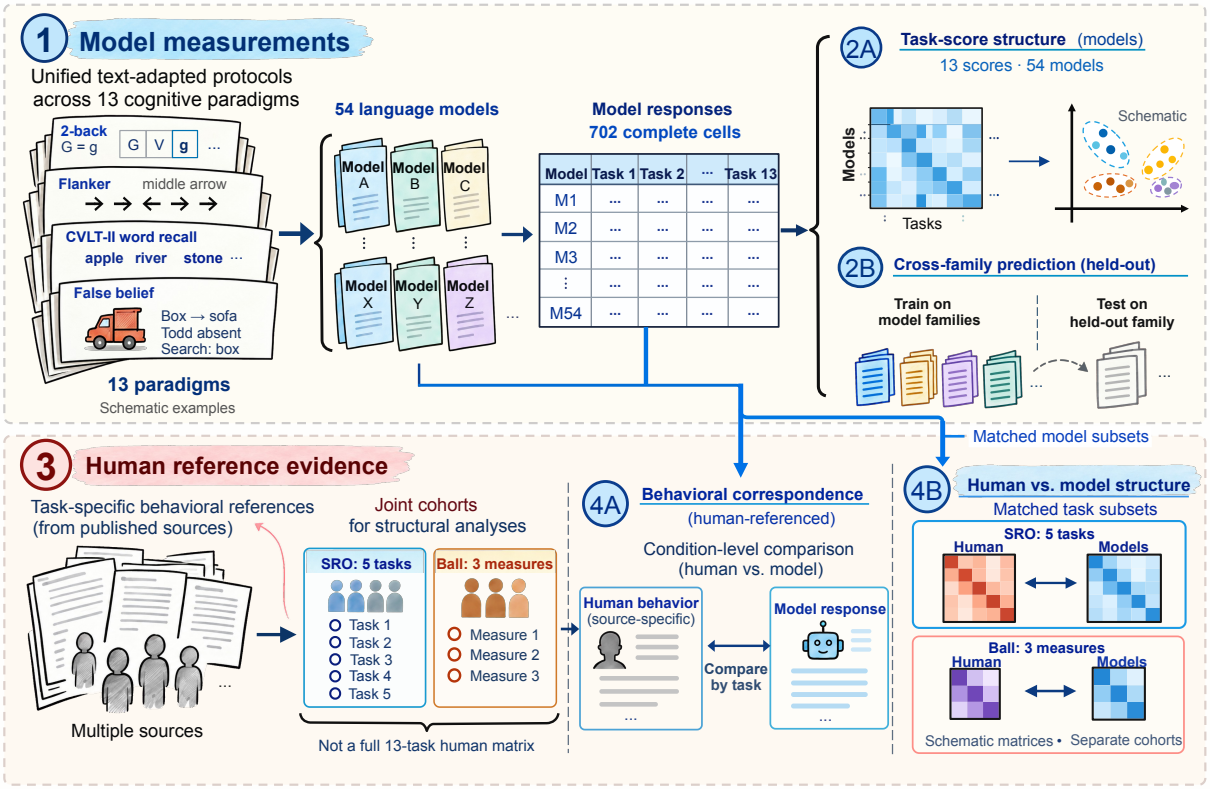

1 Model measurements
Unified text-adapted protocols across 13 cognitive paradigms
2-back
G = g
G V g ...
Flanker
middle arrow
CVLT-II word recall
apple river stone ···
False belief
Box → sofa
Todd absent
Search: box
13 paradigms
Schematic examples
54 language models
Model A
Model B
Model C
Model X
Model Y
Model Z
Model responses
702 complete cells
Model Task 1 Task 2 ··· Task 13
M1 ... ...
M2 ... ...
M3 ... ...
⋮ ... ...
M54 ... ...
2A Task-score structure (models)
13 scores · 54 models
Models
Tasks
Schematic
2B Cross-family prediction (held-out)
Train on model families
Test on held-out family
Matched model subsets
3 Human reference evidence
Task-specific behavioral references (from published sources)
Multiple sources
Joint cohorts for structural analyses
SRO: 5 tasks
Task 1
Task 2
Task 3
Task 4
Task 5
Ball: 3 measures
Measure 1
Measure 2
Measure 3
Not a full 13-task human matrix
4A Behavioral correspondence
(human-referenced)
Condition-level comparison (human vs. model)
Human behavior
(source-specific)
Compare by task
Model response
4B Human vs. model structure
Matched task subsets
SRO: 5 tasks
Human
Models
Ball: 3 measures
Human
Models
Schematic matrices • Separate cohorts