Supplementary Material

Evaluating the Effectiveness of Error Bars and Multilevel Confidence Bands in Conformalized Regression

M. Dossi[1], J. T. Lin[1], A. M. Smits[2], A. Chatzimparmpas[1]

[1] Department of Information and Computing Sciences, Utrecht University, The Netherlands
[2] Department of Human Centered Technologies, Hanze University of Applied Sciences, The Netherlands

1. House Visualization

Survey Interface and Results from Participant Estimates

House visualizations are presented by complexity level.

For each house, the first slide shows the uncertainty visualizations used in the survey interface, and the second slide presents the corresponding analysis results.

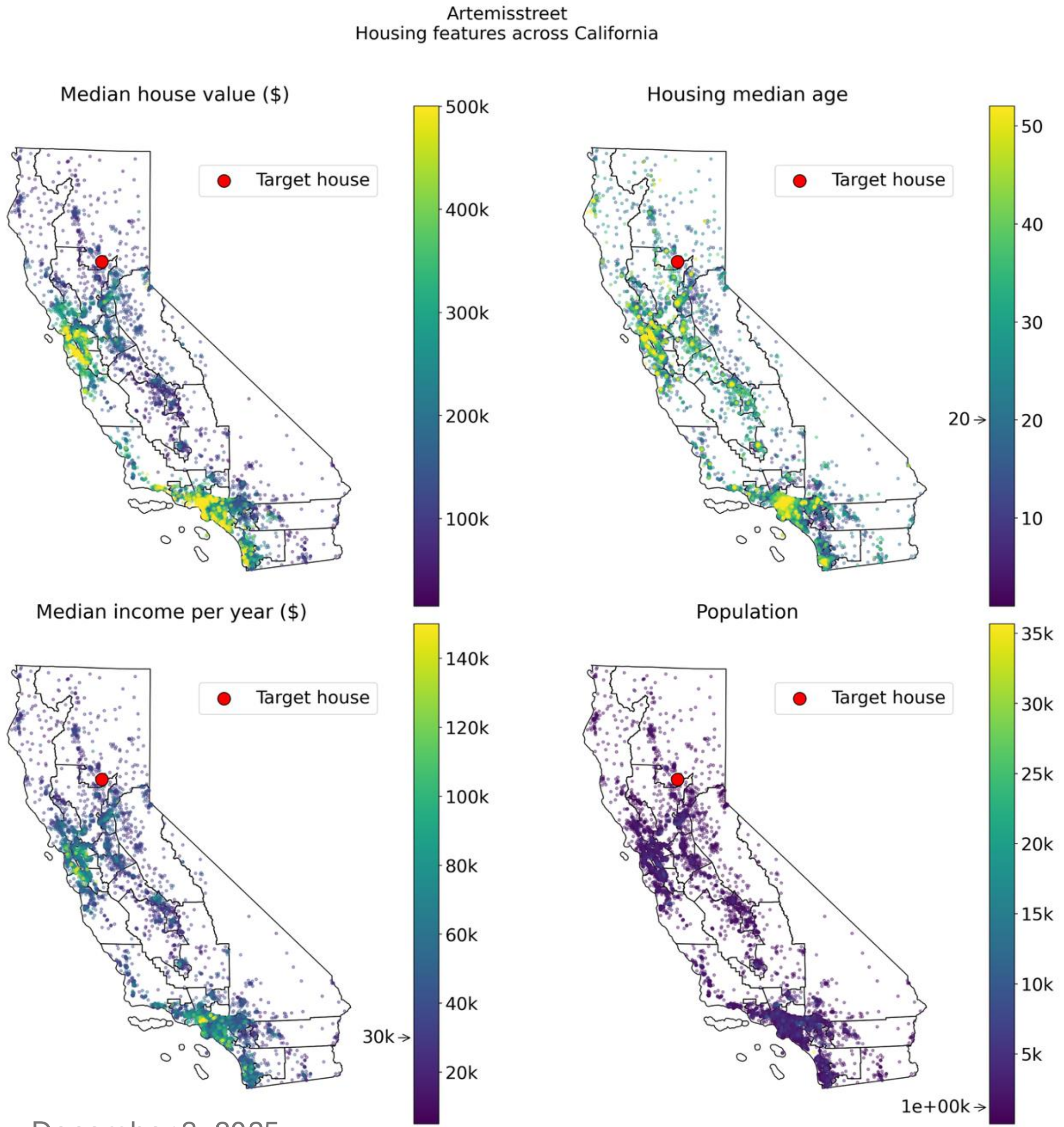
Note that, unlike in the survey, these slides include a red dot (for graded error bars) or a red dashed line (for multilevel confidence bands) to indicate the model's prediction for the target house. This cue was intentionally omitted in the survey.

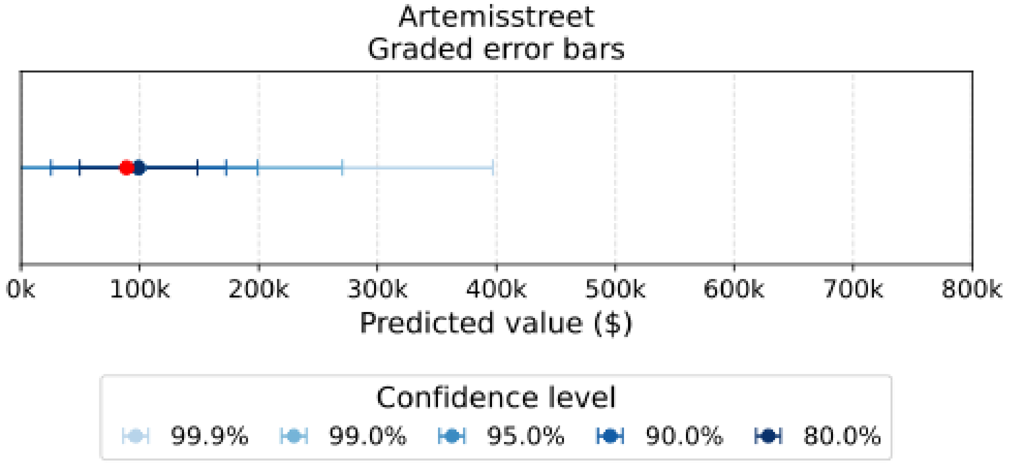
Task complexity: easy

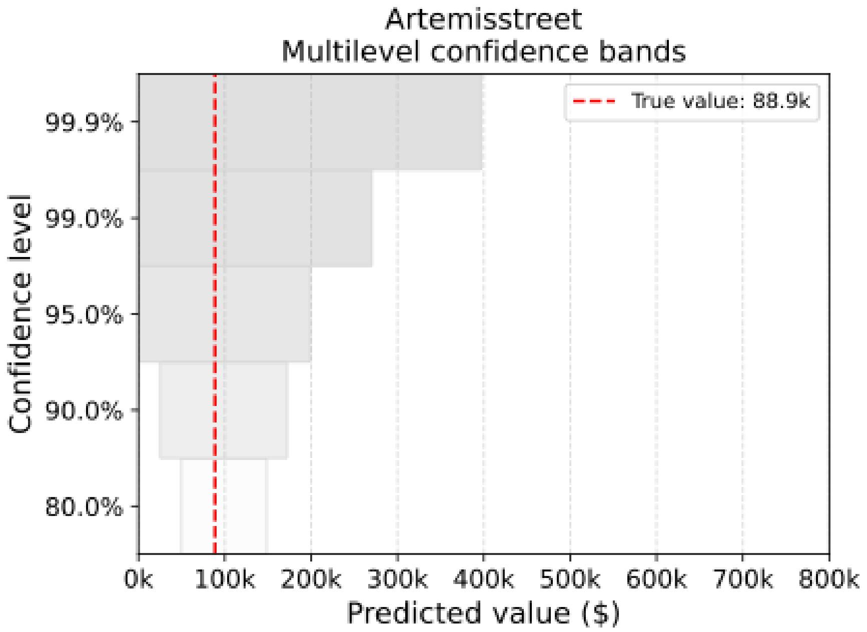
- 1) Artemisstreet
- 2) Aphroditestreet
- 3) Apollostreet

1) Low error – Artemisstreet

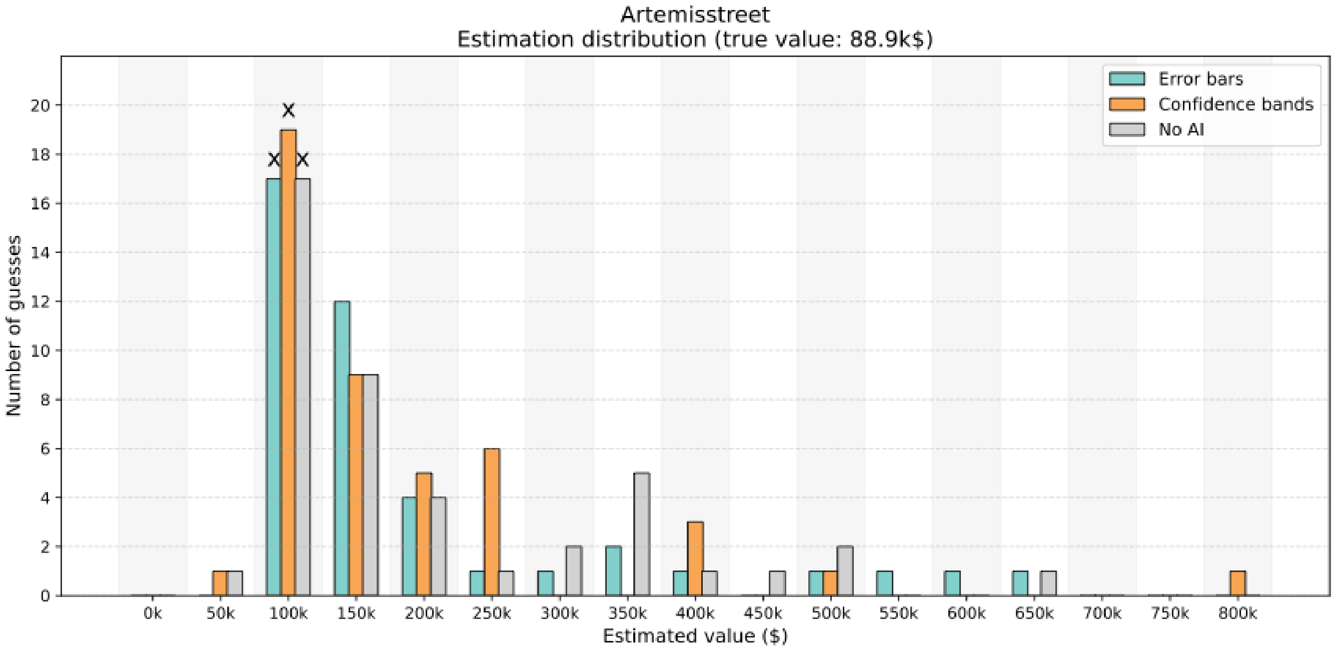
Survey interface

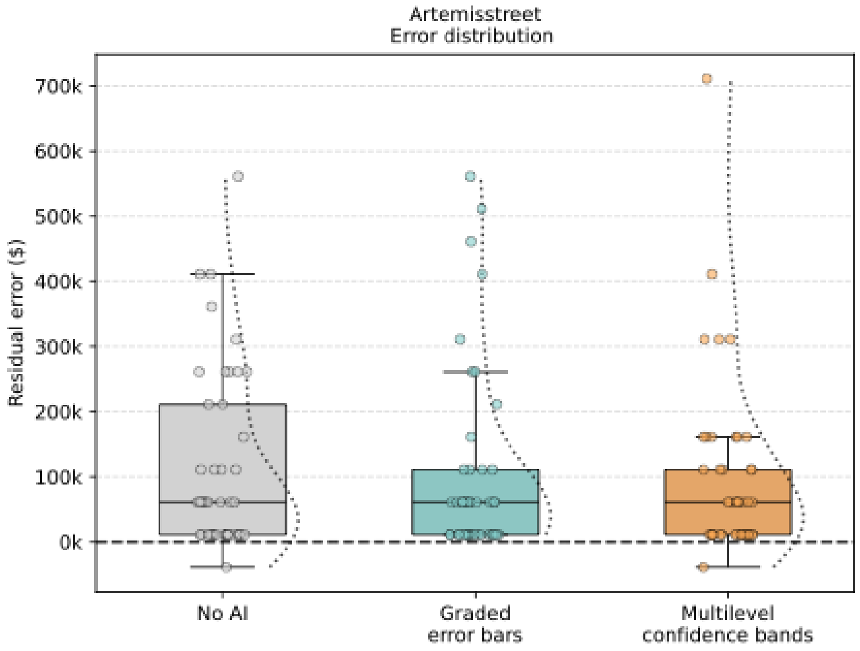




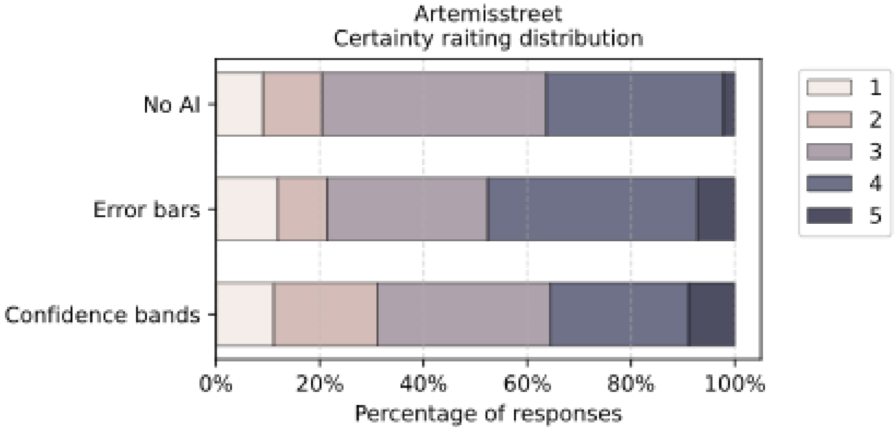


1) Low error – Artemisstreet



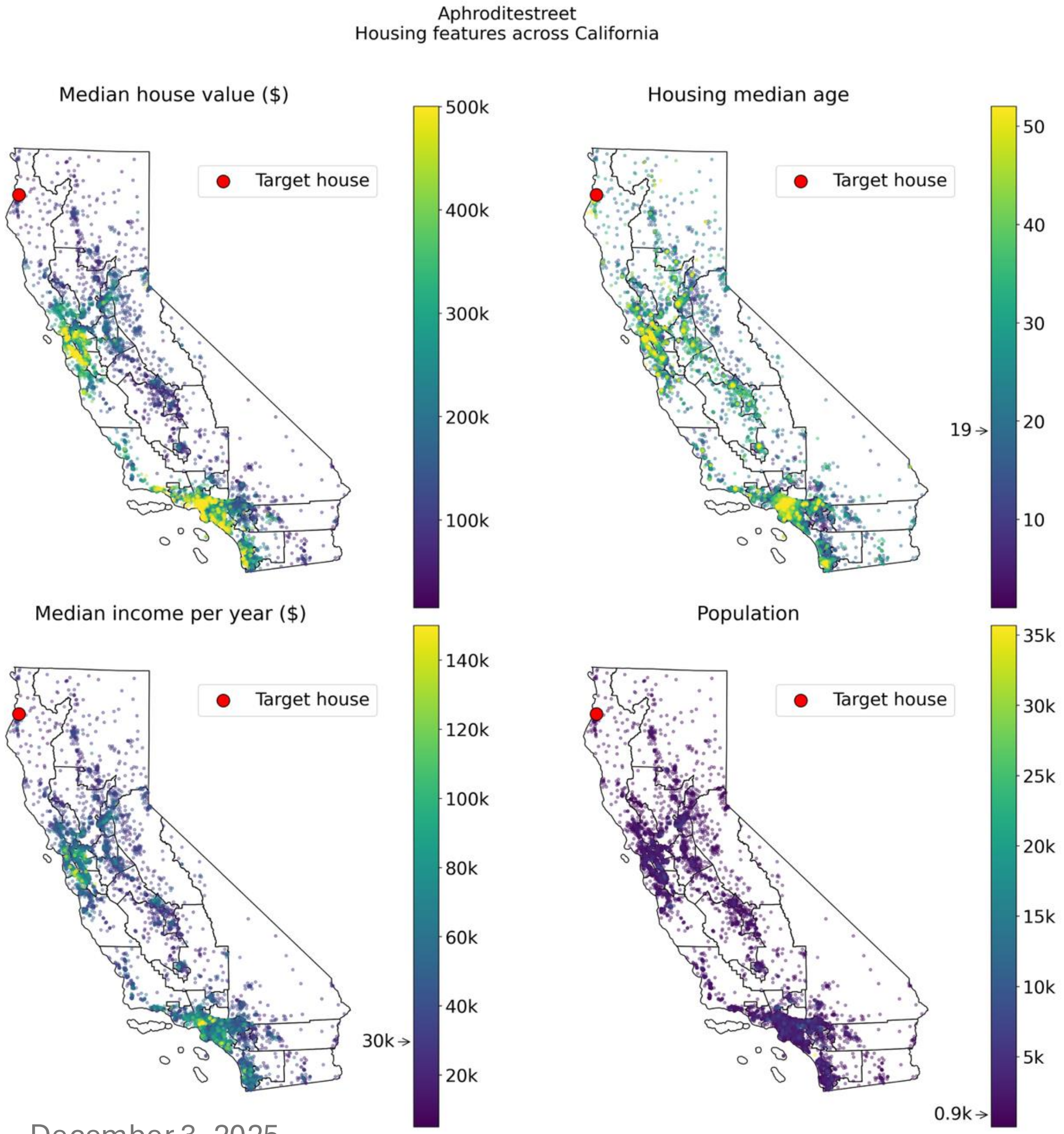


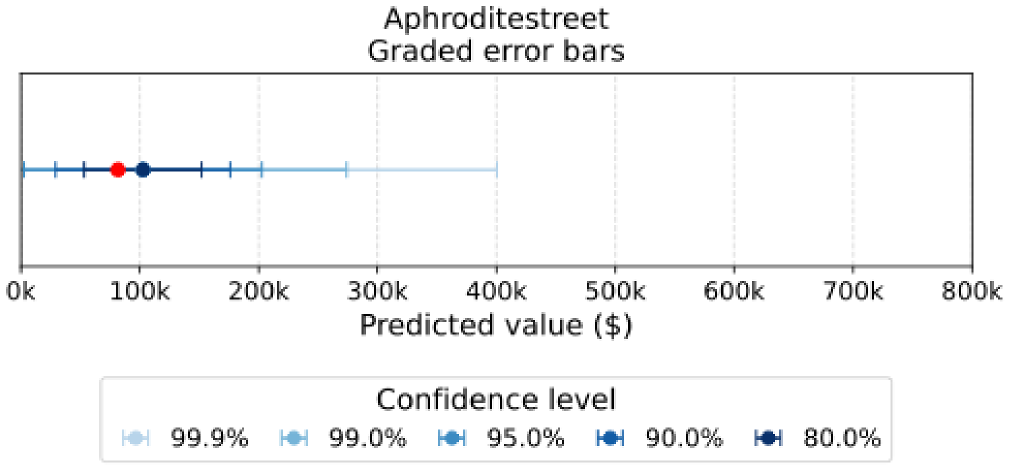
Estimates closest to the true value are marked with an “x” above the bars.

