{\pm 1.3}$	$27.9_{\pm 1.2}$	$24.6_{\pm 2.1}$	$12.3_{\pm 0.2}$
	COMPREHEND on Main Task	$50.4_{\pm 1.6}$	$35.7_{\pm 1.8}$	$37.9_{\pm 1.4}$	$28.6_{\pm 1.4}$	$27.8_{\pm 2.3}$	$15.3_{\pm 0.3}$
	EvalTree + Long Prompt	$29.9_{\pm 1.4}$	$17.6_{\pm 1.3}$	$41.7_{\pm 1.4}$	$29.5_{\pm 1.5}$	$17.3_{\pm 1.7}$	$19.4_{\pm 0.1}$
	EvalTree (MiniLM)	$39.7_{\pm 1.9}$	$24.8_{\pm 1.9}$	$21.2_{\pm 1.0}$	$14.0_{\pm 0.8}$	$17.7_{\pm 2.2}$	$5.5_{\pm 0.1}$
	EvalTree + Long Prompt (MiniLM)	$30.6_{\pm 1.5}$	$17.5_{\pm 1.3}$	$37.4_{\pm 1.4}$	$26.8_{\pm 1.5}$	$18.0_{\pm 1.6}$	$19.5_{\pm 0.1}$
UF256	EvalTree	$36.2_{\pm 1.9}$	$23.1_{\pm 1.9}$	$26.1_{\pm 1.4}$	$18.5_{\pm 1.2}$	$14.4_{\pm 2.4}$	$5.4_{\pm 0.1}$
	EvalTree + Our Prompts	$41.5_{\pm 1.7}$	$26.9_{\pm 1.7}$	$31.6_{\pm 1.4}$	$22.0_{\pm 1.3}$	$20.0_{\pm 2.1}$	$12.5_{\pm 0.3}$
	EvalTree + Our Prompts + Org.	$52.0_{\pm 1.8}$	$37.6_{\pm 2.0}$	$36.8_{\pm 1.5}$	$27.3_{\pm 1.5}$	$24.0_{\pm 2.4}$	$14.5_{\pm 0.4}$
	COMPREHEND - Main - Eval	$\mathbf{55.6}_{\pm 1.7}$	$41.1_{\pm 1.9}$	$41.1_{\pm 1.5}$	$31.8_{\pm 1.6}$	$28.4_{\pm 2.4}$	$14.3_{\pm 0.3}$
	COMPREHEND - Main	$54.4_{\pm 1.7}$	$40.3_{\pm 2.0}$	$43.7_{\pm 1.6}$	$34.5_{\pm 1.6}$	$31.1_{\pm 2.5}$	$14.5_{\pm 0.3}$
	COMPREHEND	$55.3_{\pm 1.7}$	$\mathbf{41.8}_{\pm 2.0}$	$\mathbf{44.5}_{\pm 1.6}$	$\mathbf{35.5}_{\pm 1.7}$	$\mathbf{31.4}_{\pm 2.4}$	$14.0_{\pm 0.4}$
	COMPREHEND on Capability	$49.2_{\pm 1.7}$	$32.9_{\pm 1.9}$	$36.9_{\pm 1.4}$	$26.4_{\pm 1.3}$	$22.1_{\pm 2.4}$	$10.8_{\pm 0.2}$
	COMPREHEND on Main Task	$46.9_{\pm 1.8}$	$33.0_{\pm 2.0}$	$33.8_{\pm 1.5}$	$25.7_{\pm 1.4}$	$25.7_{\pm 2.2}$	$12.1_{\pm 0.3}$
	EvalTree + Long Prompt	$30.2_{\pm 1.6}$	$19.9_{\pm 1.5}$	$42.8_{\pm 1.7}$	$31.9_{\pm 1.7}$	$14.9_{\pm 1.9}$	$18.9_{\pm 0.1}$
	EvalTree (MiniLM)	$37.6_{\pm 1.9}$	$24.5_{\pm 1.9}$	$25.5_{\pm 1.3}$	$18.0_{\pm 1.2}$	$15.9_{\pm 2.4}$	$5.6_{\pm 0.1}$
	EvalTree + Long Prompt (MiniLM)	$31.6_{\pm 1.6}$	$19.3_{\pm 1.5}$	$40.7_{\pm 1.6}$	$28.8_{\pm 1.7}$	$18.2_{\pm 1.8}$	$18.9_{\pm 0.1}$

Table 2: Comparison of EvalTree, COMPREHEND, and its ablations on 256 randomly sampled prompts from WildBench (WB) and UltraFeedback (UF). Values are mean ± standard error. Entailment and informativeness are percentages, Specificity is in percentage points, and Len is in words. The best score for each metric within each dataset is highlighted.

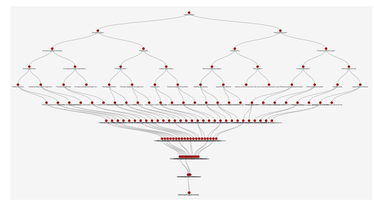
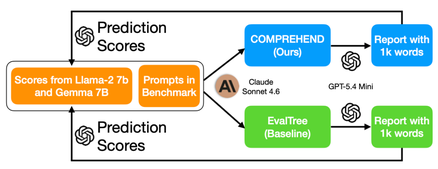


Figure 5: EvalTree for the dissimilar subset.



	RMSE (↓)	Pearson (%)
Llama-2 7B		
EvalTree	$1.83_{\pm 0.17}$	36.0
COMPREHEND	$\mathbf{1.74}_{\pm 0.17}$	**45.0**
Gemma 7B		
EvalTree	$2.26_{\pm 0.16}$	18.9
COMPREHEND	$\mathbf{2.18}_{\pm 0.17}$	**20.2**

Take Home Messages

- **Existing taxonomy structure is either constructed by humans or clustering-based methods, which lead to very a coarse-grained taxonomy.**
- **Don't cluster first and generate description later. Alternate matching and summarizing from the bottom up.**
- **Our method could be adapted to construct an ultra-fine-grained taxonomy / topic model for an arbitrary set of texts (you just need a proper similarity model and a few examples in the common summary generation prompt). Try our methods!**