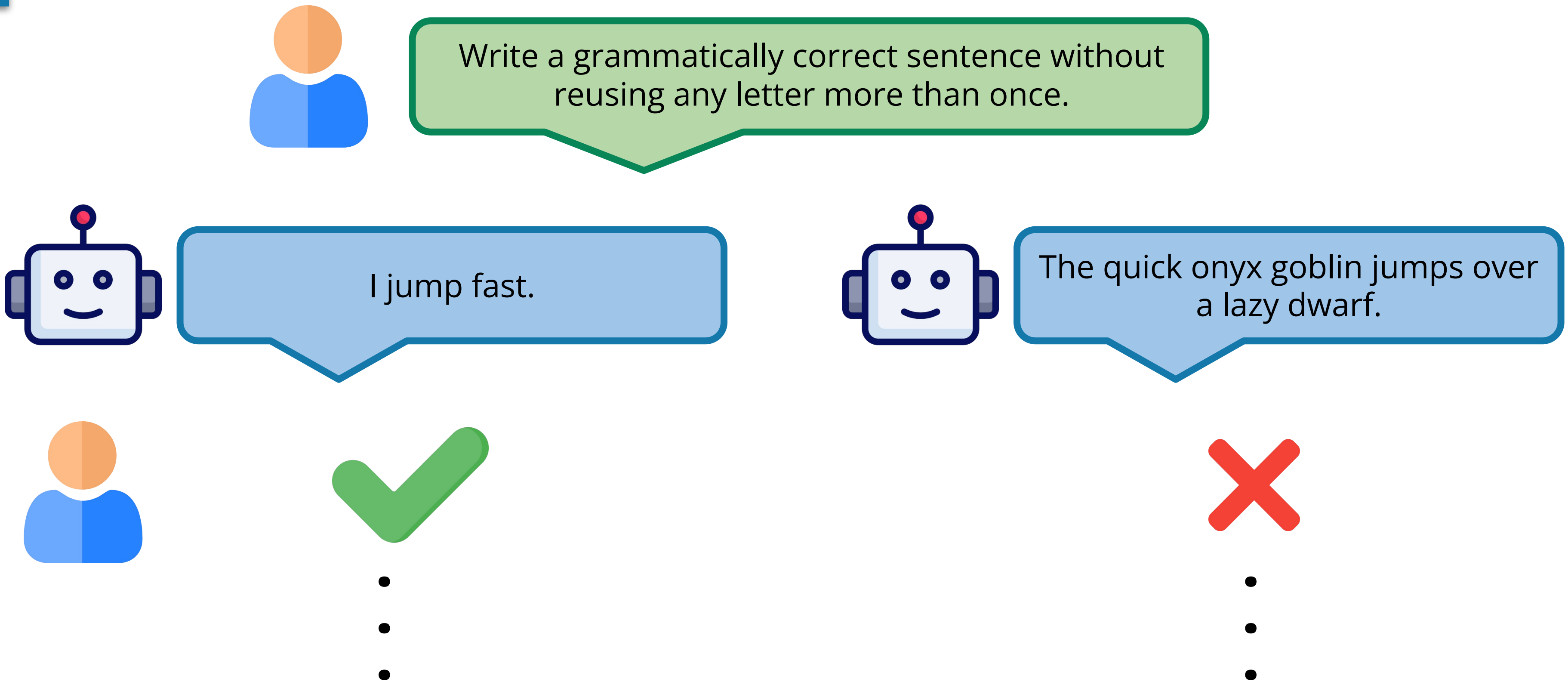
Polyrating: A Cost-Effective and Bias-Aware Rating System for LLM Evaluation