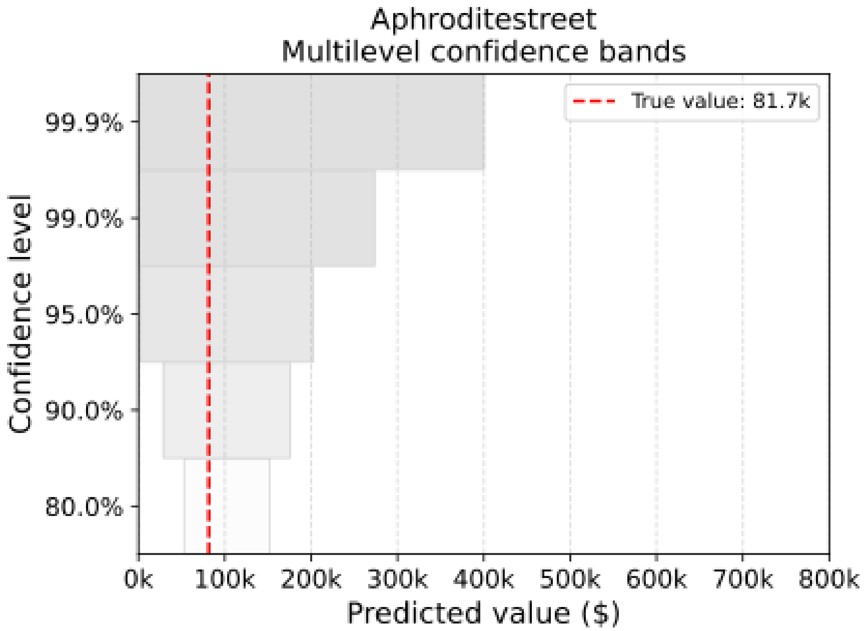


2) Low error – Aphroditestreet

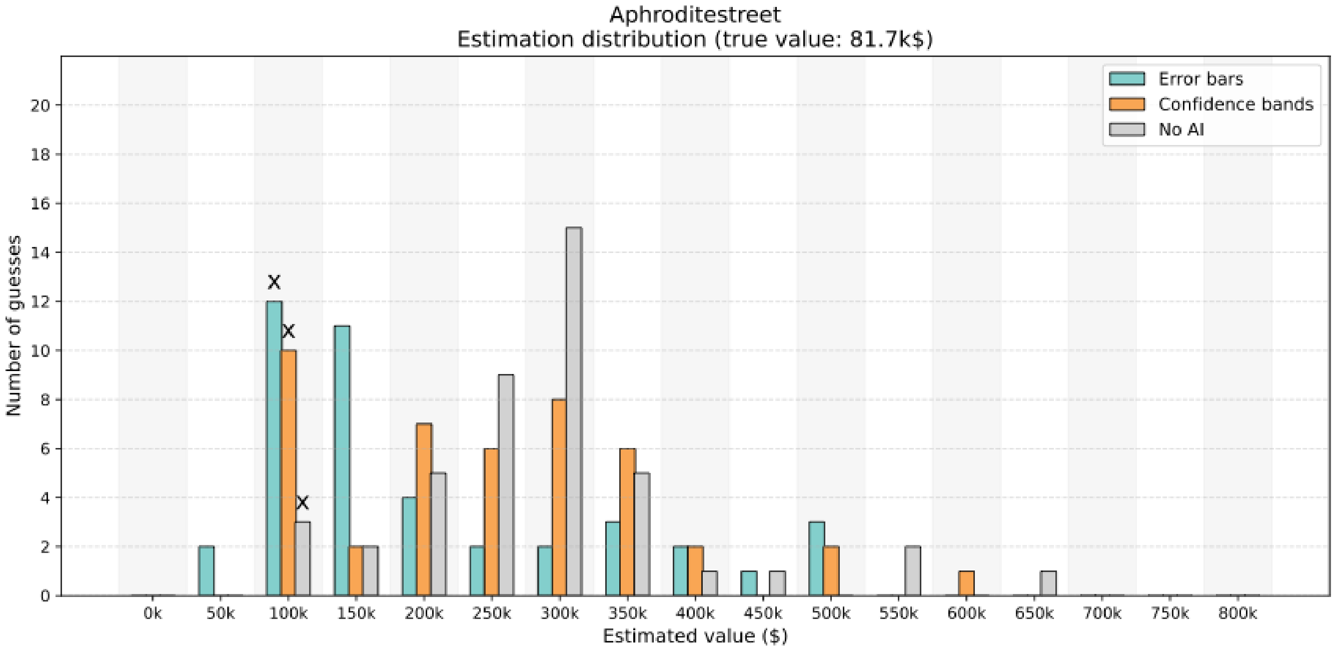
Survey interface

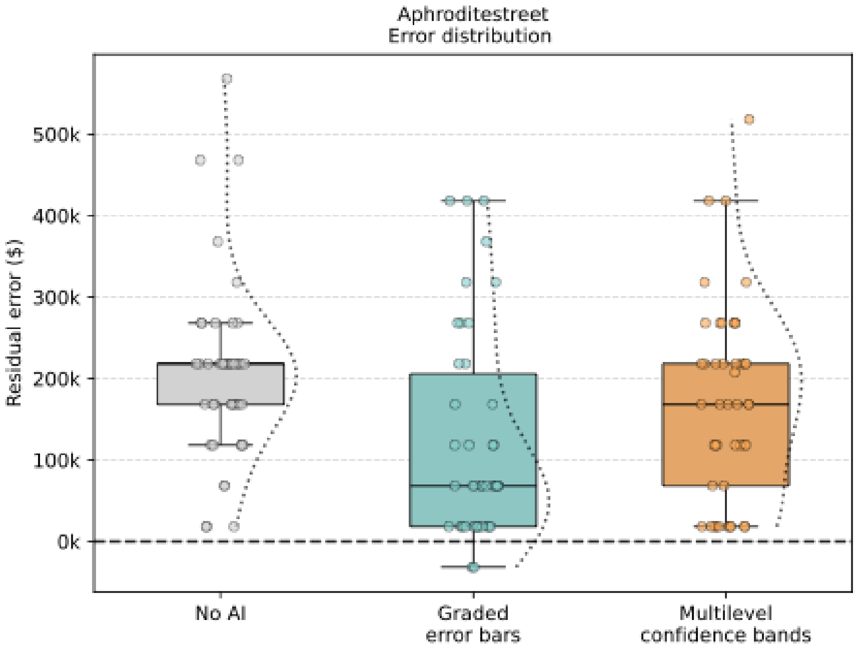


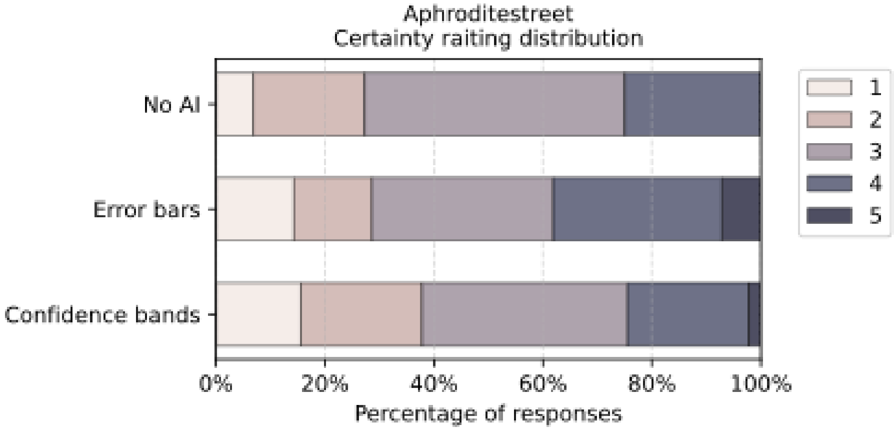




2) Low error – Aphroditestreet



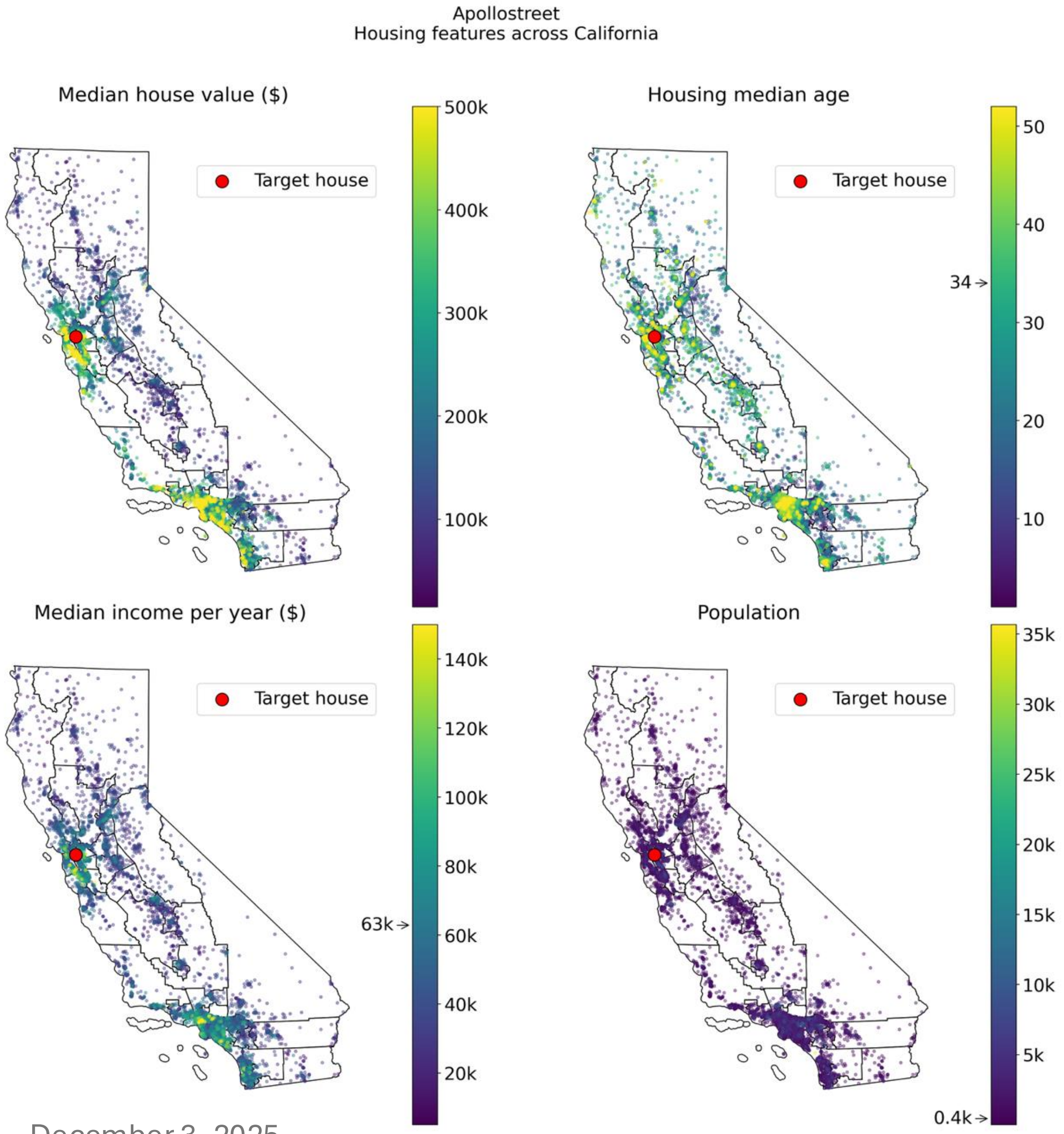


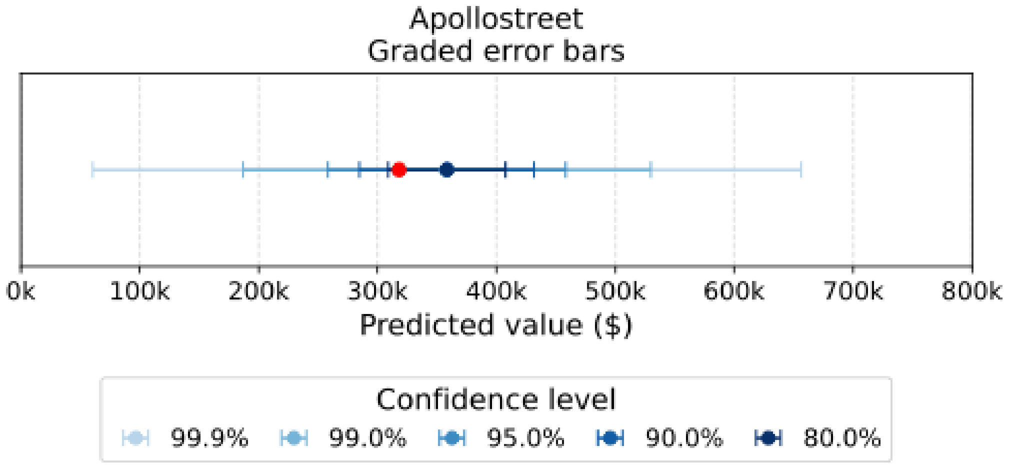


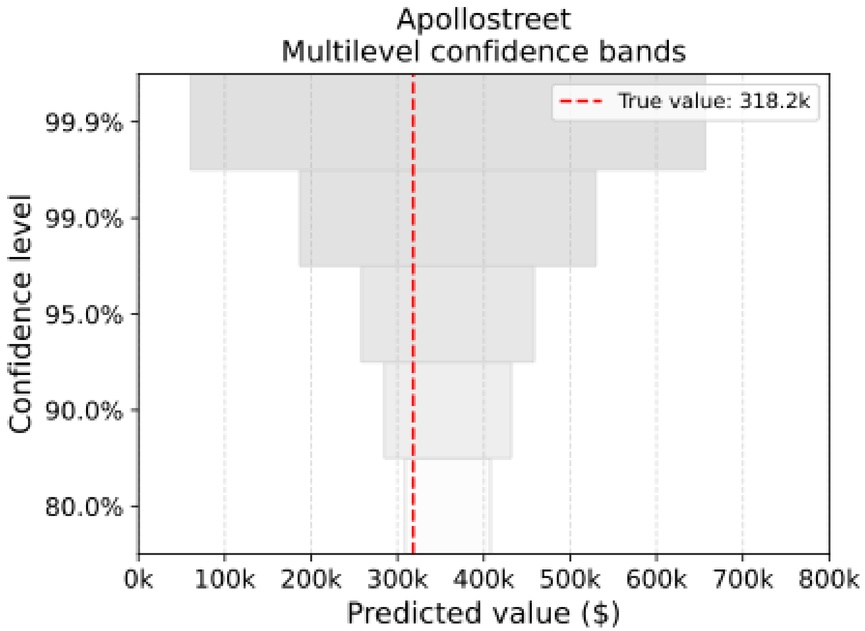
Estimates closest to the true value are marked with an "x" above the bars.

