

Evaluating Language Model Assessment of Students' Writing

Shamya Karumbaiah

Assistant Professor, Learning Sciences

Director, The Responsible AI for Learning (TRAIL) Lab

Department of Educational Psychology

University of Wisconsin-Madison

trail.wcer.wisc.edu

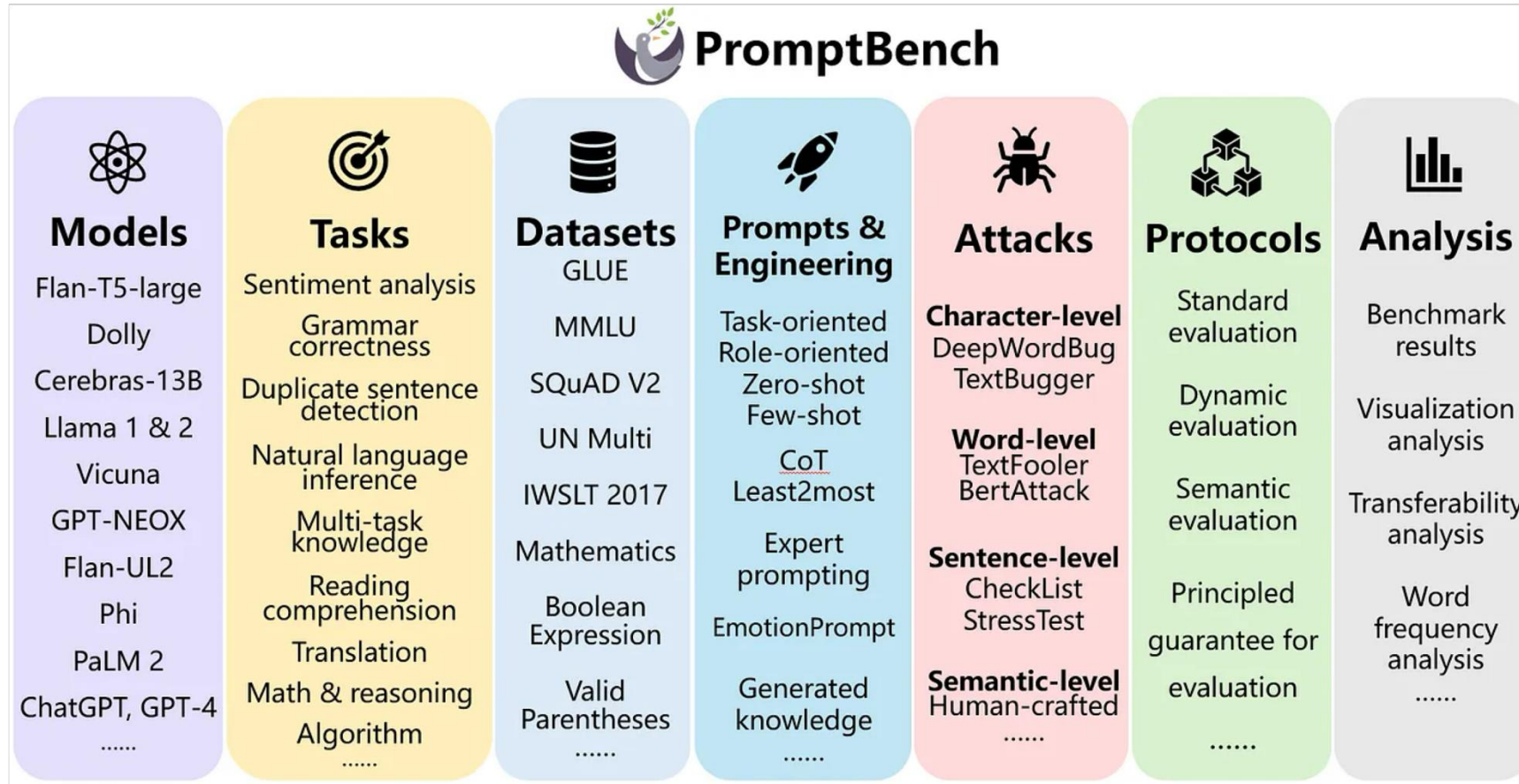
As you settle in, login to AIBAT (AI Behavior Analysis Tool)

trail.wcer.wisc.edu/aibat

[Turn & Talk for 2 minutes]

How do you evaluate AI in your context?

AI Evaluation



AI Evaluation: What We Will Do Today

 Models Flan-T5-large Dolly Cerebras-13B Llama 1 & 2 Vicuna GPT-NEOX Flan-UL2 Phi PaLM 2 ChatGPT, GPT-4	 Tasks Sentiment analysis Grammar correctness Duplicate sentence detection Natural language inference Multi-task knowledge Reading comprehension Translation Math & reasoning Algorithm	 Datasets GLUE MMLU SQuAD V2 UN Multi IWSLT 2017 Mathematics Boolean Expression Valid Parentheses	 Prompts & Engineering Task-oriented Role-oriented Zero-shot Few-shot CoT Least2most Expert prompting EmotionPrompt Generated knowledge	 Attacks Character-level DeepWordBug TextBugger Word-level TextFooler BertAttack Sentence-level CheckList StressTest Semantic-level Human-crafted	 Protocols Standard evaluation Dynamic evaluation Semantic evaluation Principled guarantee for evaluation	 Analysis Benchmark results Visualization analysis Transferability analysis Word frequency analysis
---	--	---	--	--	---	---