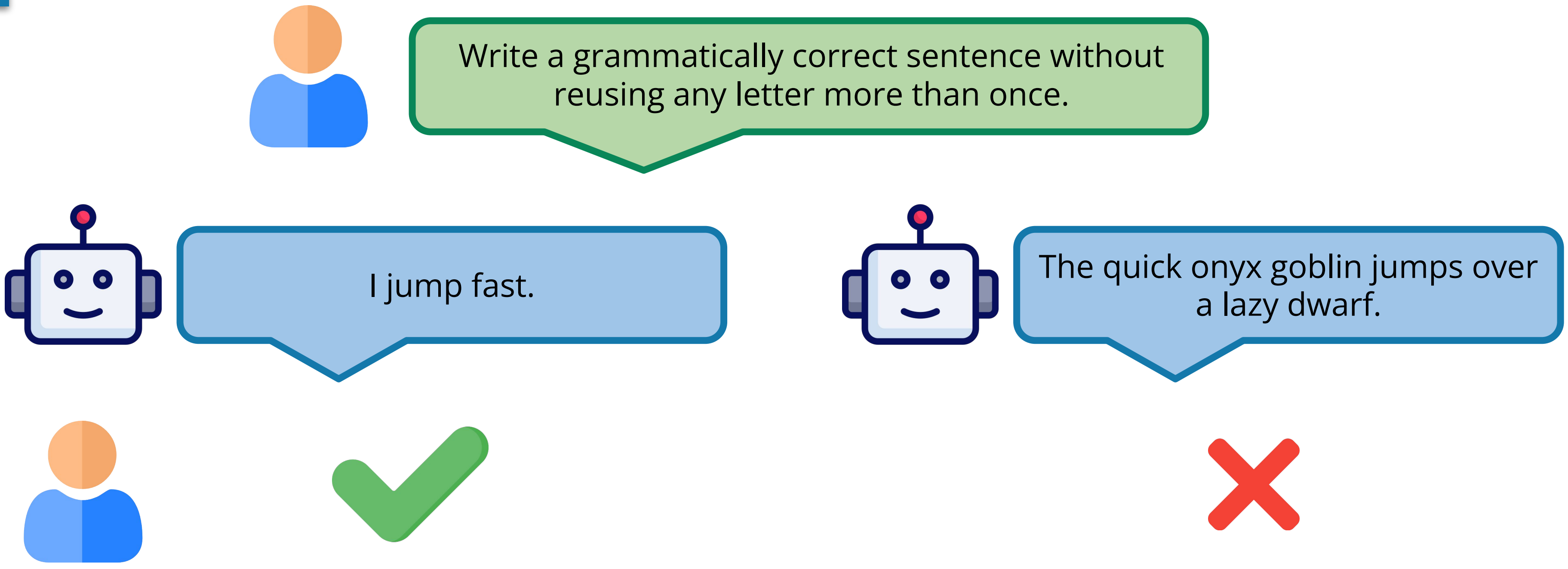
Jasper Dekoninck Maximilian Baader Martin Vechev

ETH Zurich, Switzerland

Judge-Based Evaluation



Judge-Based Evaluation



$$P(M_1 > M_2) = \frac{1}{1+\exp\left(\frac{R_2-R_1}{400}\right)}$$

$R_1 = 1400$

$R_2 = 1000$

Problem #1: Expensive Evaluation

Human evaluation is expensive.

Current rating systems are inefficient.

Problem #1: Expensive Evaluation

Human evaluation is expensive.

Current rating systems are inefficient.

$5,000 - $50,000

$0.10 - $1.00

50,000 samples

Problem #2: Judge Biases Affect Rating

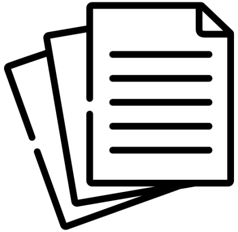
Judges may base their decisions on biases or prejudices.

Current rating systems cannot detect biases effectively.

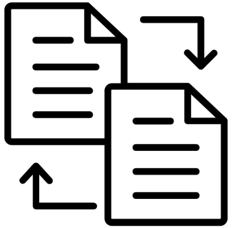
Problem #2: Judge Biases Affect Rating

Judges may base their decisions on biases or prejudices.

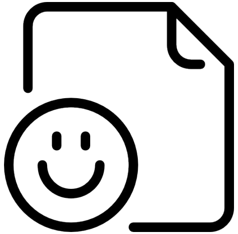
Current rating systems cannot detect biases effectively.



Length



Position



Sentiment

Formality

Icons from https://www.flaticon.com/ and created by (in order of appearance): Freepik, vectorsmarket15, justicon, Freepik

Problem #3: Unaligned Ratings

Rating comparisons between leaderboards are meaningless.

Current rating systems compute ratings against an arbitrary reference.

Problem #3: Unaligned Ratings

Rating comparisons between leaderboards are meaningless.

Current rating systems fix ratings using an arbitrary reference.