3) Low error – Apollostreet

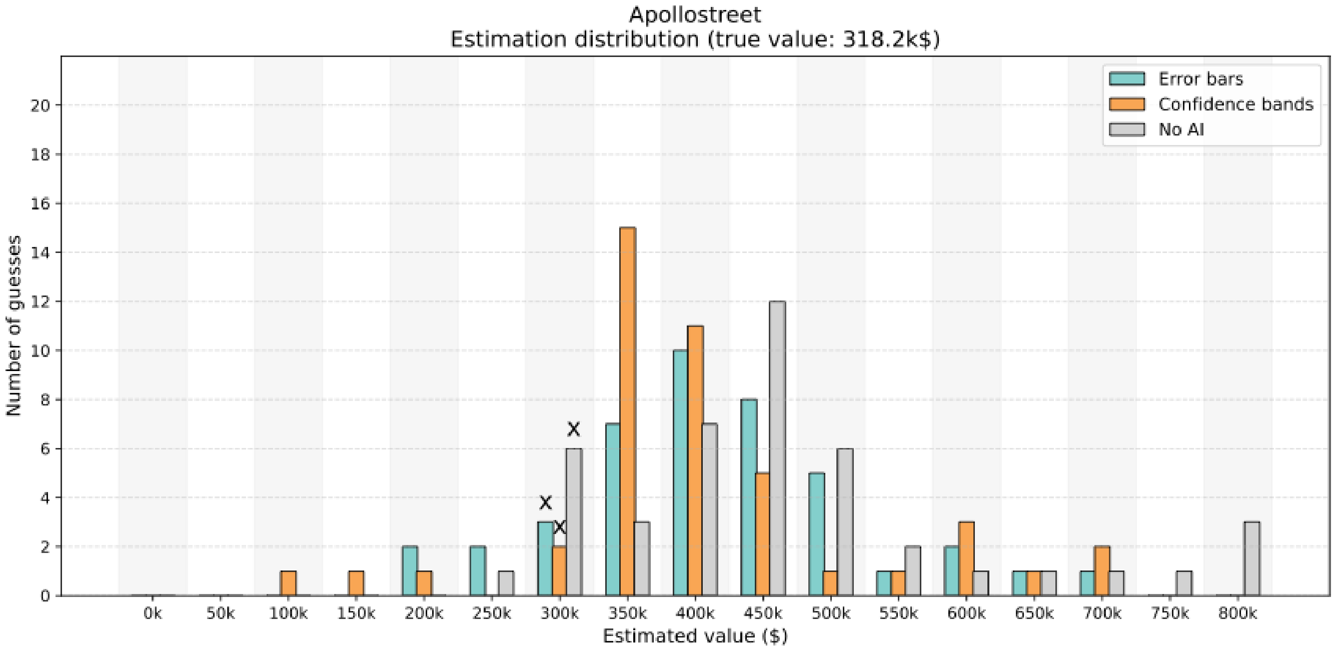
Survey interface


Apollostreet
Housing features across California

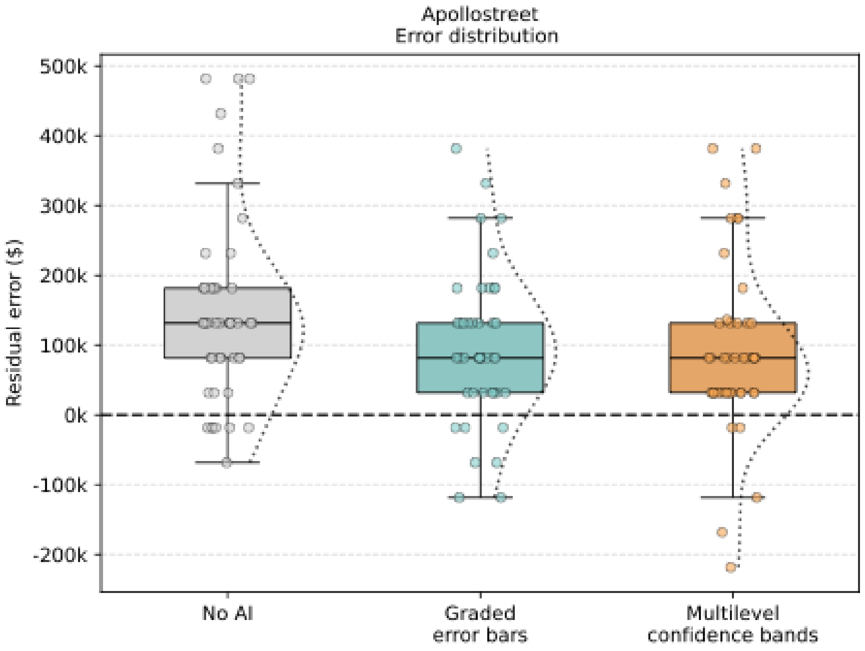

Apollostreet
Graded error bars

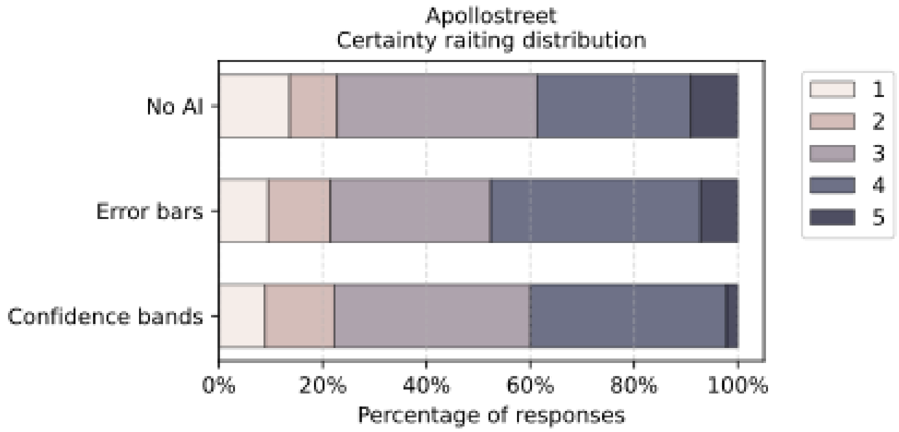

Apollostreet
Multilevel confidence bands

3) Low error – Apollostreet



Estimates closest to the true value are marked with an “x” above the bars.



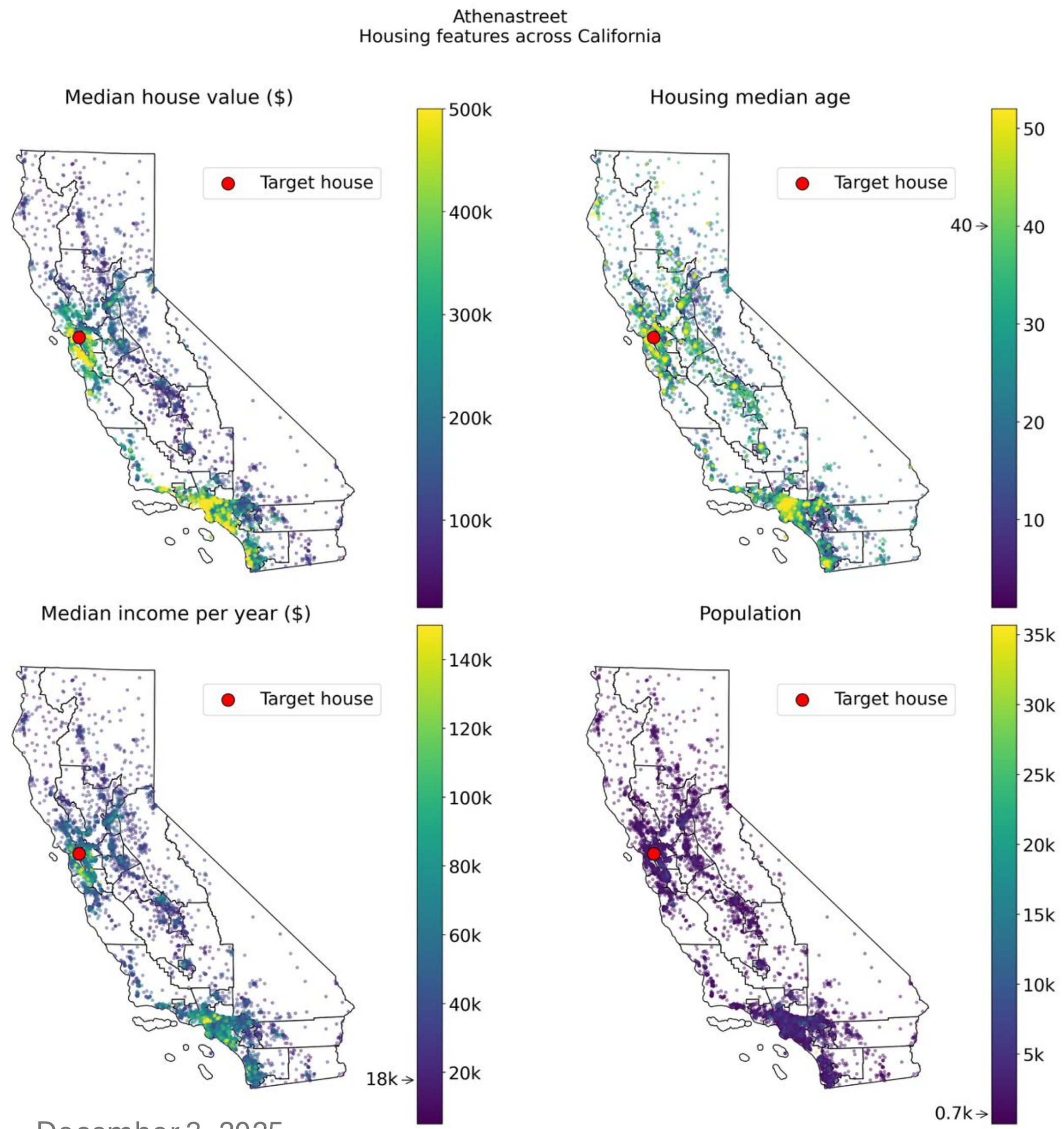


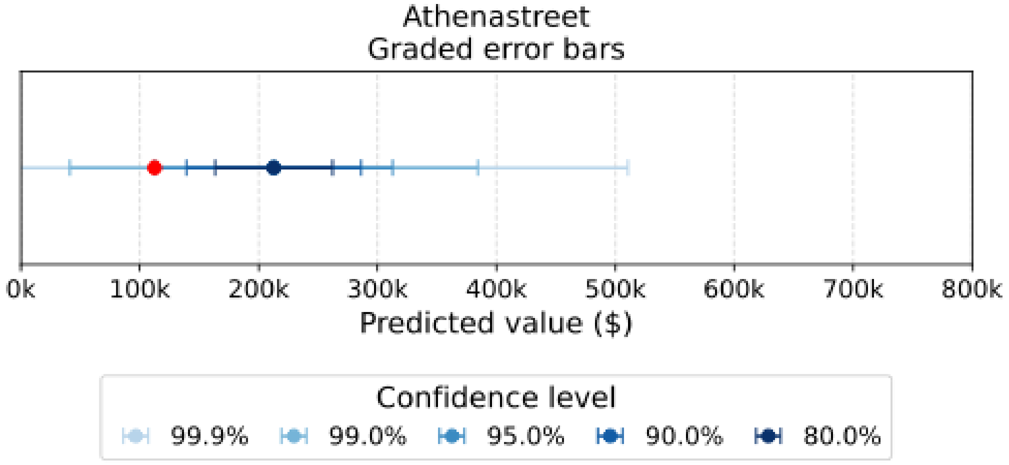
Task complexity: medium

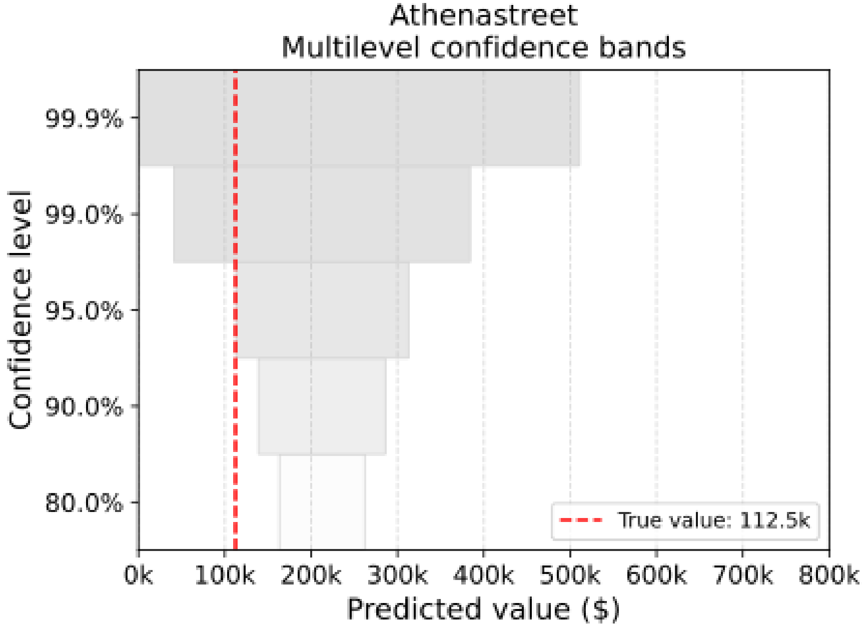
- 4) Athenastreet
- 5) Poseidonstreet
- 6) Aresstreet