The Challenge: Evaluating AI and auditing for biases that reflects stakeholders' **experiences** while accounting for varying pedagogical **contexts** and shifting **populations** over time

Generalization —> Contextual Evaluation

differences in stakeholders' needs, experiences, and priorities define AI biases and failures differently

**Get into pairs or triads. One person in a group,
login to AIBAT (AI Behavior Analysis Tool)**

trail.wcer.wisc.edu/aibat

Context

- Design-based roller coaster unit for motion, forces & energy (8th-grade physics)
- ~12 weeks · 3 teachers · 238 consenting students, 2 Midwest districts
- Students use CoMPASS (a concept-map e-textbook) and an online roller-coaster simulator to gather evidence
- Final task: write an essay justifying their roller coaster design using physics evidence from their trials

The screenshot shows the CoMPASS e-textbook interface. On the left is a navigable concept map for momentum. The central text explains momentum in linear motion, defining it as mass times velocity ($p = m \times v$). It includes a paragraph about the metric unit (kg m/s) and a detailed explanation of Newton's 1st Law, illustrating how momentum is transferred during a collision between an arrow and a target. The interface also includes a search bar, a 'Go to: Linear motion' button, and a 'History' button.

momentum in Linear motion

Momentum is defined as the **mass** of an object times its **velocity** or

$$p = m \times v$$

where p is momentum of a moving object, m is the mass of the object, and v is the velocity of the object. The metric unit for momentum is the kilogram meter/second or kg m/s.

All moving objects have momentum. Momentum, like velocity, is a vector quantity that has both magnitude (amount) and direction. The direction of momentum is the same as the velocity of the object. The more momentum an object has, the harder it is to stop, as explained by **Newton's 1st Law**. For example, we all know it is harder to stop semi-truck hauling logs than it is to stop a small compact car if they are traveling at the same speed. Changes in momentum may occur when there is a change in either the mass of the object, the velocity with which it is traveling, or both.

How can momentum help us understand and predict what happens during collisions? The Law of Conservation of momentum states that the total momentum of a system is the same before and after the collision. Let's use the example of shooting an arrow into a target that is able to move along a metal track with very little friction. Just before the collision, the arrow has momentum based on its mass and velocity and the target has none because it is at rest. After the collision the arrow and the target both travel together as the target moves along the track. The arrow transfers momentum to the target so that both now have momentum. The loss in momentum of the arrow is equal to the gain in momentum by the target. The arrow striking the target provided the **impulse** necessary to change the momentum of both the arrow and the target.

The screenshot shows the Roller Coaster Simulation interface. The top section displays the track design with a car starting at a lift. The track includes a hill, a drop, and another hill. The car's mass is 0.5 kg. The simulation controls include a 'Car Starting Point' dropdown set to 'car lift', a 'Scale' of 1:1, and a 'Max Acceleration' of 40 m/s/s. The bottom section shows the 'Set Up' and 'Measurements' panels. The 'Set Up' panel includes sliders for 'Hill Height' (0.25 m) and 'Hill Friction' (0). The 'Measurements' panel shows 'Applied Force' (3.49 N), 'Work' (0 J), and 'Potential Energy' (0 J).

Roller Coaster Simulation

Car Starting Point: car lift

Scale = 1:1

Max Acceleration: 40 m/s/s

Largest acceleration reached: 0.04 m/s/s

Largest net force reached: 0.02 N

Car mass: 0.5 kg

Set Up: hill Track Type: hill Hill Type: gradual

Hill Height: 0.25 m Hill Friction: 0

Measurements: Applied Force: 3.49 N Work: 0 J Potential Energy: 0 J

CoMPASS e-textbook with navigable concept maps (left) and the roller coaster simulation students used to run experiments (right)

Human Annotations

CONTENT UNIT KEY



CU0

Greater height → greater Potential Energy



CU1

Down the track: PE decreases as KE increases



CU2

CU2 Total energy (PE + KE) stays constant



CU3

Law of Conservation of Energy



CU4

Hill height should be lower than the initial drop



CU5

Mass → energy

E1: The drop height of my roller coaster is going to be 5 meters. This is because the more height the car has the more potential energy. For example, when my drop height was 2 meters the potential energy was 977 joules. When I increased the drop height to 3 meters the potential energy increased 489 joules. When the car got to the bottom of the drop, the kinetic energy was greater the more height it had at the beginning.

Kinetic energy and Potential energy add up to the total amount of energy. The conservation of energy law states that energy cannot be created or destroyed. At the top of drop the potential energy has the total amount of energy, but as the car goes down the drop it transfers into kinetic energy. Once it is at the very bottom, potential energy has 0 joules and kinetic energy has the total amount because kinetic energy is involved with velocity. Potential energy is based on three things, one of them is height. The greater the height the more potential energy. When the car starts 3 meters tall the potential energy is 1466 joules. If you put a hill that is 4 meters tall then there will simply not be enough to get over the hill. The energy can't be created or destroyed, the car only has enough to get up 3 meters. Also the more mass the car has the more energy because potential energy is height x mass x gravity. The greater the mass the greater the energy.

My roller coaster will have an initial drop of 5 meters and will have a hill right after that is only 2 meters high. The car will make it over easily especially because the mass is 50 kg. After the hill the car will be able to go farther with hills, loops, etc. So it is really fun and fast but it is still safe.