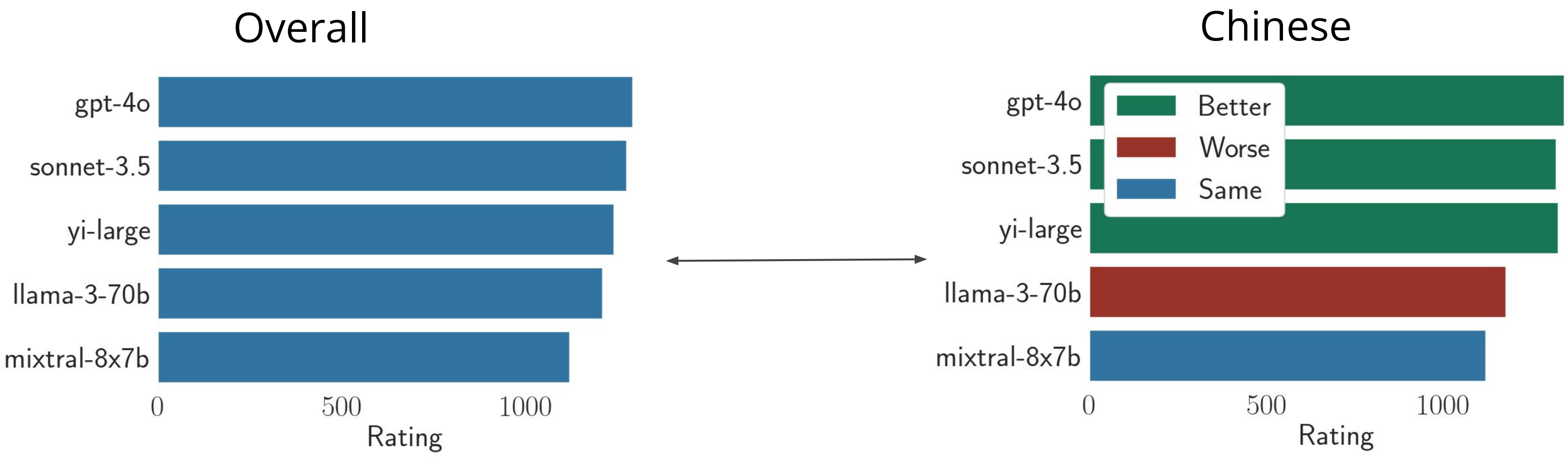
$$\exp\left(\frac{R_2 - R_1}{400}\right) = \exp\left(\frac{(R_2 + C) - (R_1 + C)}{400}\right)$$

$$\begin{aligned} R_1 &= 1400 \\ R_2 &= 1000 \end{aligned} \quad \Leftrightarrow \quad \begin{aligned} R_1 &= 1200 \\ R_2 &= 800 \end{aligned}$$

Problem #3: Unaligned Ratings

Rating comparisons between leaderboards are meaningless.

Current rating systems compute ratings against an arbitrary reference.



97/114 models “improve”

Insight #1: Multivariate Modelling to Mitigate Unaligned Ratings

Ratings can be modeled as multivariate objects.

Enables accurate rating comparisons.

Insight #1: Multivariate Modelling to Mitigate Unaligned Ratings

Ratings can be modeled as multivariate objects.

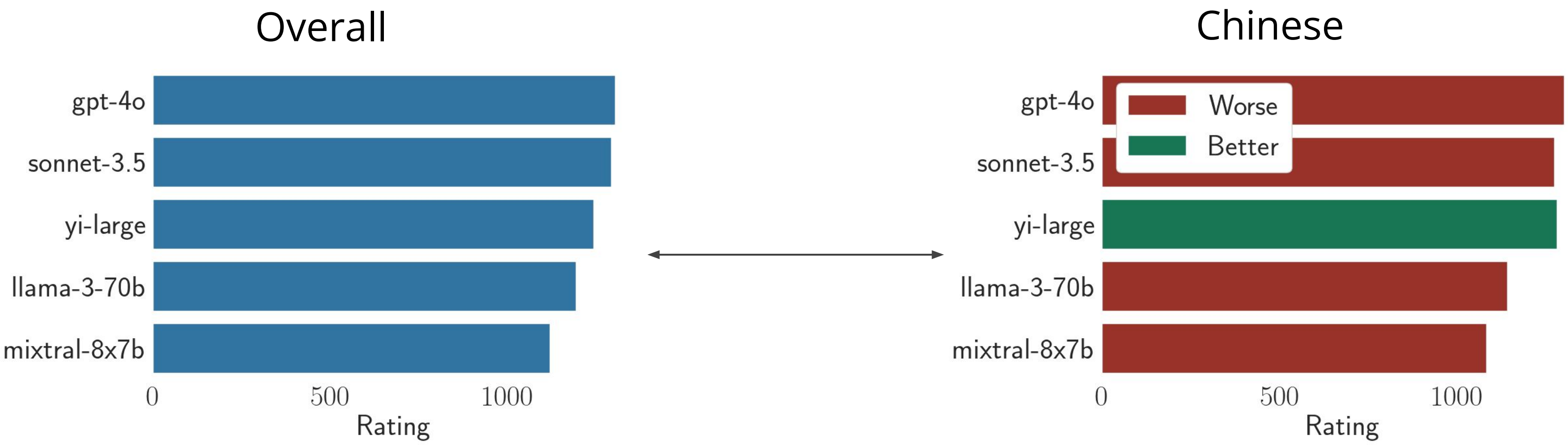
Enables accurate rating comparisons and continuous dependencies.