4) Medium error – Athenastreet

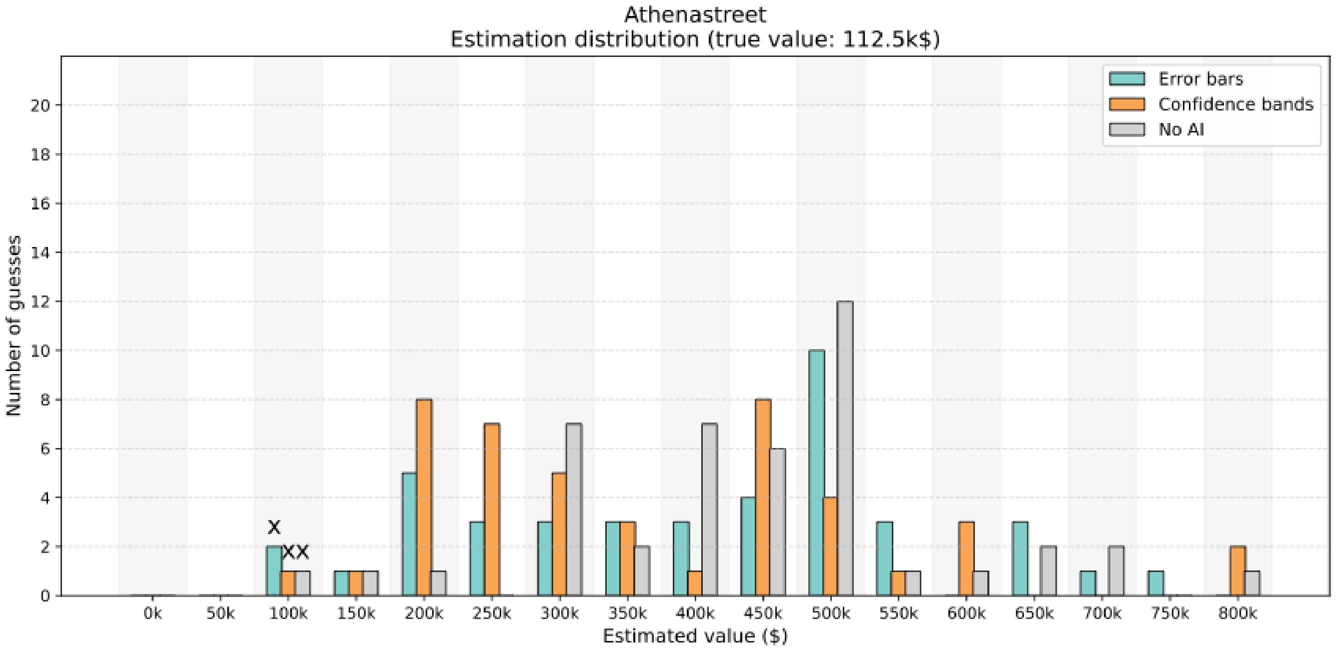
Survey interface

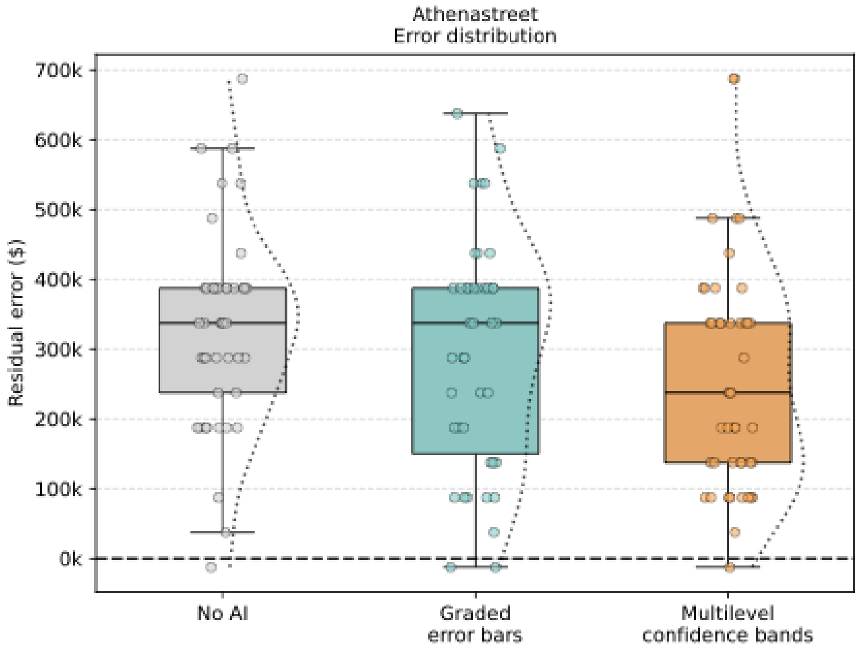




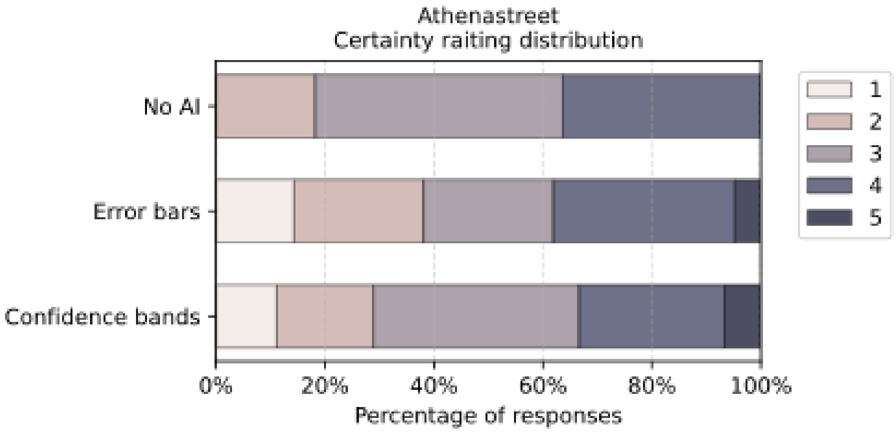


4) Medium error – Athenastreet



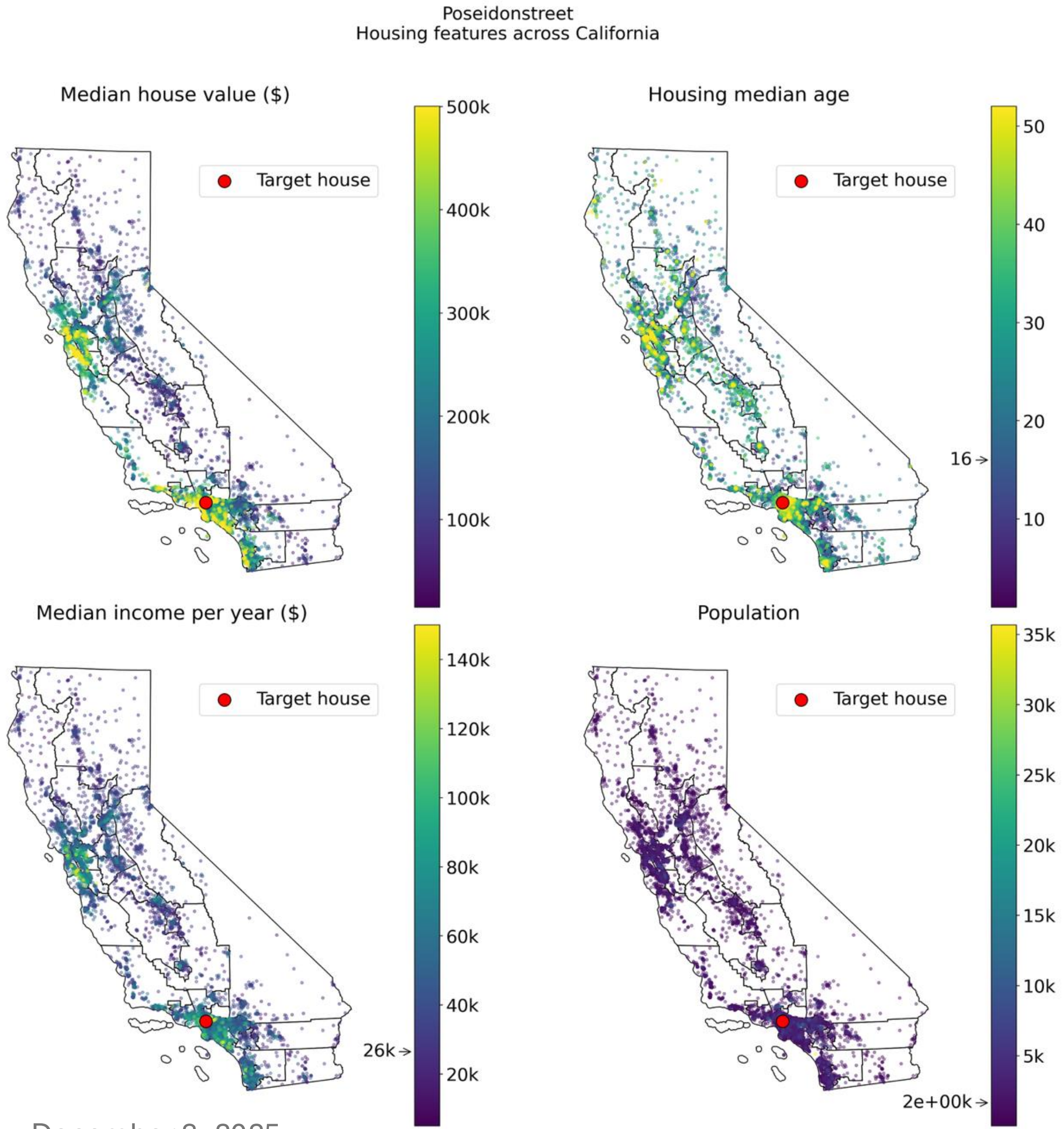


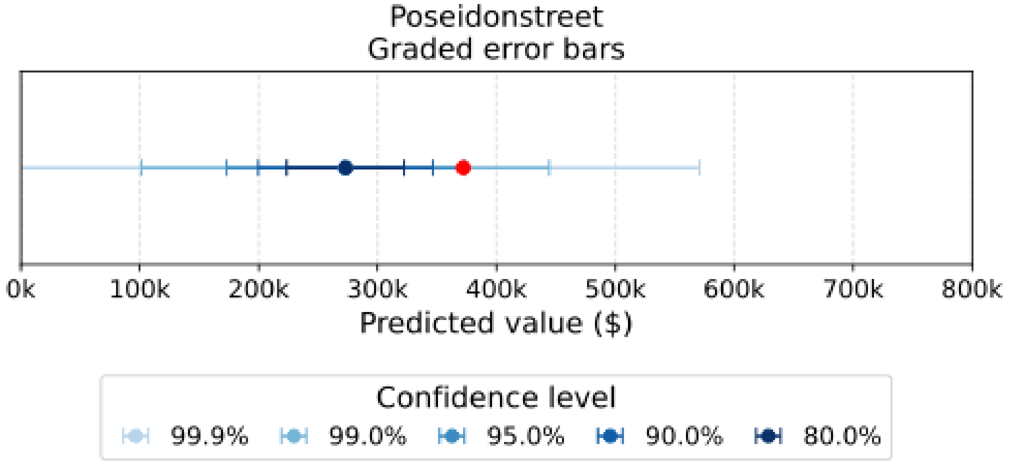
Estimates closest to the true value are marked with an “x” above the bars.

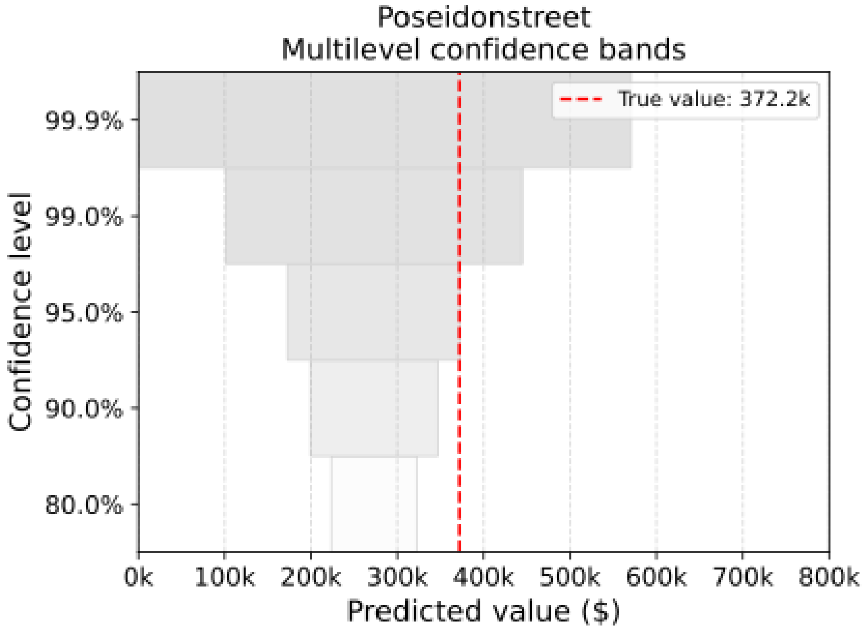


5) Medium error – Poseidonstreet

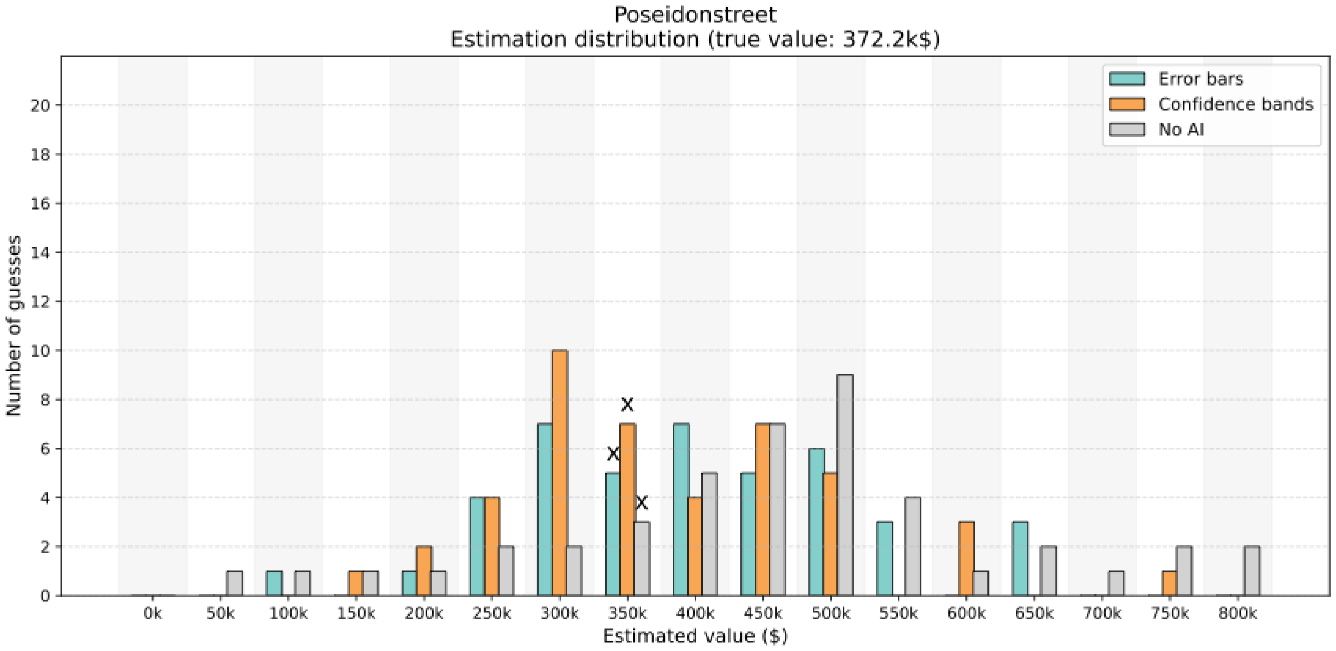
Survey interface



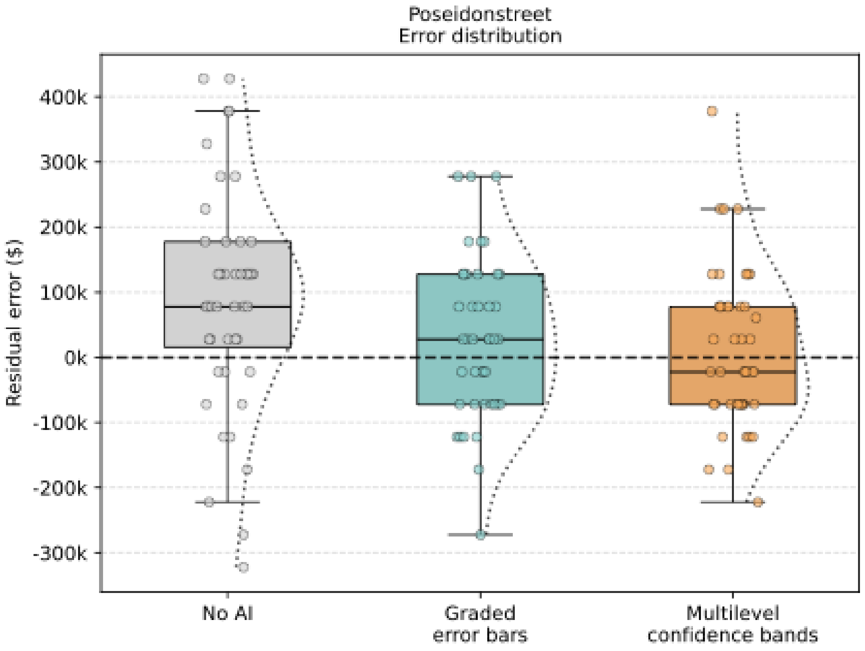


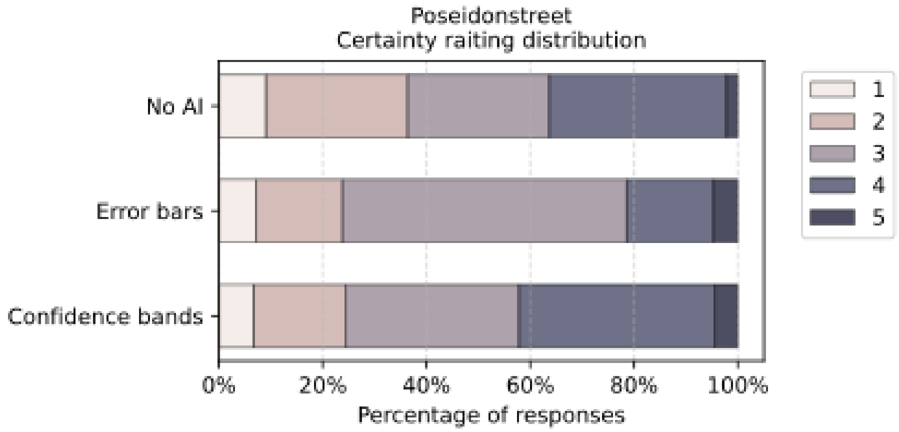


5) Medium error – Poseidonstreet



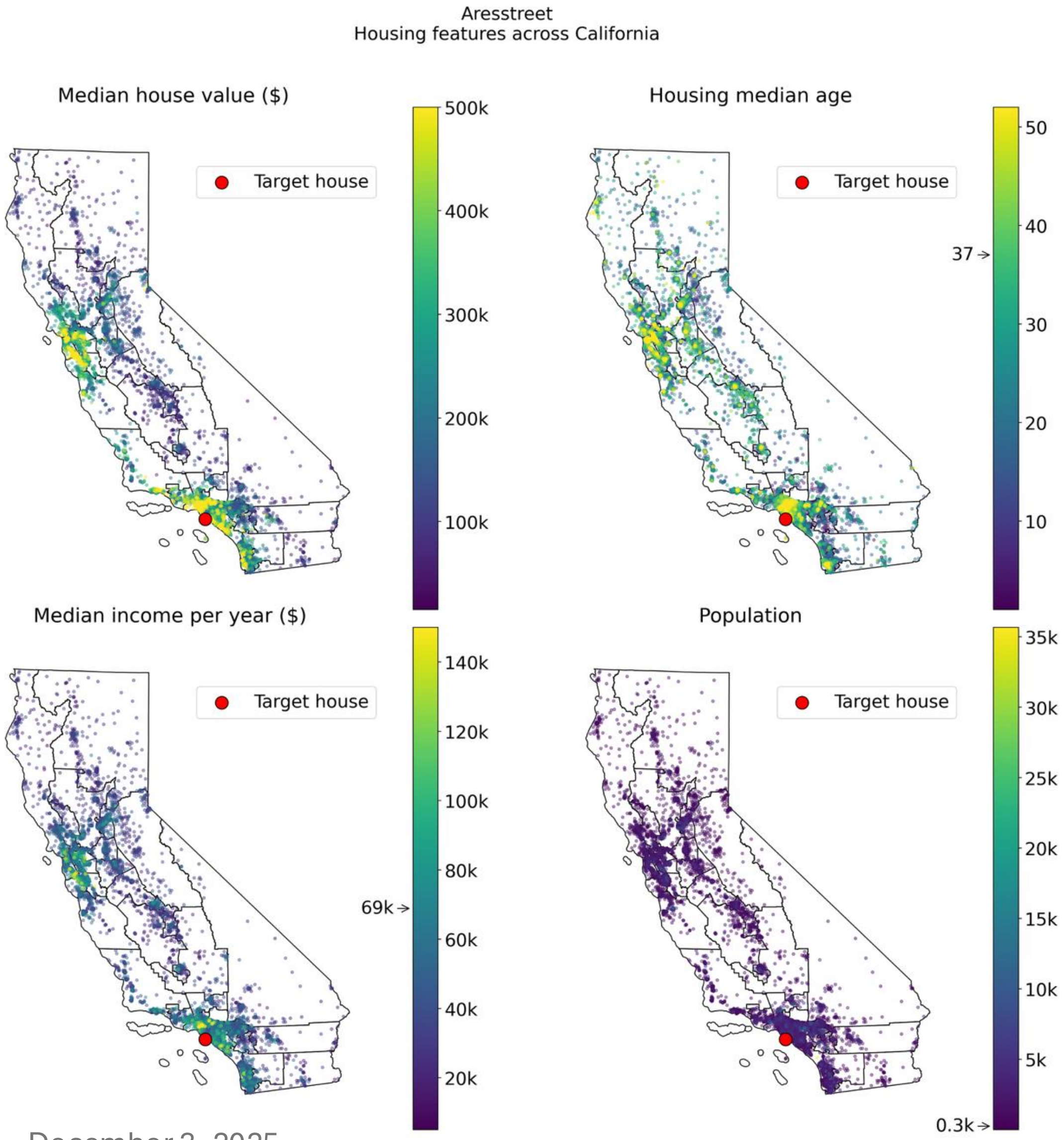
Estimates closest to the true value are marked with an “x” above the bars.