STUDY 1

Prompt vs. Supervise

RESEARCH QUESTION

How effective are pre-trained and finetuned general-purpose language models, compared to feature-engineered models, at assessing inquiry-based science explanations?

How the essays were annotated

1

Coder 1 — Domain Expert

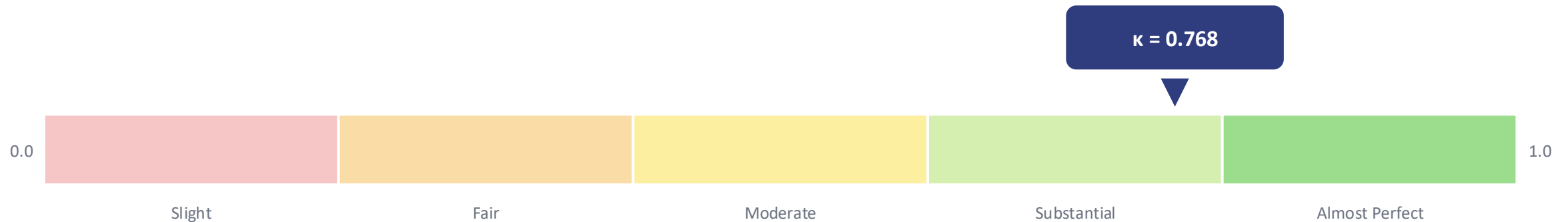
- 8th-grade science teacher for a decade
- Now a science education researcher
- Authored the CU rubric

2

Coder 2 — Research Assistant

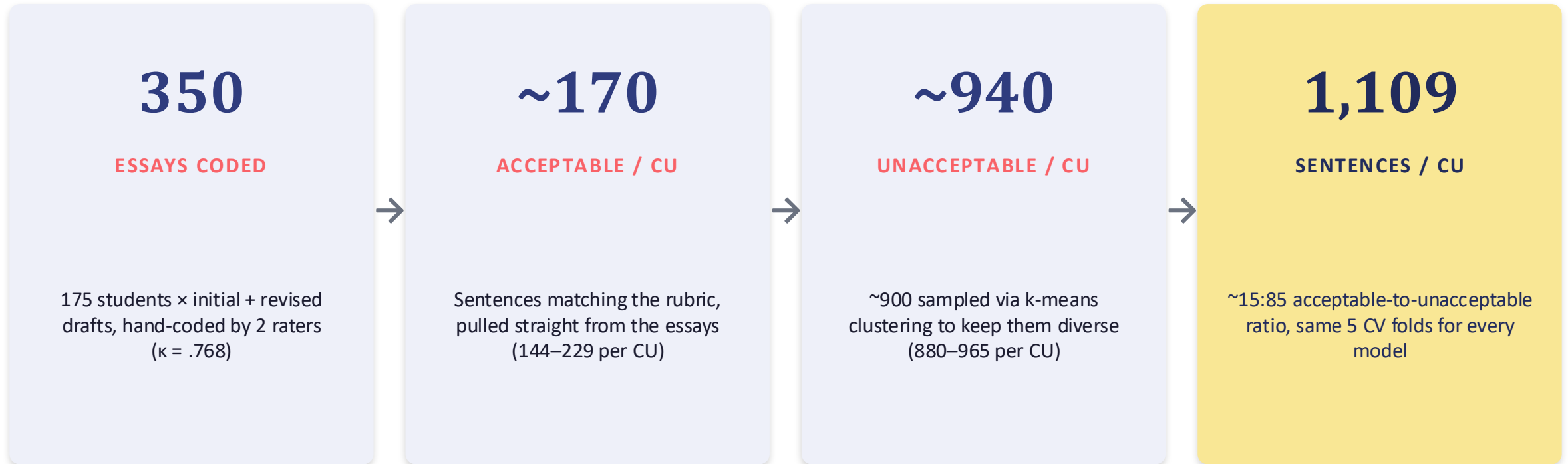
- Undergraduate research assistant
- Trained directly by Coder 1 on the rubric
- Coded independently to check consistency

BOTH CODERS INDEPENDENTLY ANNOTATED THE SAME 10% OF ESSAYS, THEN COHEN'S KAPPA MEASURED THEIR AGREEMENT



$\kappa = 0.768 \rightarrow$ **substantial agreement (Landis & Koch, 1977)**. Disagreements were resolved through discussion; one coder then annotated the remaining ~90% of essays alone, highlighting the exact sentence matching each concept unit.

From 350 essays to a labeled dataset



Two open-source LMs

Llama 3 (8B-Instruct)

- Decoder-only transformer from Meta AI
- Pretrained on 15 trillion tokens + RLHF instruction-tuning
- Used for BOTH prompting and finetuning
- Runs locally on a single Nvidia A40 GPU

RoBERTa-large

- Encoder-only transformer, BERT family
- Masked-language-model pretraining, dynamic masking, no next-sentence prediction
- Used for supervised finetuning AND entailment-based zero-shot
- Widely used already for classification in education research

Why local, open models instead of GPT-4? Keeps sensitive student writing off third-party servers, and avoids per-token API costs.