$$R_m(q) = R_{m,\text{base}} + R_{m,\text{chinese}}[\![q \in Q_{\text{chinese}}]\!] + \dots$$

Insight #1: Multivariate Modelling to Mitigate Unaligned Ratings

Ratings can be modeled as multivariate objects.

Enables accurate rating comparisons.



Half of the models improve
Only bilingual models improve significantly

Insight #2: Shared Parameters to Measure Judge Biases

Sharing parameters between models measures overall influences.

Enables measuring the influence of judge biases.

Insight #2: Shared Parameters to Measure Judge Biases

Sharing parameters between models measures overall influences.

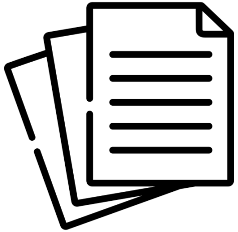
Enables measuring the influence of judge biases.

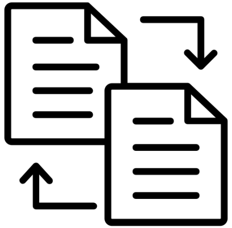
$$R_m(q, \text{response}) = R_{m,\text{base}} + \alpha_{\text{length}} \text{Length}(\text{response}) + \ldots$$

Insight #2: Shared Parameters to Measure Judge Biases

Sharing parameters between models measures overall influences.

Enables measuring the influence of judge biases.





	Length	Position
LLM	+48.5	+37.5
Human	+40.8	+2.7

Icons from https://www.flaticon.com/ and created by (in order of appearance): Freepik, vectorsmarket15

Insight #3: Regularization for Improved Convergence

Using priors on the parameters acts as regularization.

Enables faster convergence by using other (cheap) data.

Insight #3: Regularization for Improved Convergence

Using priors on the parameters acts as regularization.

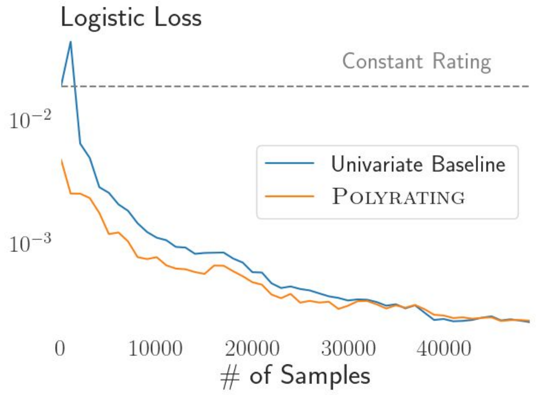
Enables faster convergence by using other (cheap) data.

$$R_m(q) = R_{m,\mathrm{LLM}} + R_{m,\mathrm{human}}[\![q \in Q_{\mathrm{human}}]\!] + \dots$$

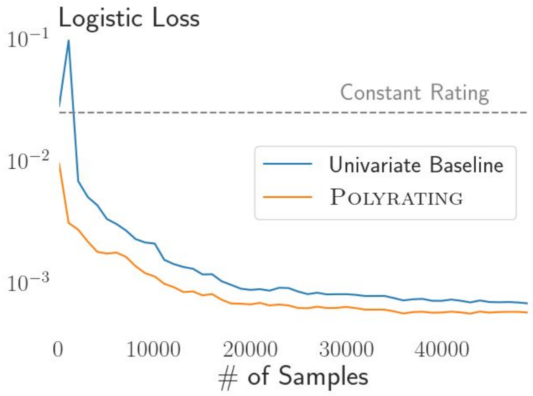
Insight #3: Regularization for Improved Convergence

Using priors on the parameters acts as regularization.

Enables faster convergence by using other (cheap) data.



Prior on cheap benchmark



Prior on model versions