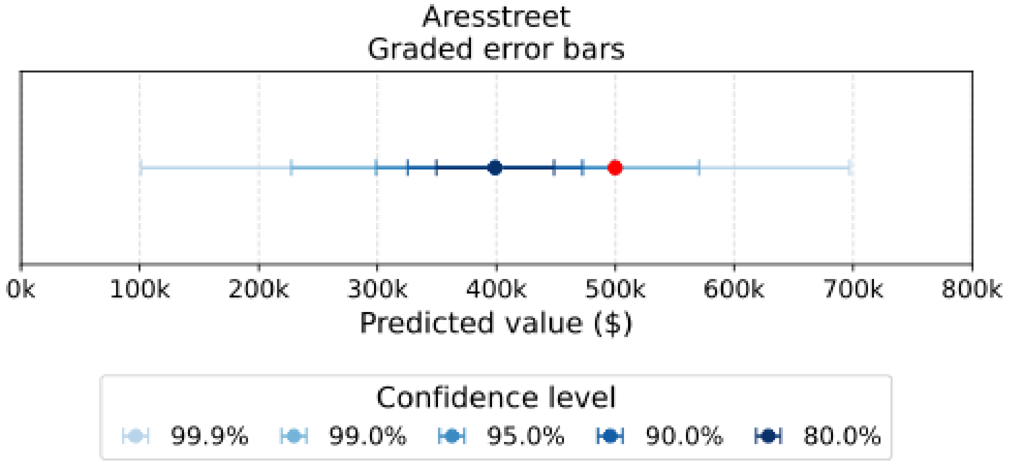


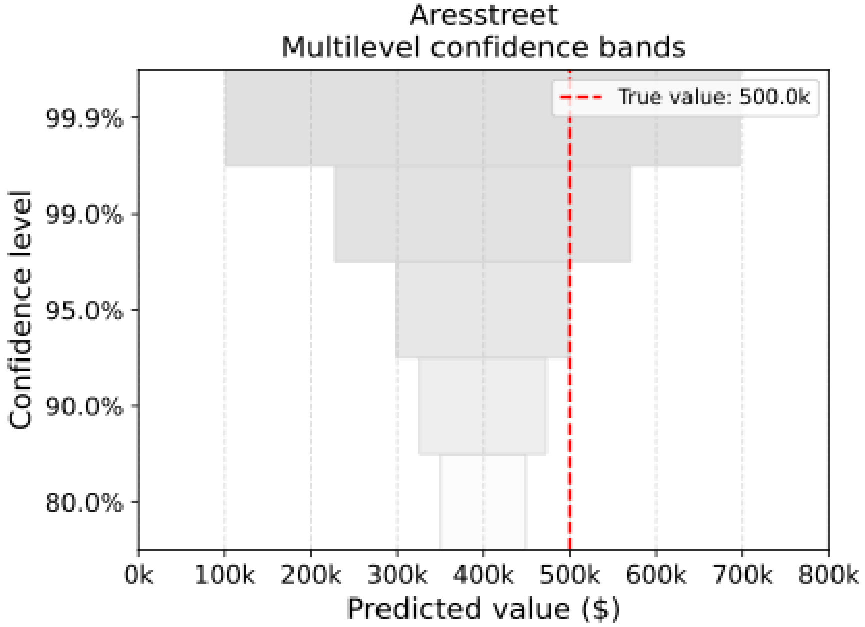


6) Medium error – Aresstreet

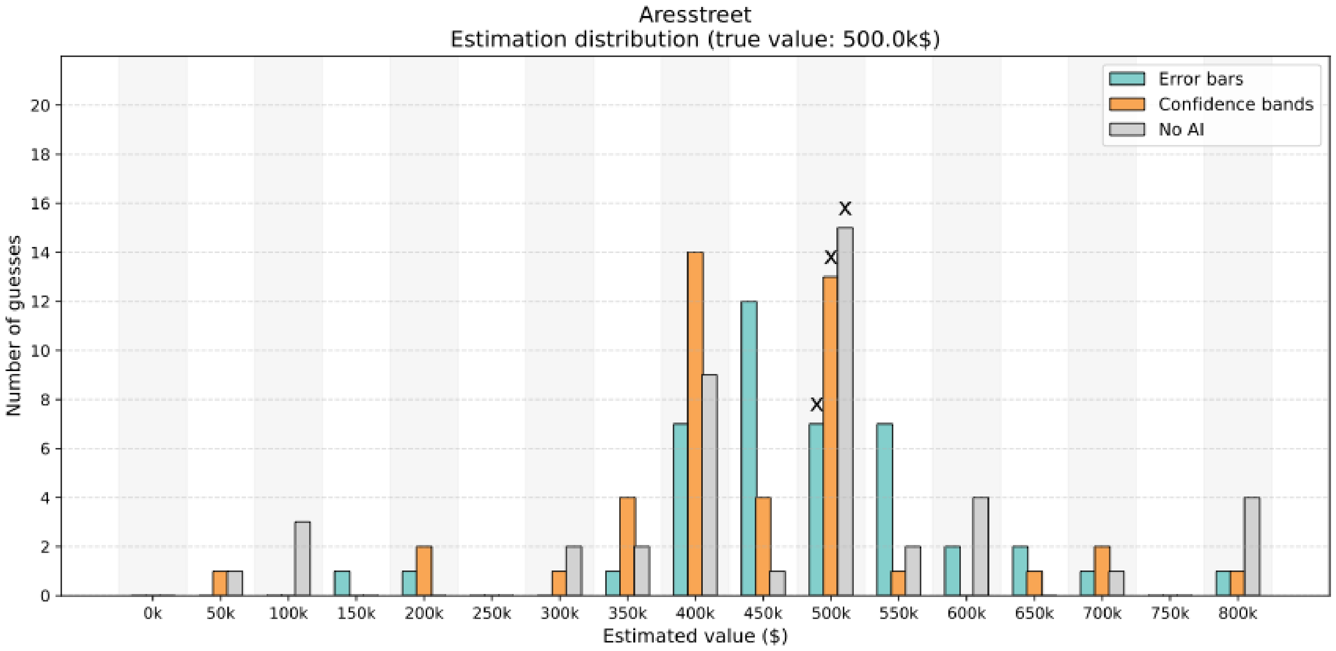
Survey interface



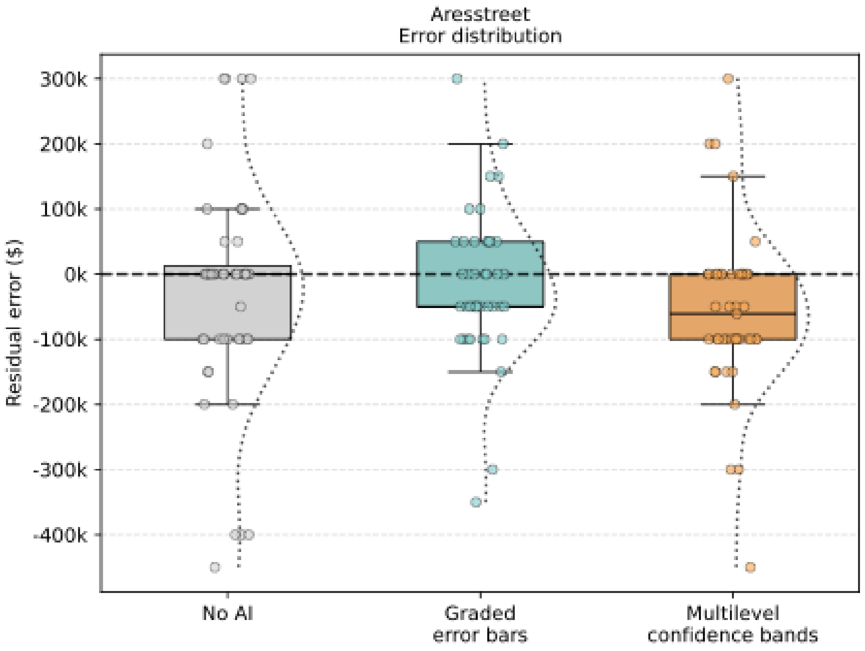


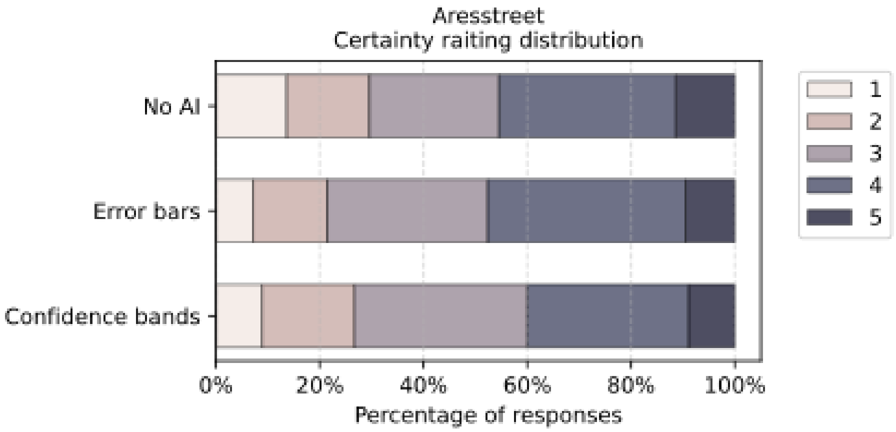


6) Medium error – Aresstreet



Estimates closest to the true value are marked with an “x” above the bars.