Two ways to get a model to grade a sentence



Prompting

- Write instructions in plain text; the pretrained model predicts directly
- No labeled data or training required
- Highly flexible - redefine the task by editing the prompt
- Output can be inconsistent (“incorrect” vs. “not correct”)

Zero-shot → few-shot → chain-of-thought

Zero-Shot

just the rubric

“Does the sentence contain the physics concept: ‘Greater height means greater Potential Energy.’ If yes, return acceptable. Otherwise, return unacceptable:”

[+ target sentence appended here]

Few-Shot

+ 2 worked examples

Task: Does the Answer describe the given Concept?

Ex. 1 — Concept: height → PE. Answer: “The more mass an object has, the more PE it has.” Decision: Unacceptable

Ex. 2 — Concept: height → PE. Answer: “The total energy will never change.” Decision: Unacceptable

[+ target sentence to classify]

Chain-of-Thought

+ reasoning first

Same setup as few-shot, plus:

“Explain your reasoning first, then make your decision.”

Ex. 1 — Answer: “...the more height, the more PE it has.” Explanation: correctly links height with PE. Decision: Acceptable

Ex. 2 — Answer: “The total energy will never change.” Explanation: mentions energy, not PE or height. Decision: Unacceptable

Two ways to get a model to grade a sentence



Prompting

- Write instructions in plain text; the pretrained model predicts directly
- No labeled data or training required
- Highly flexible - redefine the task by editing the prompt
- Output can be inconsistent (“incorrect” vs. “not correct”)



Supervised Learning

- Show the model many labeled examples; it learns task-specific weights
- Needs annotated data + training compute
- Rigid - new task means new labels and retraining
- Consistent, evaluable output (fixed label set)

Teaching the models directly on human annotated data

RoBERTa - full finetuning

- Classification head + softmax layer on top of roberta-large
- Trained with cross-entropy loss, up to 20 epochs
- Early stopping (patience = 8) to avoid overfitting
- One model per concept unit per fold ($6 \times 5 = 30$ models)

Llama 3 - LoRA finetuning

- Parameter-efficient finetuning via Low-Rank Adaptation (LoRA)
- Implemented with Unsloth to cut training memory
- 3 epochs, learning rate = $2e-4$
- Updates small adapter weights, not all 8B parameters

Teaching the models directly on human annotated data

RoBERTa - full finetuning

- Classification head + softmax layer on top of roberta-large
- Trained with cross-entropy loss, up to 20 epochs
- Early stopping (patience = 8) to avoid overfitting
- One model per concept unit per fold ($6 \times 5 = 30$ models)

Llama 3 - LoRA finetuning

- Parameter-efficient finetuning via Low-Rank Adaptation (LoRA)
- Implemented with Unsloth to cut training memory
- 3 epochs, learning rate = $2e-4$
- Updates small adapter weights, not all 8B parameters

5-FOLD CROSS-VALIDATION

1

The 1,109-sentence dataset for each concept is split into 5 folds

2

Each model trains on 4 folds and is tested on the 5th, rotating until every fold is held out once

3

The reported score is the average across all 5 folds

A non-LM baseline: hand-crafted features + classic classifiers

Hand-engineered features

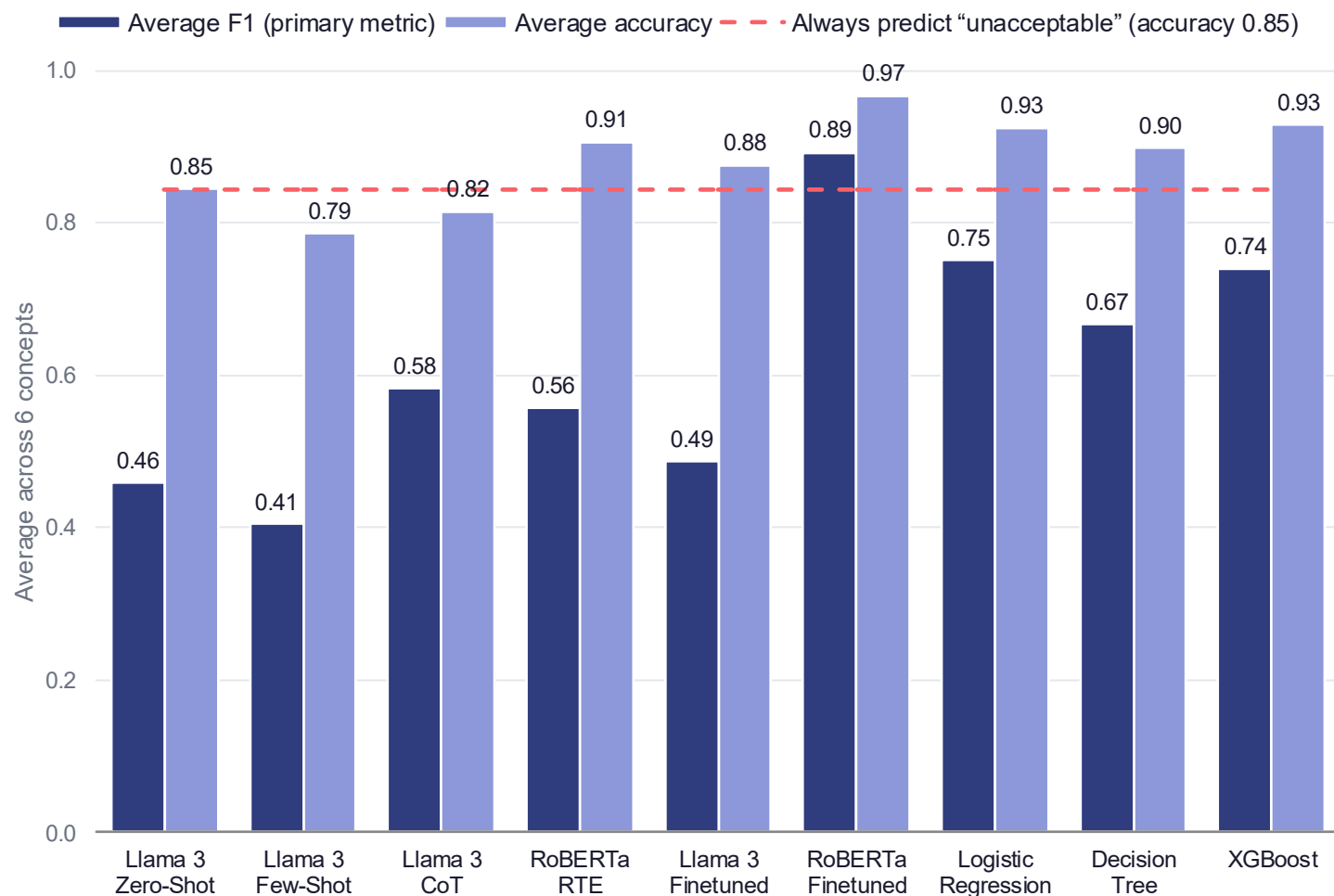
- Lexico-grammatical: sentence & token length, punctuation, parse-tree depth, type-token ratio, connectives
- Semantic: word and part-of-speech bi-/tri-gram counts
- Task-specific: TF-IDF similarity & KL divergence to each concept's core examples
- Fit on the training folds only - same 5 folds as the LMs

Three scikit-learn classifiers

- Logistic Regression (L2, C = 1.0)
- Decision Tree (Gini, library defaults)
- XGBoost (100 trees, max depth 6)
- Trains on a standard laptop - no GPU needed; coefficients & rules are inspectable

RESULTS

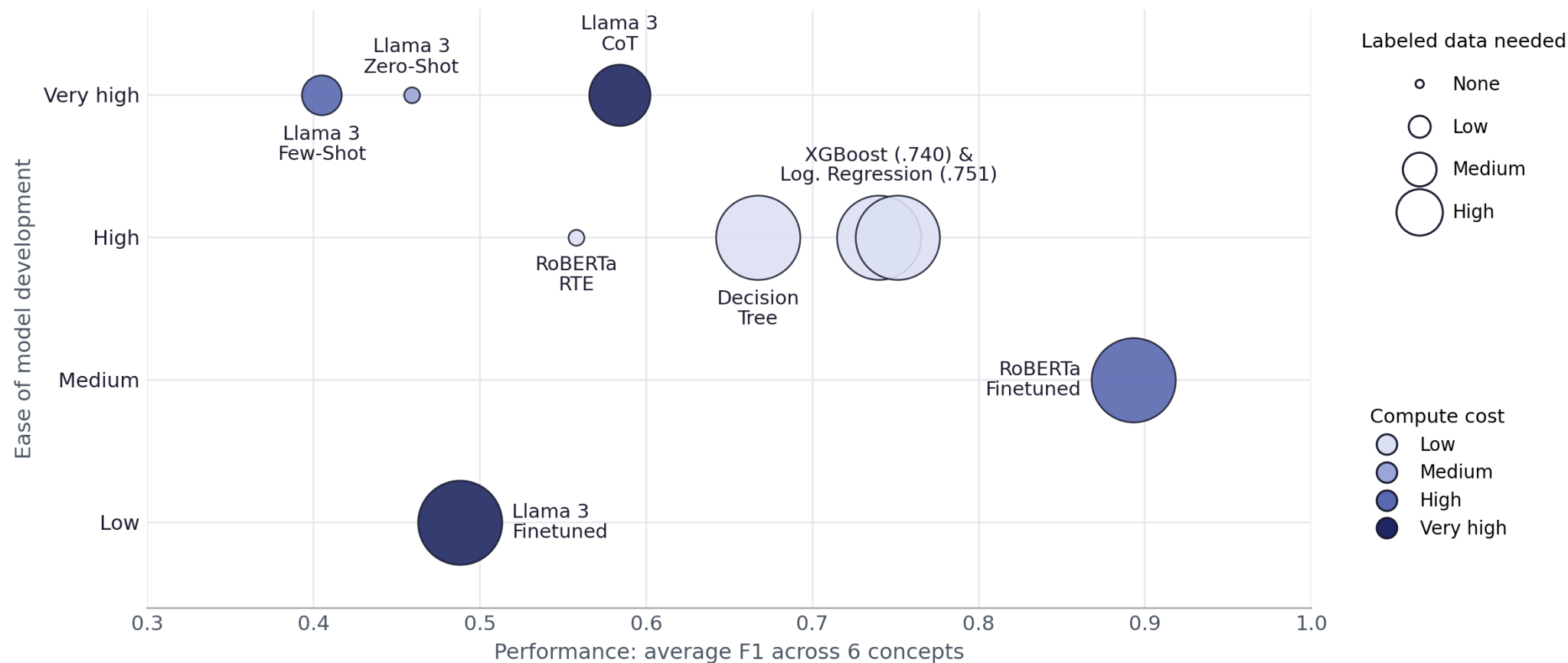
Full supervision wins in performance



KEY TAKEAWAYS

- RoBERTa finetuned is best on every concept - avg. F1 0.89 (0.86–0.94), accuracy ≥ 0.94
- Feature-engineered models are the runner-up - Logistic Regression F1 0.75, XGBoost 0.74
- Prompted Llama 3 reaches 0.79–0.85 accuracy but only 0.41–0.58 F1 - no better than always guessing "unacceptable"
- Finetuning did not help Llama 3 (F1 0.49), with large swings across folds