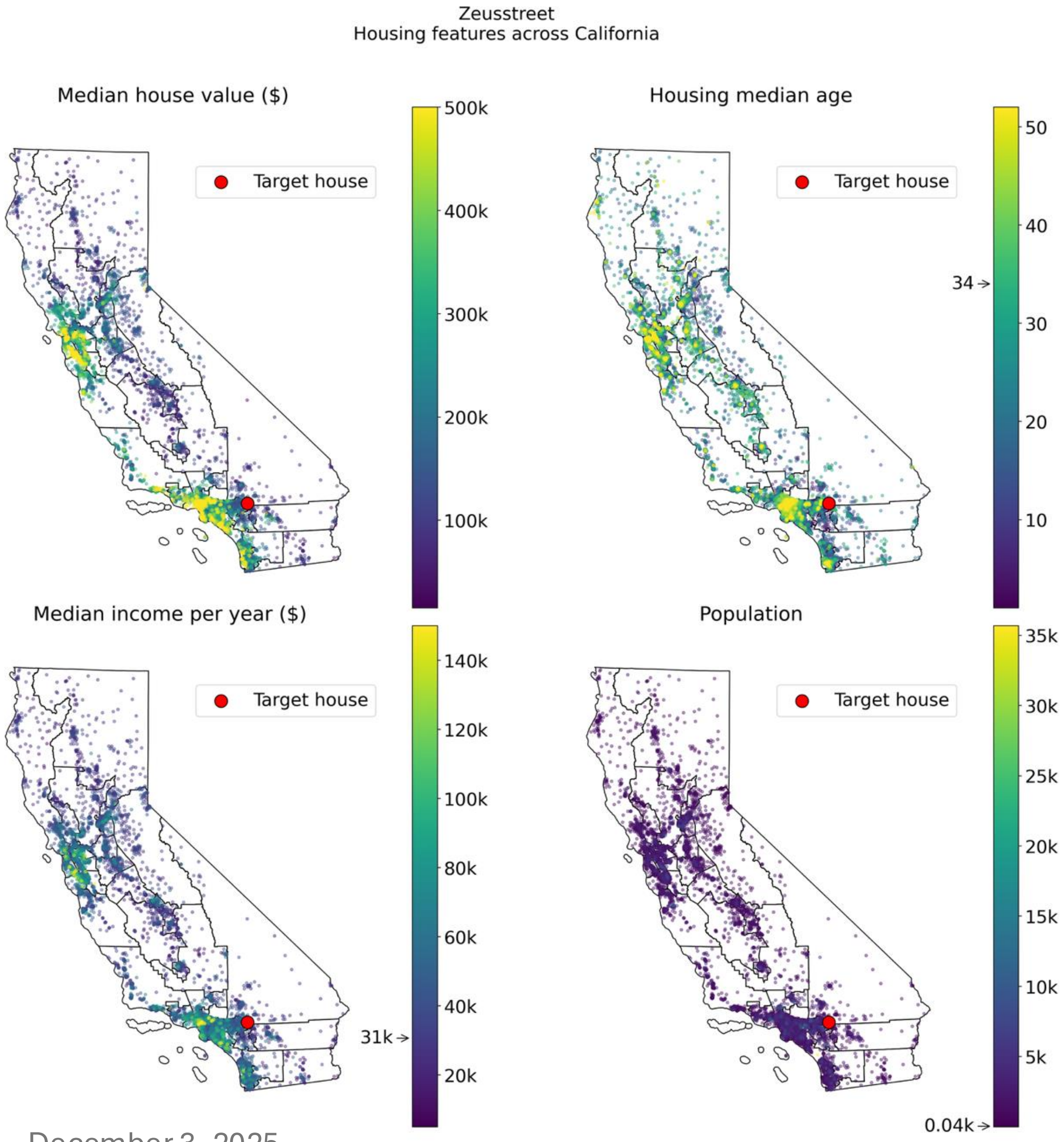


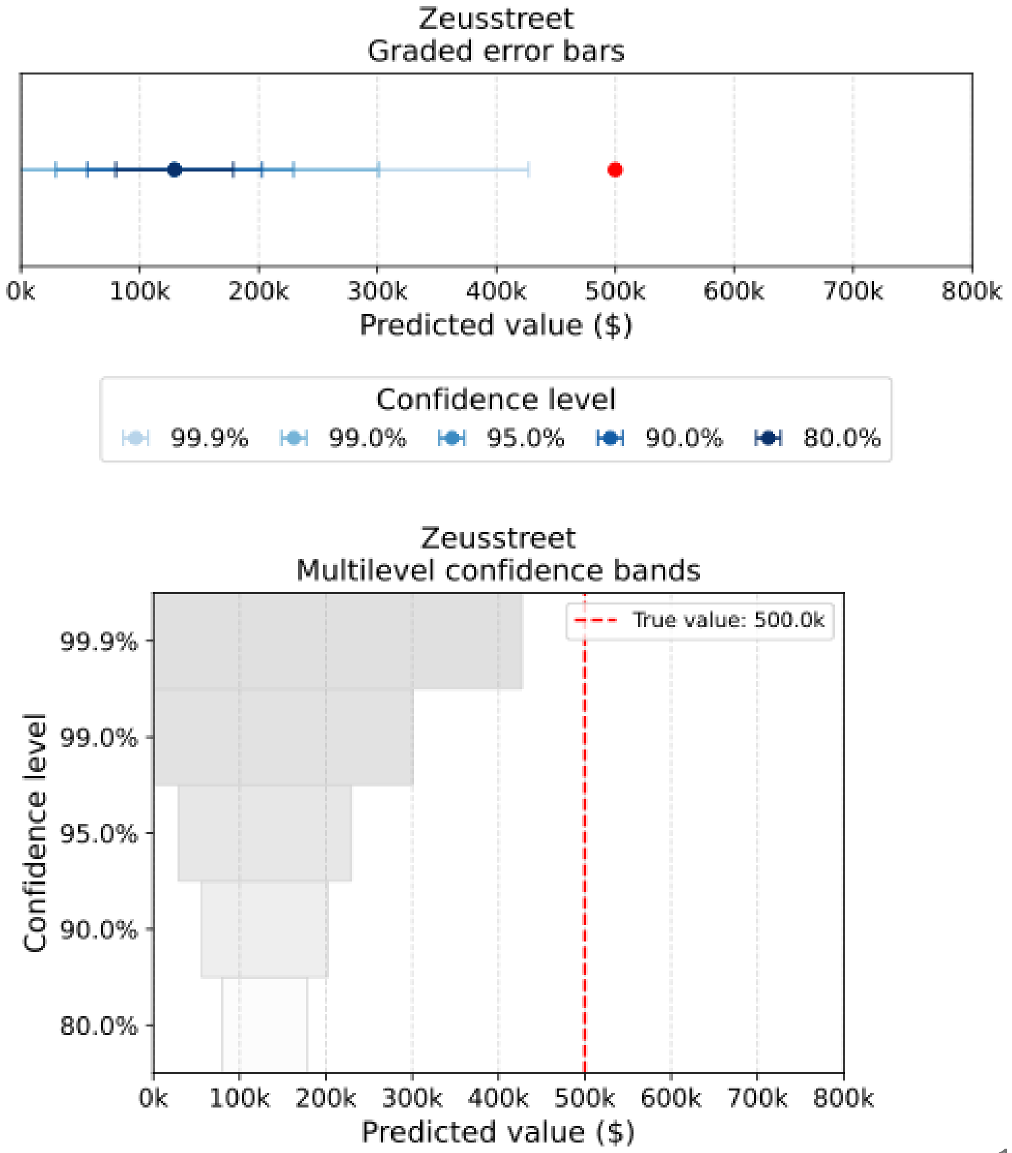
Task complexity: hard

- 7) Zeusstreet
- 8) Hephaestusstreet
- 9) Hestiastreet

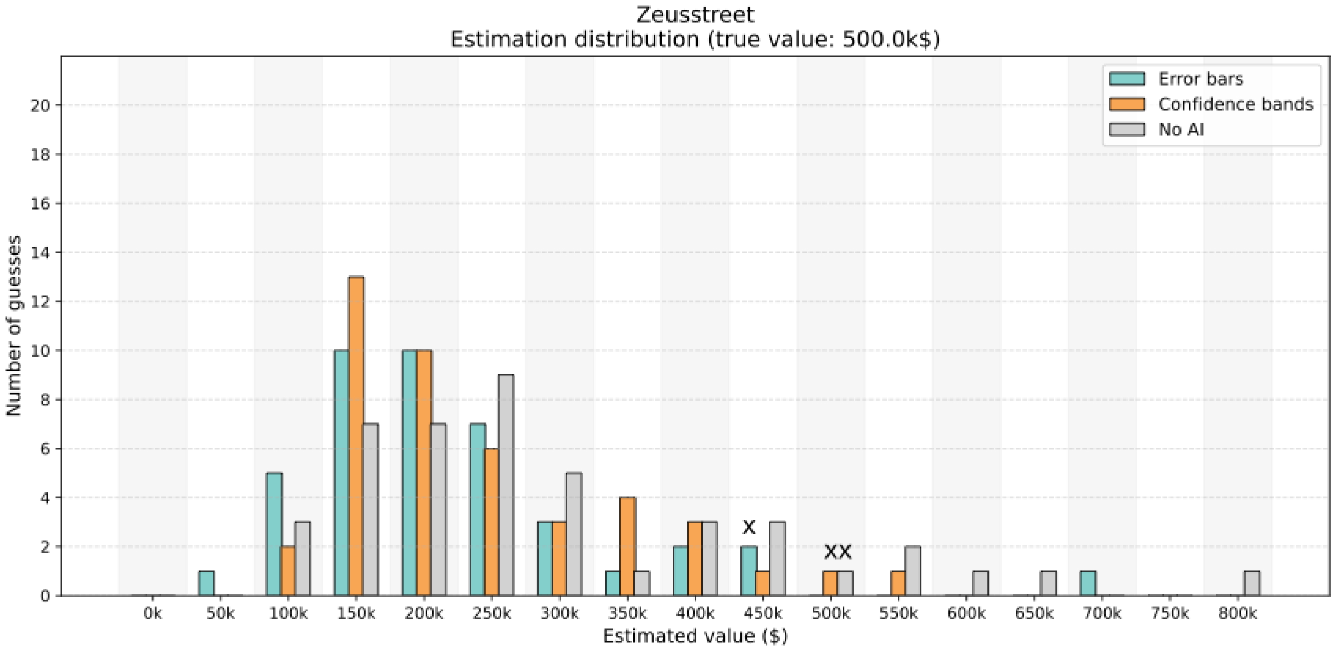
7) High error – Zeusstreet

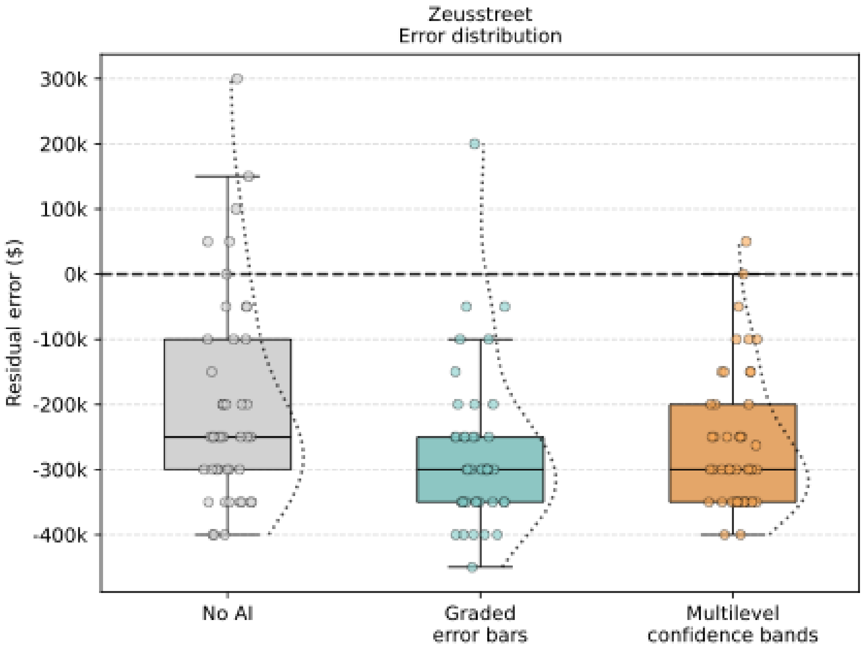
Survey interface

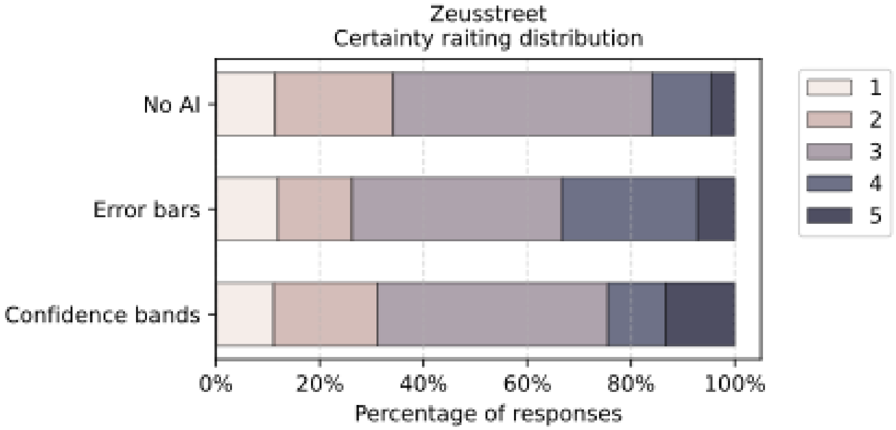




7) High error – Zeusstreet



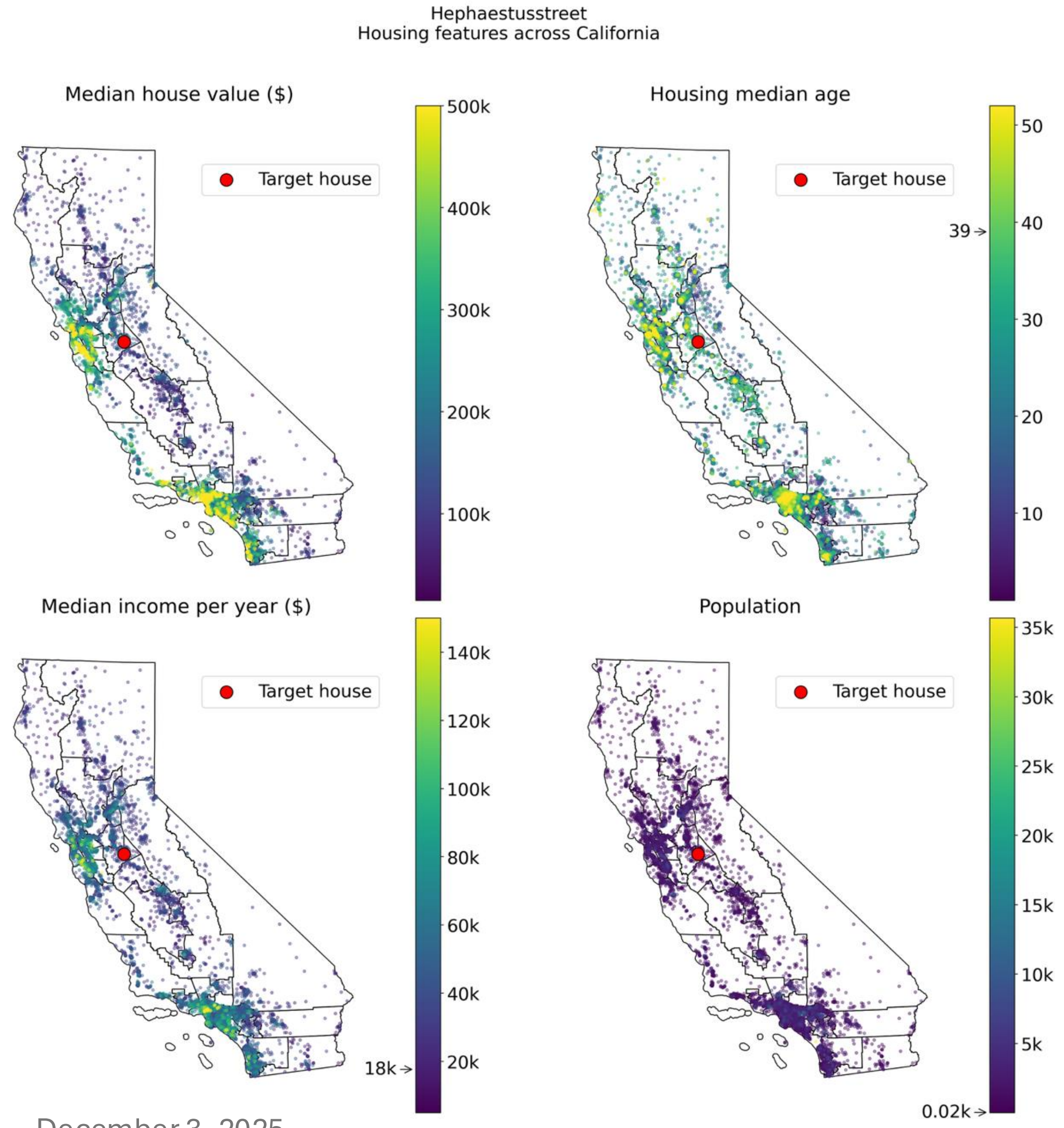


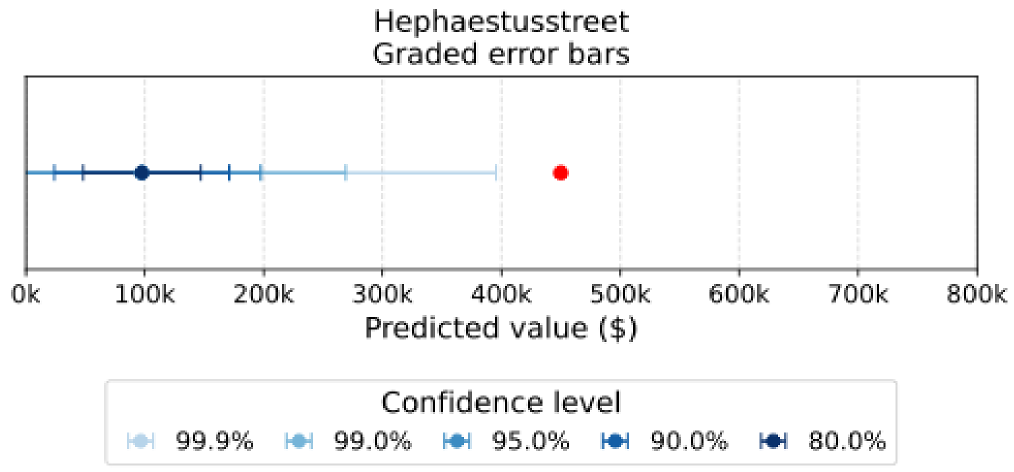


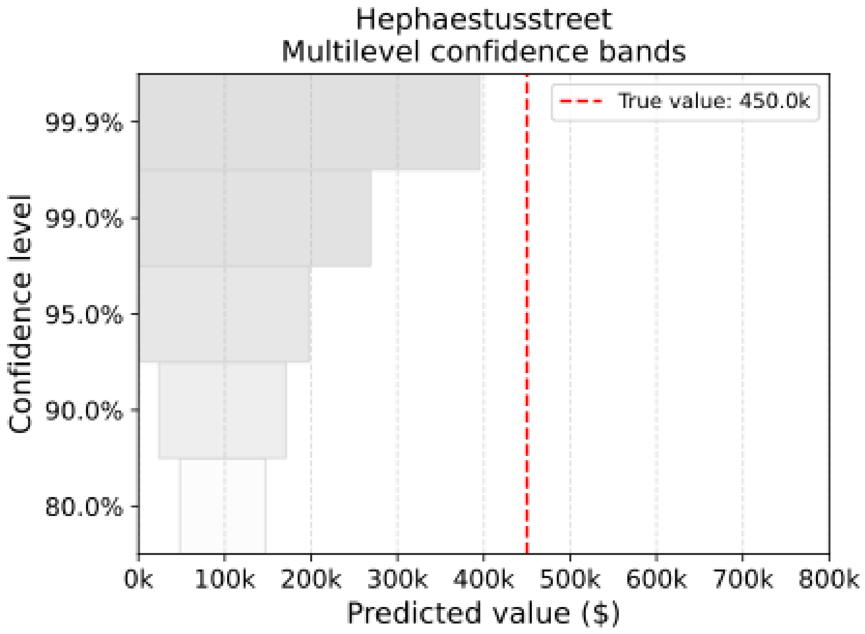
Estimates closest to the true value are marked with an “x” above the bars.

8) High error – Hephaestusstreet

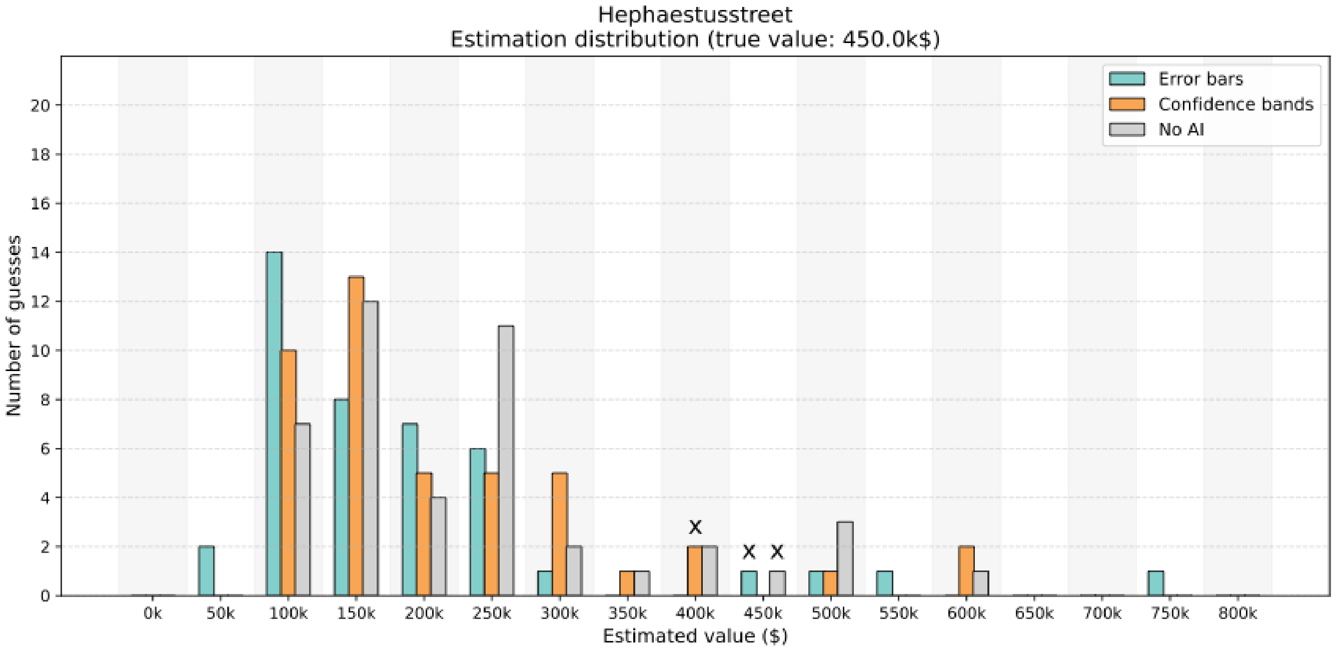
Survey interface

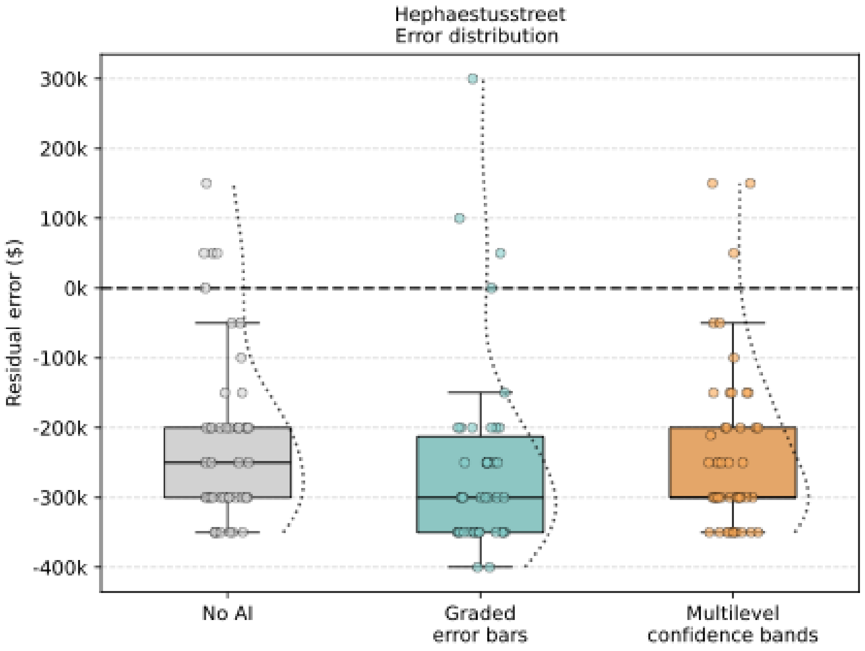


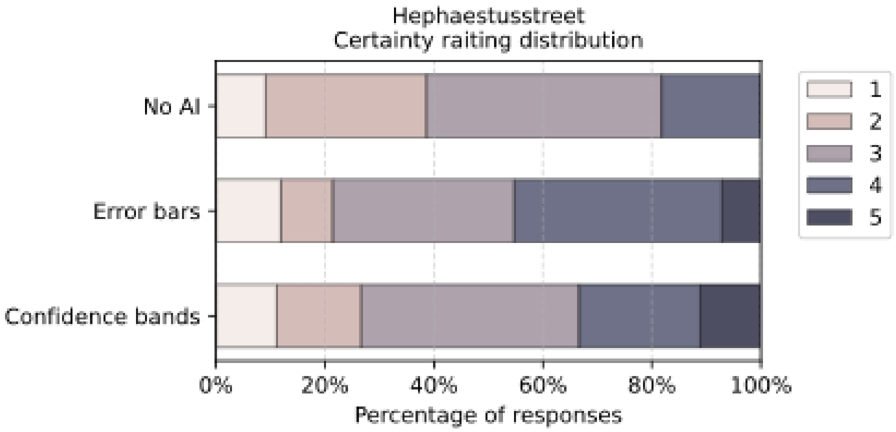




8) High error – Hephaestusstreet



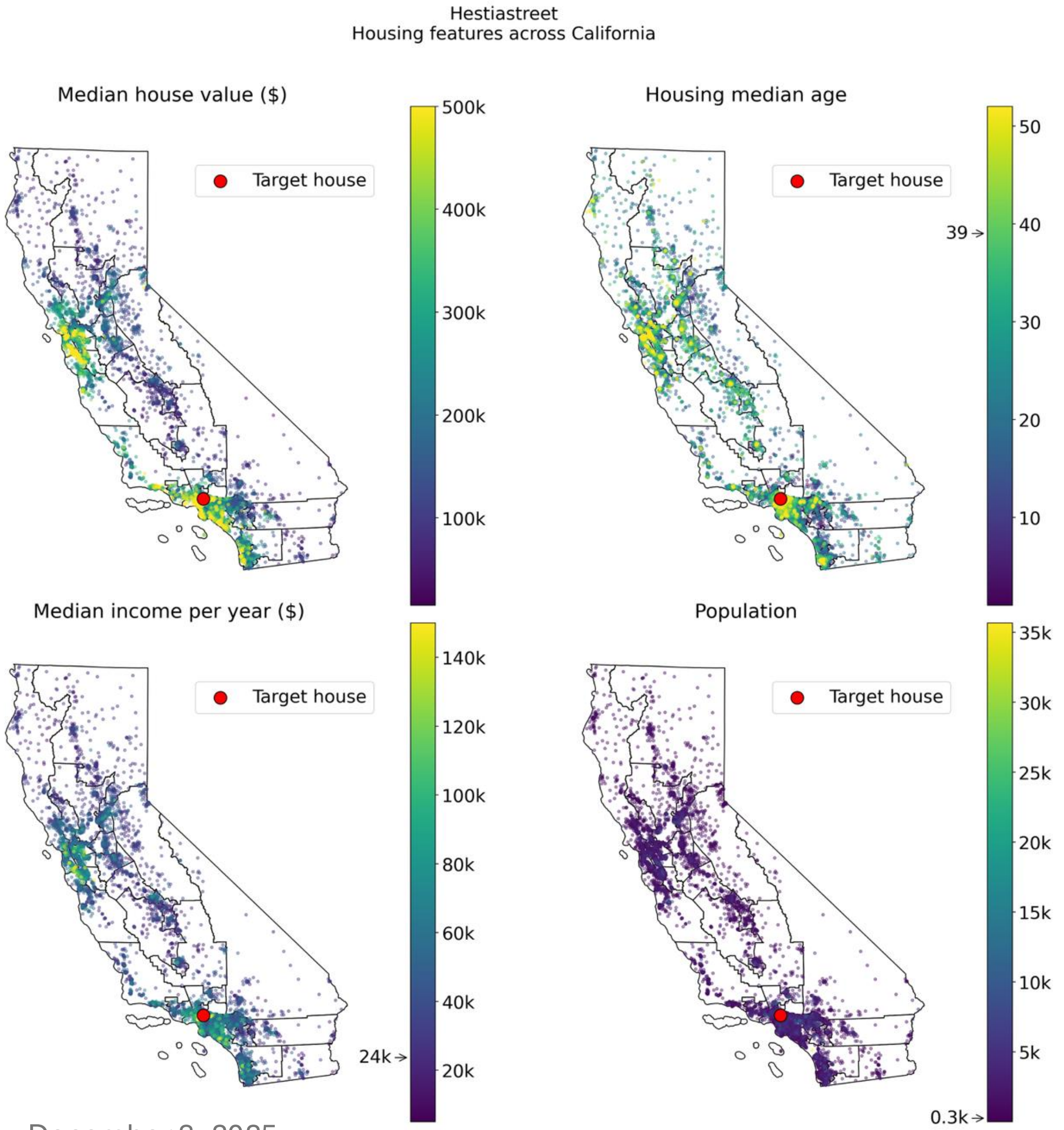


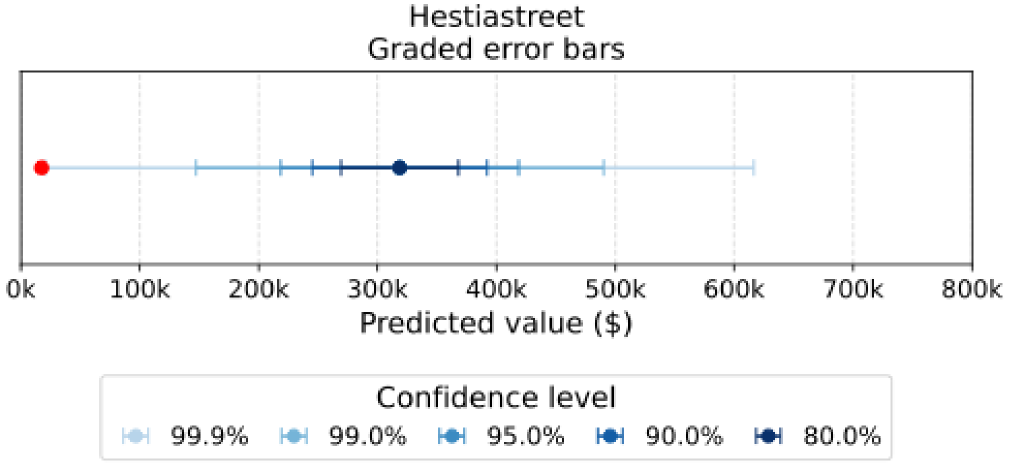


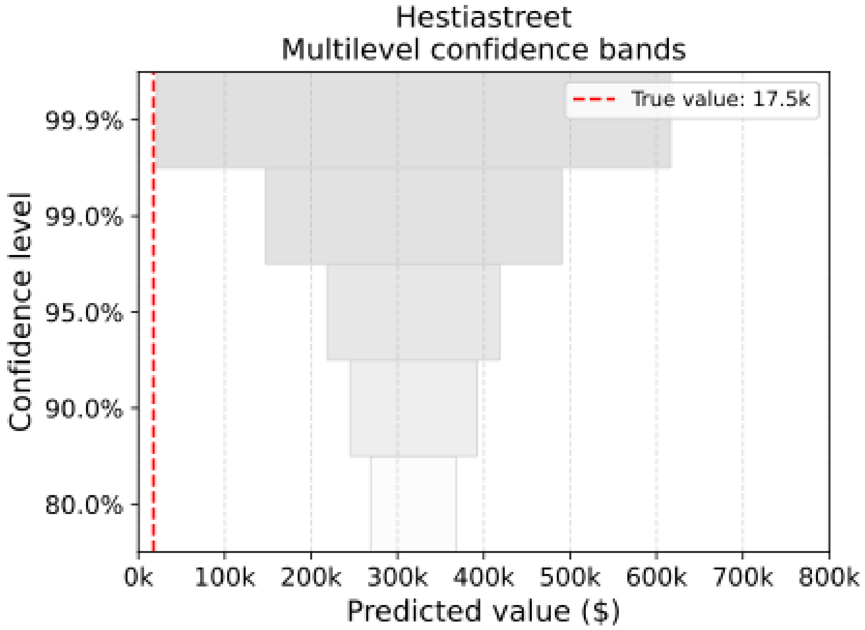
Estimates closest to the true value are marked with an “x” above the bars.