Performance vs. the resources each model demands



Qualitative schematic after the paper's Fig. 3. Ease, data, and compute levels are the authors' judgments; F1 is the 6-concept average.

STUDY 2

Contextual Evaluation of Language Models with Behavior Analysis

Generalization —> Contextual Evaluation

differences in stakeholders' needs, experiences, and priorities define AI biases and failures differently

System behaviors defined through testing can act as components of such **contracts** — particularly when they incorporate stakeholders' contextual expertise to translate **implicit expectations** (e.g., fairness) into **explicit, testable criteria**

[Turn & Talk for 2 minutes]

What linguistic variations matter in your context?

Auditing LLM Biases Using Linguistic Variations

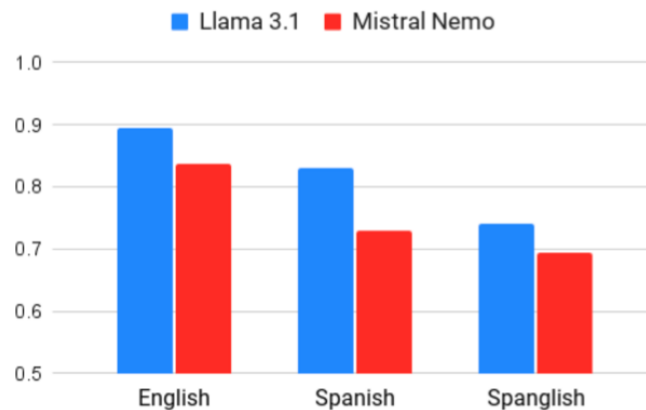


Fig. 2. Comparing pre-trained performance (AUC) of Llama 3.1 and Mistral NeMo for English, Spanish, and Spanglish grading. Both models perform best for English and worst for Spanglish

Disaggregating LLM Biases By Categorizing Variations

Table 1
Linguistic Variations and Examples

Linguistic Variation	Definition	Examples
English	A global lingua franca often used as the baseline in bi-/multilingual education.	Hola, me llamo _____. → Hi, my name is _____.
Spanish	A Romance language that is the native language for most EBLs in the U.S.	Hi, my name is _____. → Hola, me llamo _____.
Spanglish	A hybrid of English and Spanish resulting from cross-cultural interactions.	Estoy ready para ir al mall.
Nouns	The fluid use of nouns across languages, often due to lexical transfer.	The students need their libro .
Spelling	Variations in the spelling of words influenced by phonology and language learning processes.	She is frendly (She is friendly) La veo una ves a la semana (La veo una vez a la semana)
Cognates	Words in two languages with similar meaning, spelling, and pronunciation due to shared etymology.	Information → Información Potential energy → Energía potencial
Dialect	Variations of a language based on regional, social, or cultural differences.	I'm going to the store → I'm gonna go to the store Me voy a ir al supermercado → Me voy al súper
Loan words	The use of words borrowed from one language into another due to language contact.	He sent me an email → Me mandó un email (en vez de correo electrónico)

Disaggregating LLM Biases By Linguistic Variations

Table 2
Current Linguistic Variations

Linguistic Variation	Perceived Prevalence in Student Writing
Cognates	✓✓✓✓✓
Spelling (language-specific, punctuation, grammar, graphemic errors)	✓✓✓✓✓✓✓
Non-standard varieties (regionalisms, colloquialisms, tense variation)	✓✓✓✓✓✓
Translations (English, Spanish, Spanglish)	✓✓✓✓✓
Noun transfer (noun substitution)	✓✓
Loan words (lexical borrowing, <u>a</u> mericanisms)	✓✓

Table 3
Teacher Suggested Linguistic Variations

Suggested Linguistic Variation	Perceived Prevalence in Student Writing
Accentuation	✓✓
Word order (syntax/morphosyntax)	✓✓✓
Nicaraguan Mexican Puertorican Dominican Castellano	✓ ✓ ✓ ✓ ✓
Synonyms	✓
Phonetically-influenced orthography (sound-influenced spellings)	✓✓
Non-binary words	✓
Idioms	✓
Novel linguistic forms	✓✓✓✓

Disaggregating LLM Biases By Categorizing Variations

Table 3

AUC Reports of Fine-tuned Performance on Linguistic Variations.