9) High error – Hestiastreet

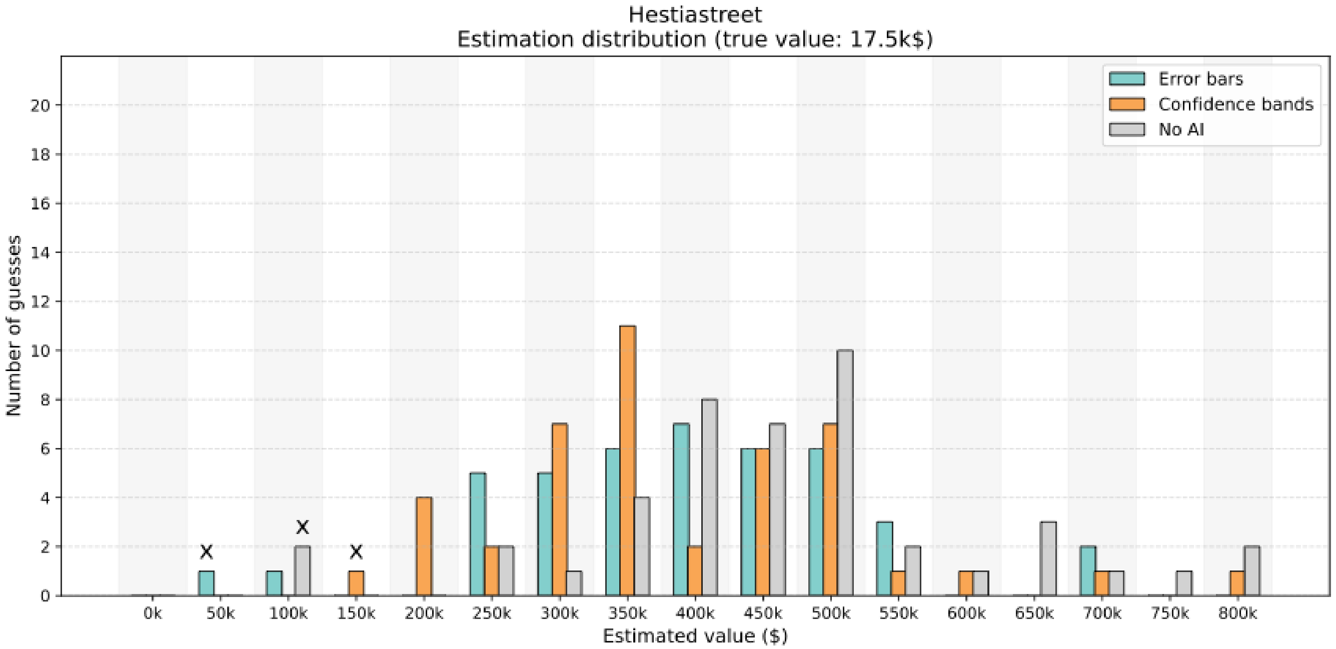
Survey interface

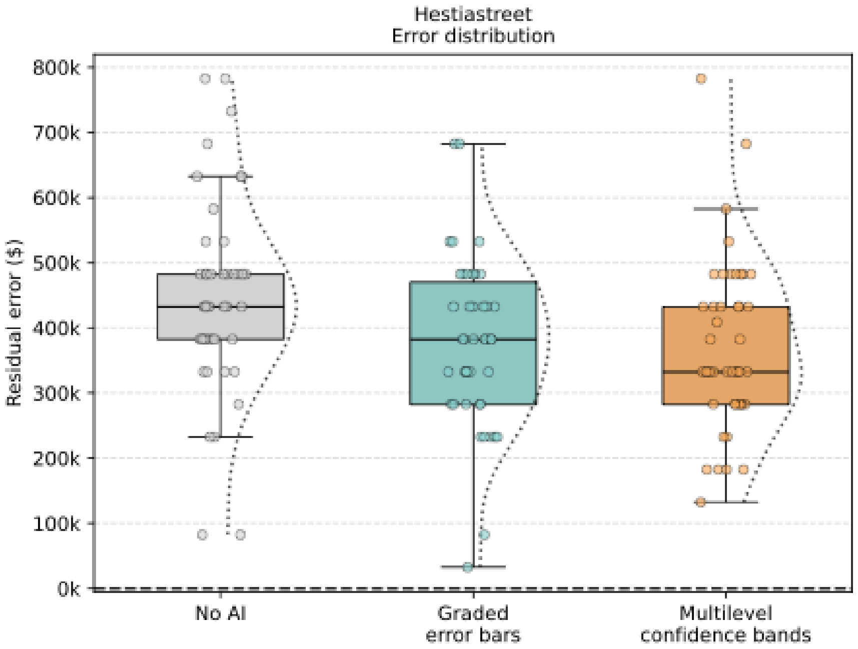


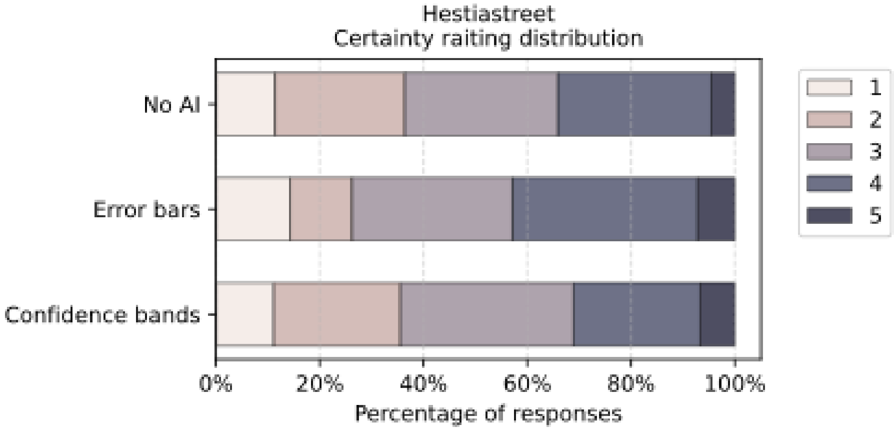




9) High error – Hestiastreet



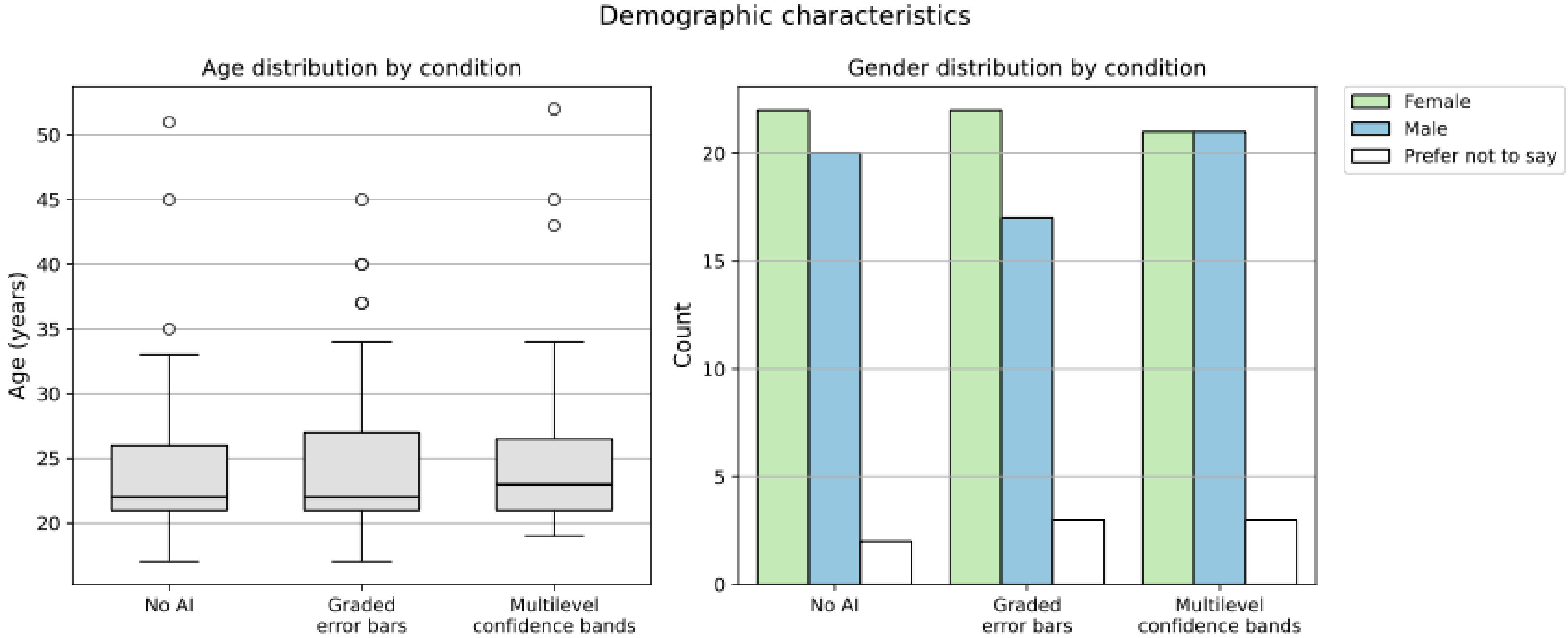




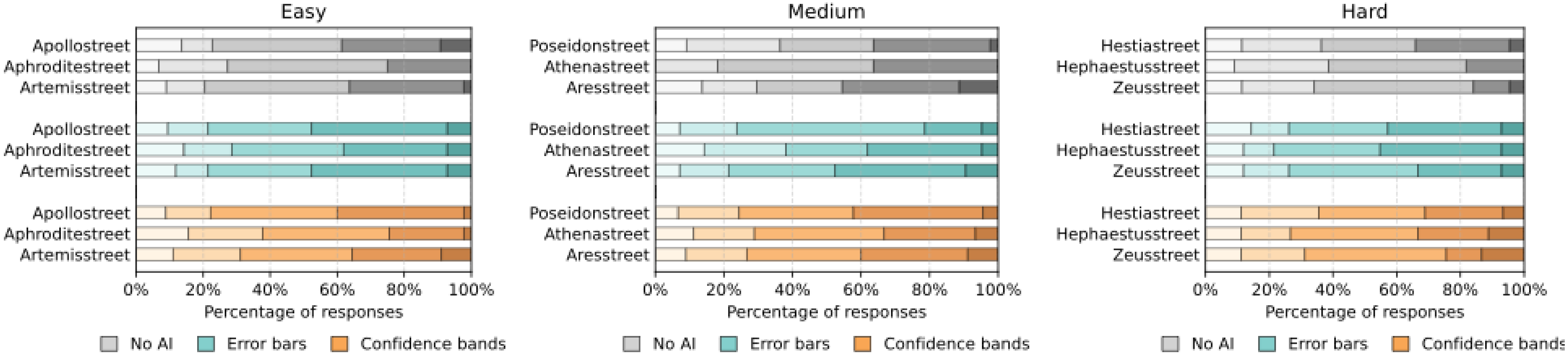
Estimates closest to the true value are marked with an "x" above the bars.

2. Additional charts

Survey Data: Demographic Statistics

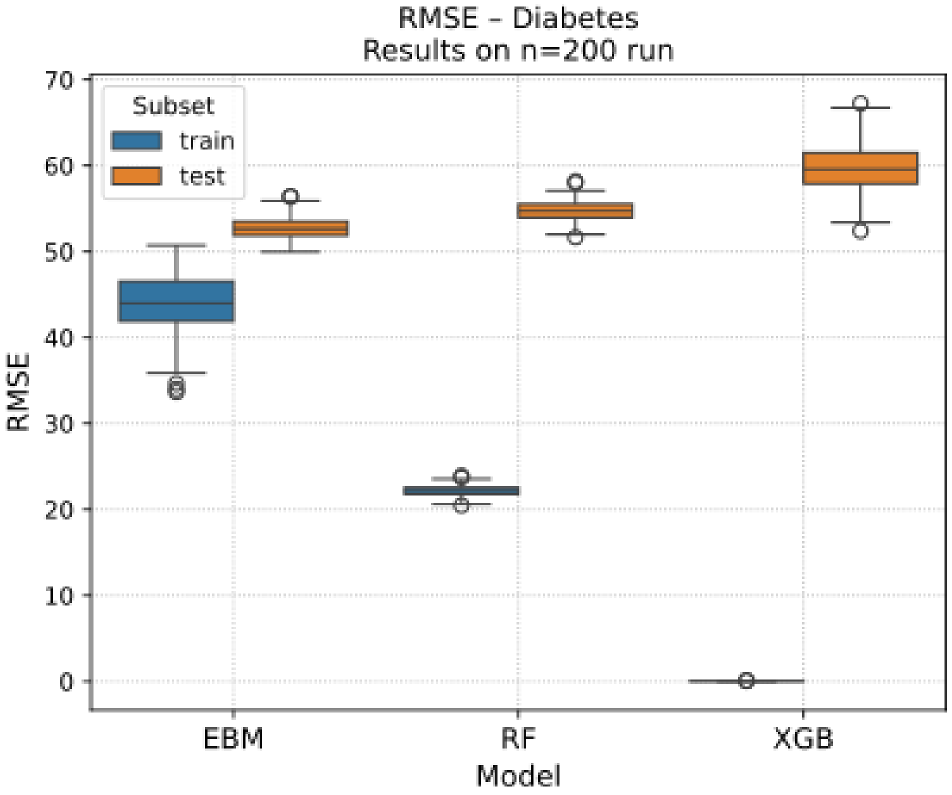


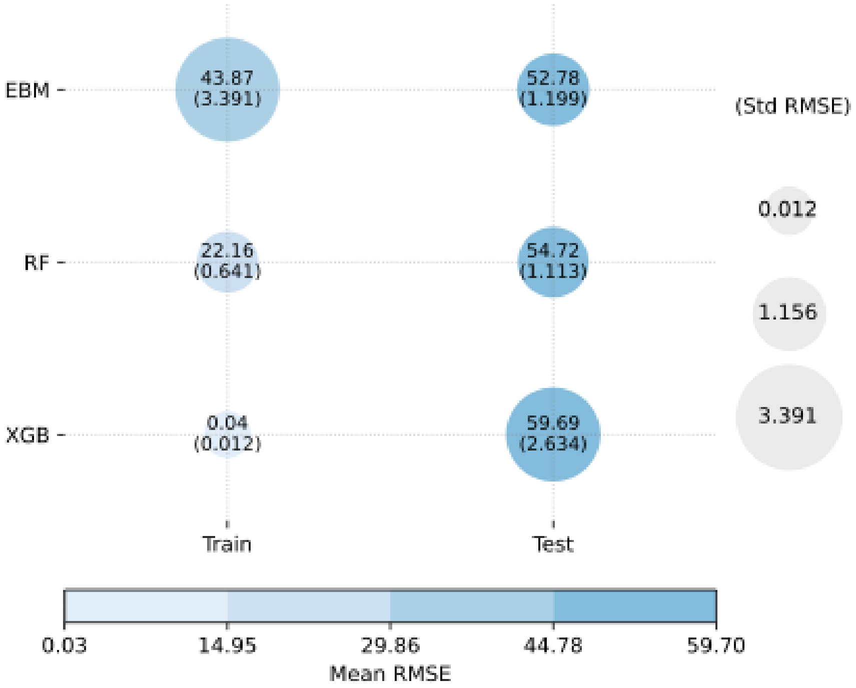
Survey Results: Certainty Ratings by House, Visualization Condition and Complexity Level