Variation	Baseline	ENG-FT	ESP-FT	SPANG-FT
Original-English	0.884	0.943	0.940	0.955
Default-Spanglish	0.735	0.771	0.819	0.924
Dialect-Mexican	0.771	0.855	0.871	0.945
(0.036)*				
<i>Cross-Lingually (Dis)Similar Linguistic Variations</i>				
Spelling Errors	0.874	0.917	0.908	0.931
PIO (-0.085)*	0.789	0.806	0.804	0.923
Cognates	0.825	0.865	0.869	0.937
Noun Transfer (-0.092)*	0.733	0.832	0.843	0.933

*Denotes the difference in baseline model AUC scores between nested pairs (ie. spelling error vs. PIO). Bold reflects the base variation in a nested pair.

**Get into pairs or triads. One person in a group,
login to AIBAT (AI Behavior Analysis Tool)**

trail.wcer.wisc.edu/aibat

Contextual Evaluation

- Performance metrics on data drawn from the same distribution as training
- *What happens when you use the model in a new context?*
- Multilingual learners, neurodivergent students, biology or history lesson
- Can we test the models for variations we expect in a new context?

Behavior Analysis

Base sentence (Acceptable): *"Kinetic energy is the energy of motion."*



Robustness (acronyms, spelling)

"KE is enrgy of motion."

→ Label should stay the same: Acceptable



Fairness (multilingualism)

"KE is energía of motion."

→ Label should stay the same: Acceptable



Taxonomy (antonym)

"KE is energy of rest."

→ Label should flip: Unacceptable

Two building blocks: what to test, how it should respond

BEHAVIOR TYPES - how output should respond

INV

Invariance

Output should NOT change - e.g. "PE" vs. "potential energy" shouldn't flip the label

DIR

Directional Change

Output SHOULD change predictably - e.g. negating a correct answer should flip it

LINGUISTIC CAPABILITIES - what to test

K

Knowledge - Basic concept correctness

R

Robustness - Spelling errors, acronyms, paraphrasing

T

Taxonomy - Synonyms & antonyms of key words

N

Negation - Adding "not" shouldn't confuse the model

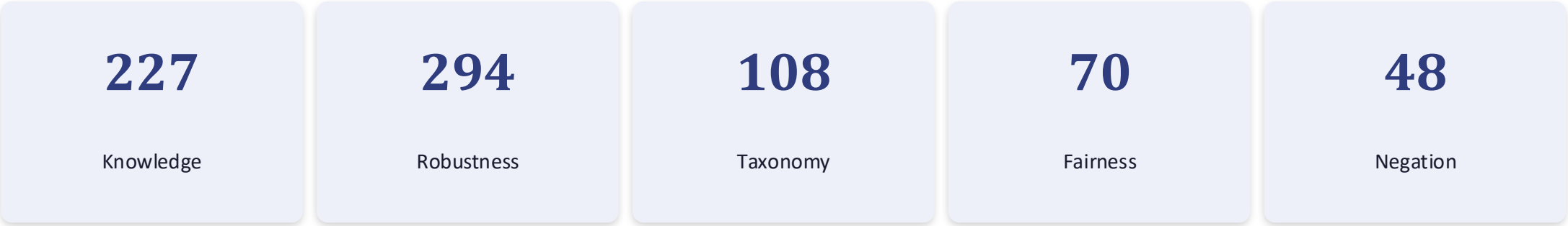
F

Fairness - Spanish-translated words, multilingual writing

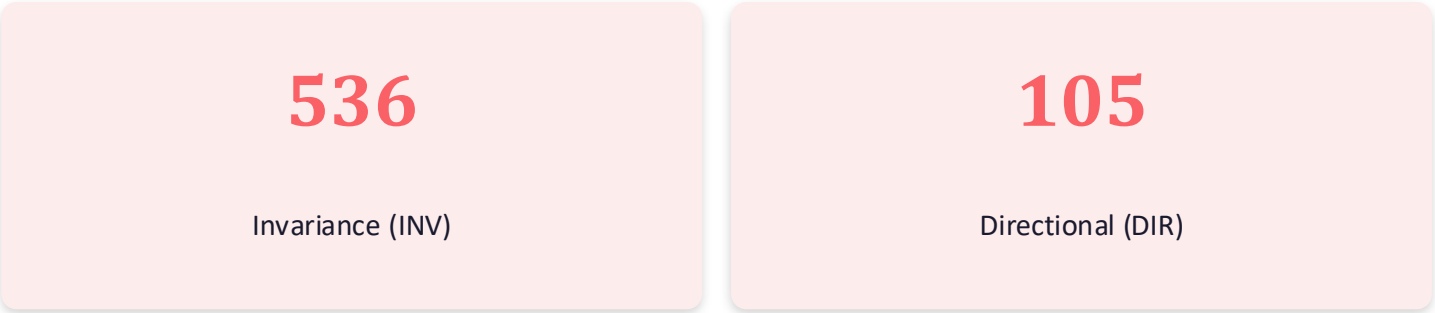
Same models, same 6 concepts - a new 747-test suite

Reuses the six language-model conditions from Part 1: Llama 3 (Zero/Few-Shot, CoT, Finetuned) and RoBERTa (RTE, Finetuned) - no new models trained.