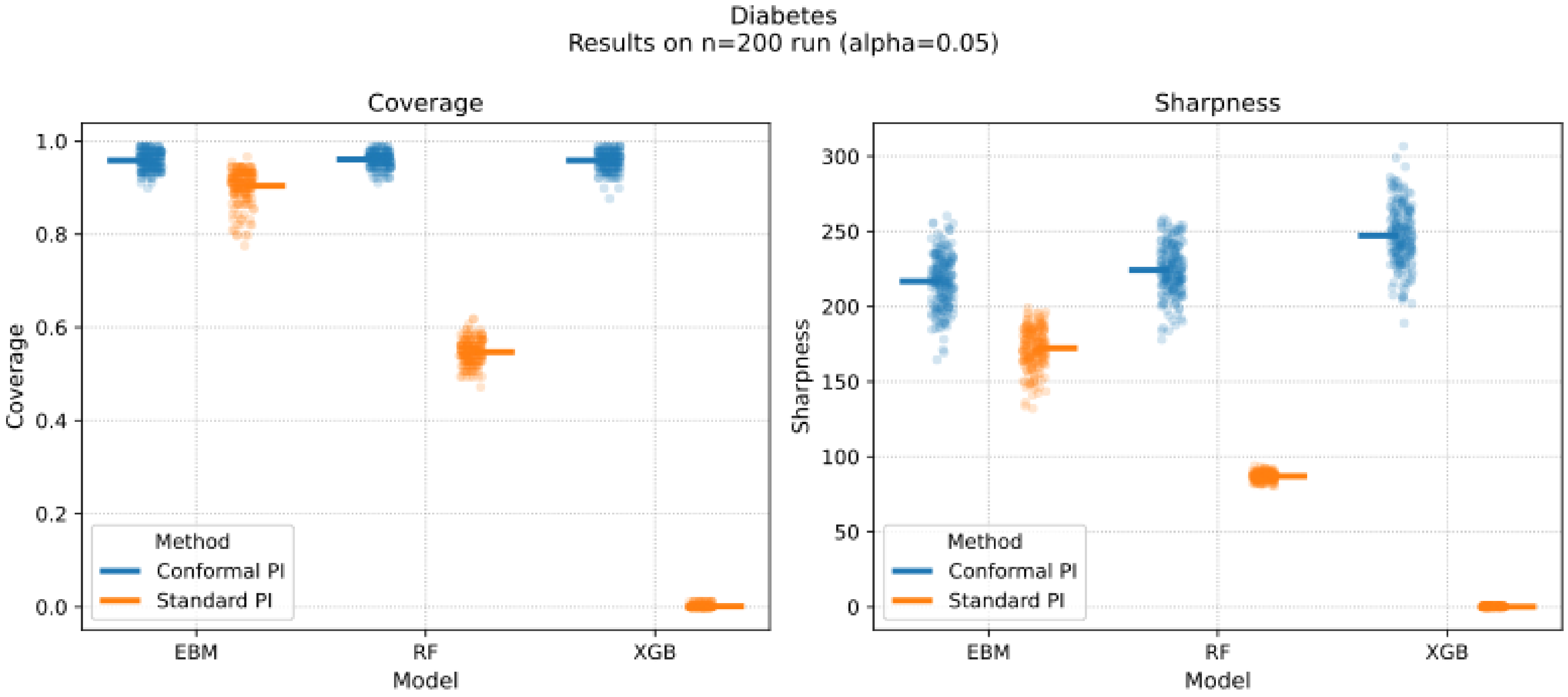
Methods: Simulation Study – RMSE on *Diabetes*

The two visualizations display the same RMSE results across 200 Monte Carlo runs.



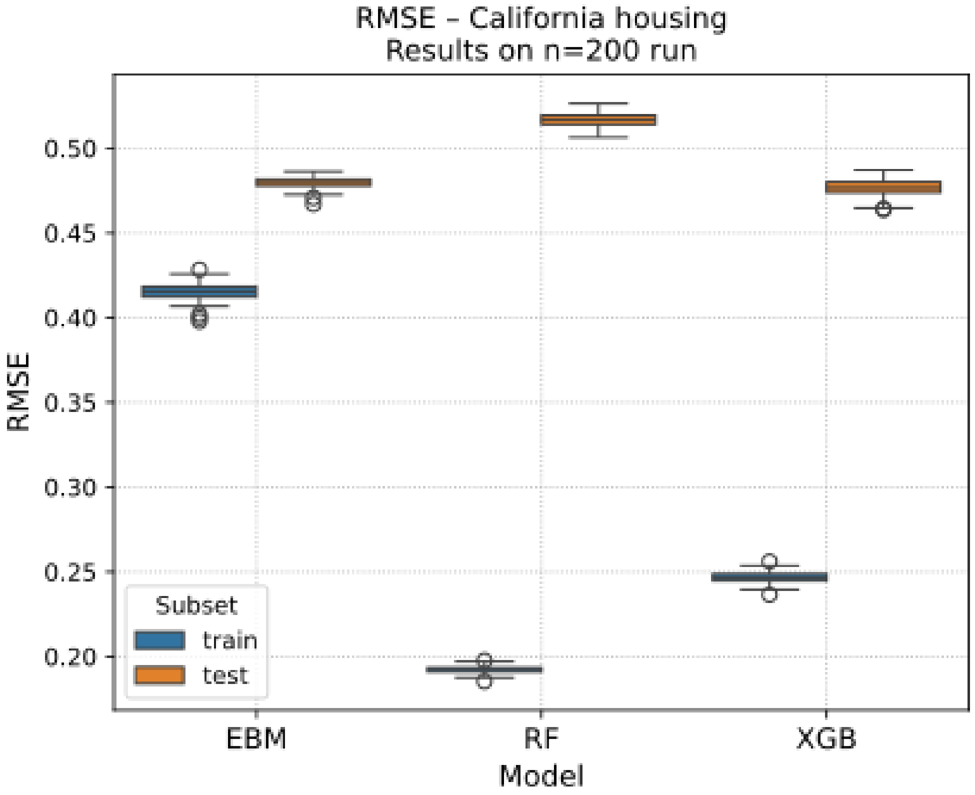


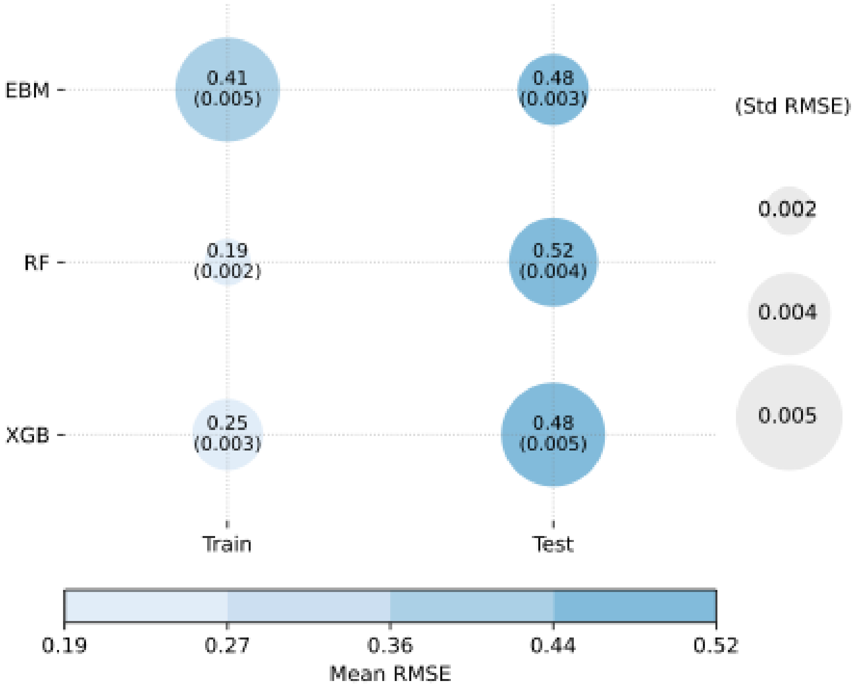
Methods: Simulation Study – *Coverage* and *Sharpness* of Conformal vs Standard PIs on *Diabetes*



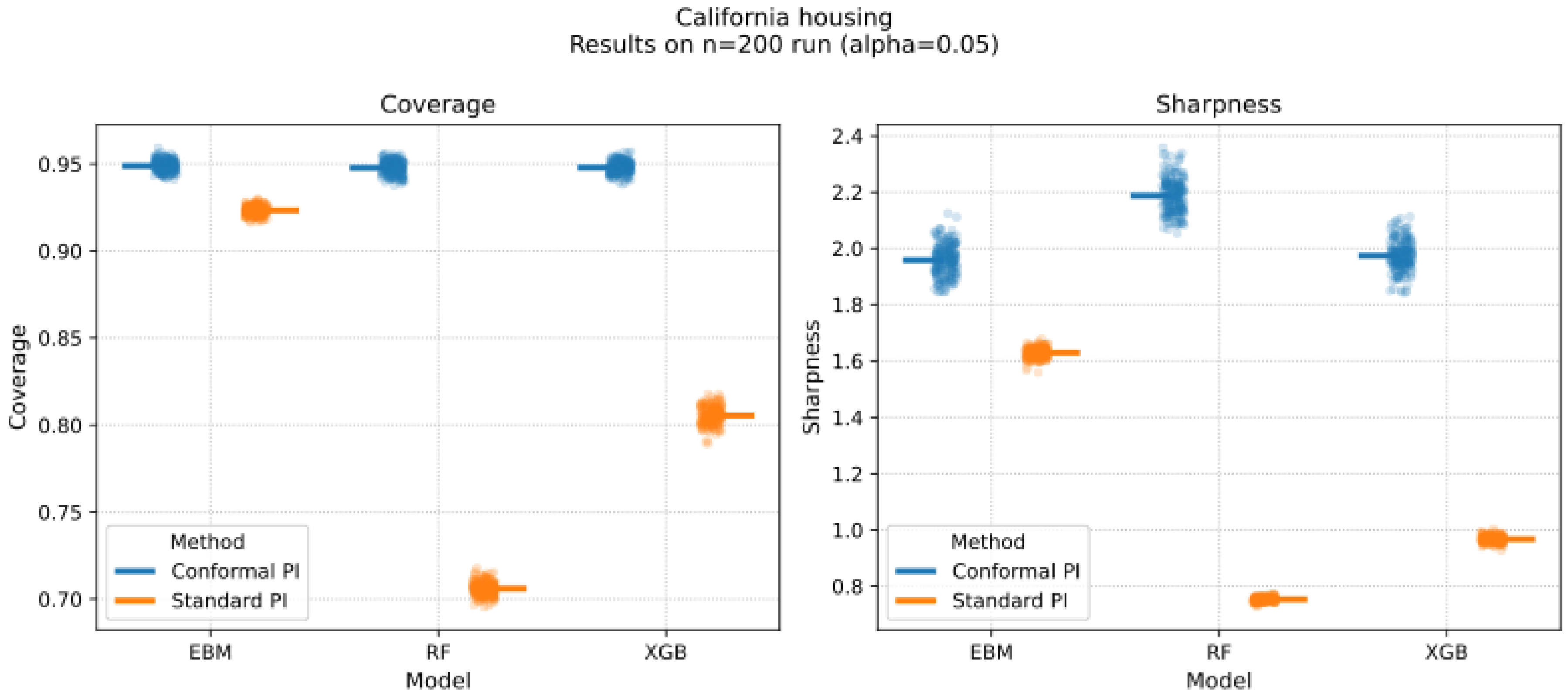
Methods: Simulation Study – RMSE on *California Housing*

The two visualizations display the same RMSE results across 200 Monte Carlo runs.

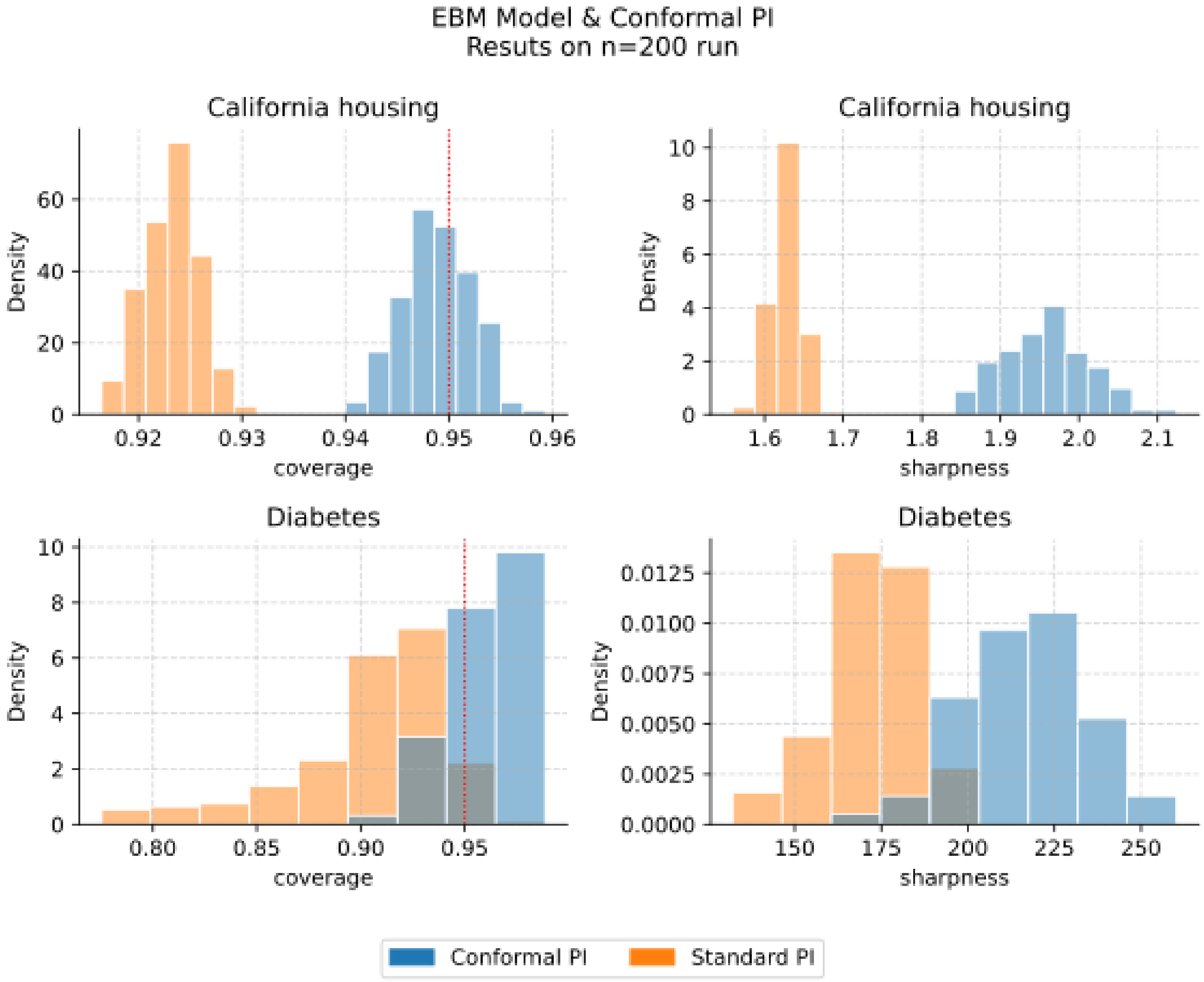




Methods: Simulation Study – *Coverage* and *Sharpness* of Conformal vs Standard PIs on *California Housing*

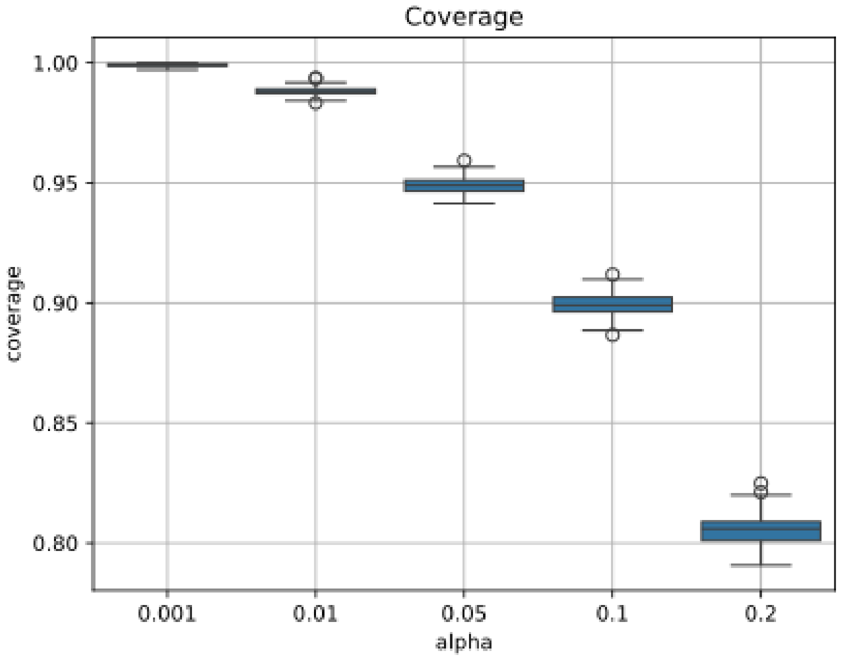


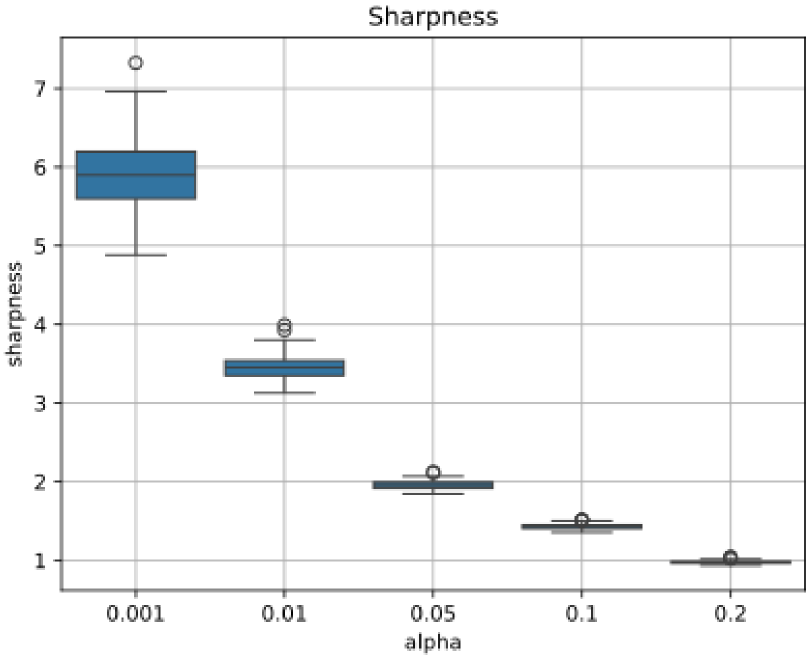
Methods: Simulation Study – Conformal PIs Statistics (on EBM)



Coverage Analysis: Simulation Study

Coverage and *sharpness* evaluated over 200 Monte Carlo simulations across five α levels (Section 4.2).


Coverage
coverage
alpha


Sharpness
sharpness
alpha