BY CAPABILITY

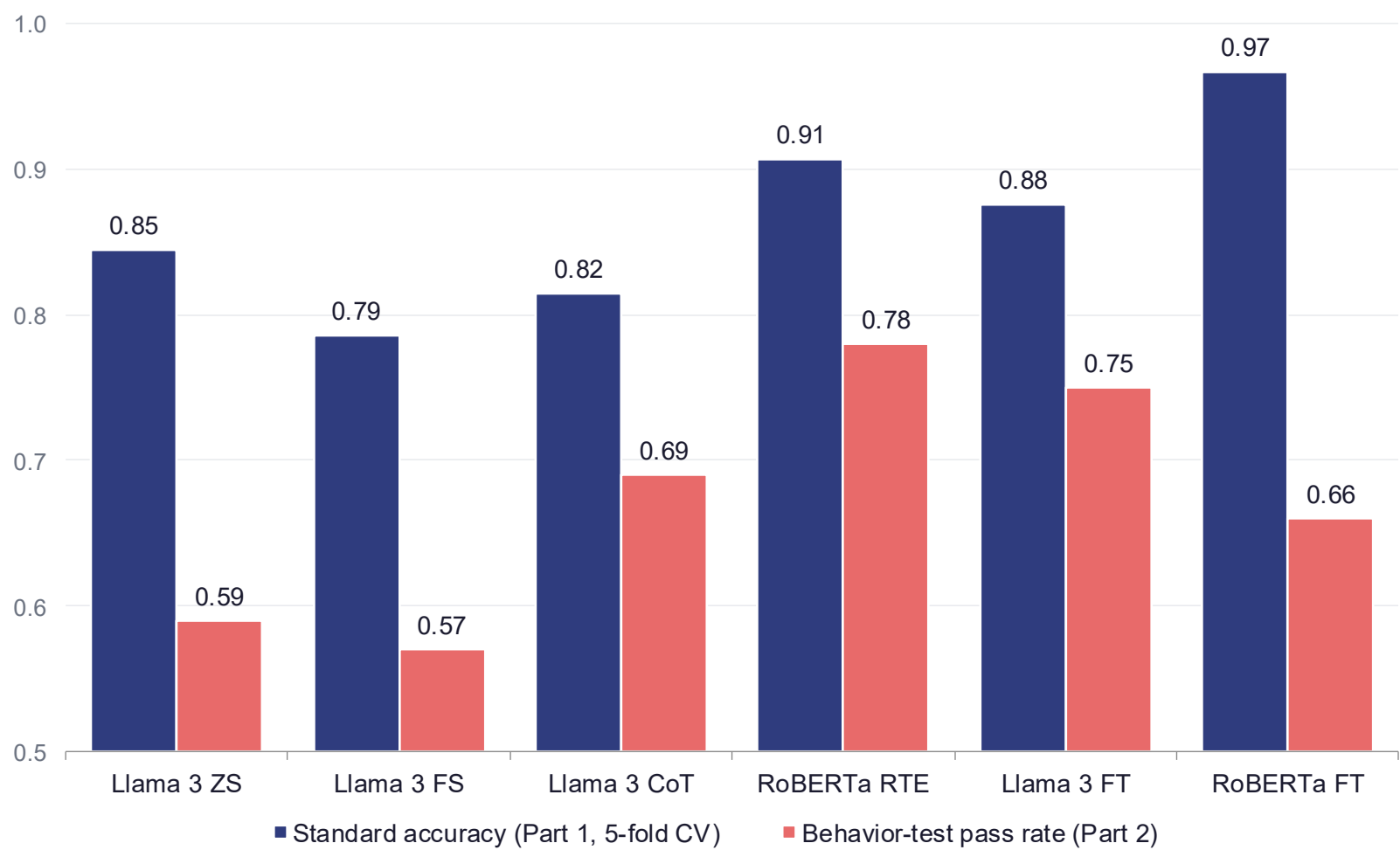


BY BEHAVIOR TYPE



RESULTS

The 0.97-accuracy model fails about a third of behavior tests



HEADLINE

RoBERTa Finetuned

0.97 → 0.66

The best-performing model in Part 1 sees the steepest drop of any model once tested for expected variations. Every model drops - about 22% on average.

TAKEAWAY

Why behavior analysis matters

1

Works without big data

2

Surfaces hidden failures

3

Localizes root causes

Together: don't just ask "how accurate is it?" - ask "accurate on what, and robust to what?"



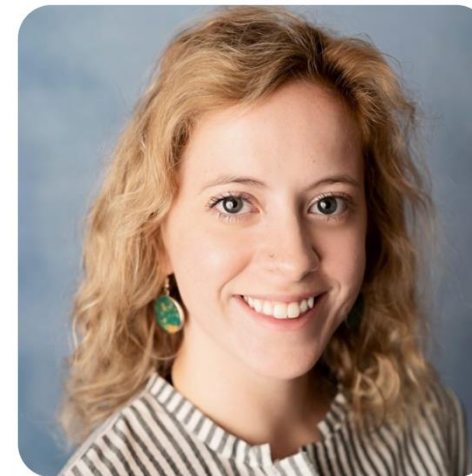
Ananya Ganesh
Postdoctoral Fellow



Anurag Maravi
Software Engineer



Yaxuan Yin
PhD Student
Advisor: Jacob Thebault-Spieker



Kaycie Barron
PhD Student



Aayush Bharadwaj
Software Engineer, Capital One



Lucas Anderson
Software Engineer, Cashmere



Nora Tseng
MS Computer Science, University of
California, San Diego (UCSD)



Vivek Saravanan
Software Engineer, Databricks



Shanglin Yang
MS, Carnegie Mellon University



Thank you

Behavior Analysis Tool: trail.wcer.wisc.edu/aibat

Questions?