Do AI Scientists Do Science?

Martiño Ríos García - 13.05.2026

AI Scientists

AI Scientists

nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › articles › article

Article | Open access | Published: 20 December 2023

Autonomous chemical research with large language models

Daniil A. Boiko, Robert MacKnight, Ben Kline & Gabe Gomes

Nature **624**, 570–578 (2023) | Cite this article

nature machine intelligence

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature > nature machine intelligence > articles > article

Article | Open access | Published: 08 May 2024

Augmenting large language models with chemistry tools

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White & Philippe Schwaller

AI Scientists

nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature > articles > article

Article | Open access | Published: 20 December 2023

Autonomous chemical research with large language models

Daniil A. Boiko, Robert MacKnight, Ben Kline & Gabe Gomes

Nature **624**, 570–578 (2023) | Cite this article

nature machine intelligence

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › nature machine intelligence › articles › article

Article | Open access | Published: 08 May 2024

Augmenting large language models with chemistry tools

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White & Philippe Schwaller

AI Scientists

nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › articles › article

Article | Open access | Published: 25 March 2026

Towards end-to-end automation of AI research

Chris Lu, Cong Lu, Robert Tjarko Lange, Yutaro Yamada, Shengran Hu, Jakob Foerster, David Ha & Jeff Clune

Nature **651**, 914–919 (2026) | Cite this article

259k Accesses | **21** Citations | **752** Altmetric | Metrics

nature

content ⌄ About the journal ⌄ Publish with us ⌄

› articles › article

Open access | Published: 20 December 2023

ɔnomous chemical research with large language lels

Boiko, Robert MacKnight, Ben Kline & Gabe Gomes

624, 570–578 (2023) | Cite this article

ccesses | **927** Citations | **974** Altmetric | Metrics

AI Scientists

nature machine intelligence

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › nature machine intelligence › articles › article

Article | Open access | Published: 08 May 2024

Augmenting large language models with tools

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White

nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › articles › article

Article | Published: 19 May 2026

Accelerating scientific discovery with Co-Scientist

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Petar Sirkovic, Artiom Myaskovsky, Grzegorz Glowaty, Felix Weissenberger, Alessio Orlandi, Dan Popovici, Anil Palepu, Keran Rong, Ryutaro Tanno, Khaled Saab, Fan Zhang, Jacob Blum, Andrew Carroll, Kavita Kulkarni, Nenad Tomašev, Dina Zverinski, Ivor Rendulic, Elahe Vedadi, Florian Hasler, Luka Rimanic, ... Vivek Natarajan

+ Show authors

Nature (2026) | Cite this article

nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › articles › article

Article | Open access | Published: 25 March 2026

Towards end-to-end automation of AI research

Chris Lu, Cong Lu, Robert Tjarko Lange, Yutaro Yamada, Shengran Hu, Jakob Foerster, David Ha & Jeff Clune

Nature **651**, 914–919 (2026) | Cite this article

259k Accesses | **21** Citations | **752** Altmetric | Metrics

content ⌄ About the journal ⌄ Publish with us ⌄

› articles › article

Open access | Published: 20 December 2023

›nomous chemical research with large language ‹lels

Boiko, Robert MacKnight, Ben Kline & Gabe Gomes

624, 570–578 (2023) | Cite this article

252k Accesses | **927** Citations | **974** Altmetric | Metrics

nature machine intelligence

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › nature machine intelligence › articles › article

Article | Open access |

Augmenting
tools

Andres M. Bran, Sam C

nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › articles › article

Article | Published: 19 May 2026

with Co-Scientist

Tu, Petar Sirkovic, Artiom Myaskovsky,
Popovici, Anil Palepu, Keran Rong, Ryutaro
Kavita Kulkarni, Nenad Tomašev, Dina
manic, ... Vivek Natarajan ✉

nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › articles › article

Article | Published: 19 May 2026

A multi-agent system for automating scientific discovery

Ali Essam Ghareeb, Benjamin Chang, Ludovico Mitchener, Angela Yiu, Caralyn J. Szostkiewicz, Dmytro Shved, Gavin J. Gyimesi, Jon M. Laurent, Samantha M. Wright, Muhammed T. Razzak, Andrew D. White ✉, Silvia C. Finnemann, Michaela M. Hinks ✉ & Samuel G. Rodrigues ✉

Nature (2026) | Cite this article

AI Sci

nature

Explore content ⌄ About the journ

nature › articles › article

Article | Open access | Published: 25

Towards end-to-end

Chris Lu, Cong Lu, Robert Tjarko Lange, Yutaro Yamada ✉, Shengran Hu, Jakob Foerster, David Ha ✉ & Jeff Clune ✉

Nature **651**, 914–919 (2026) | Cite this article

259k Accesses | **21** Citations | **752** Altmetric | Metrics

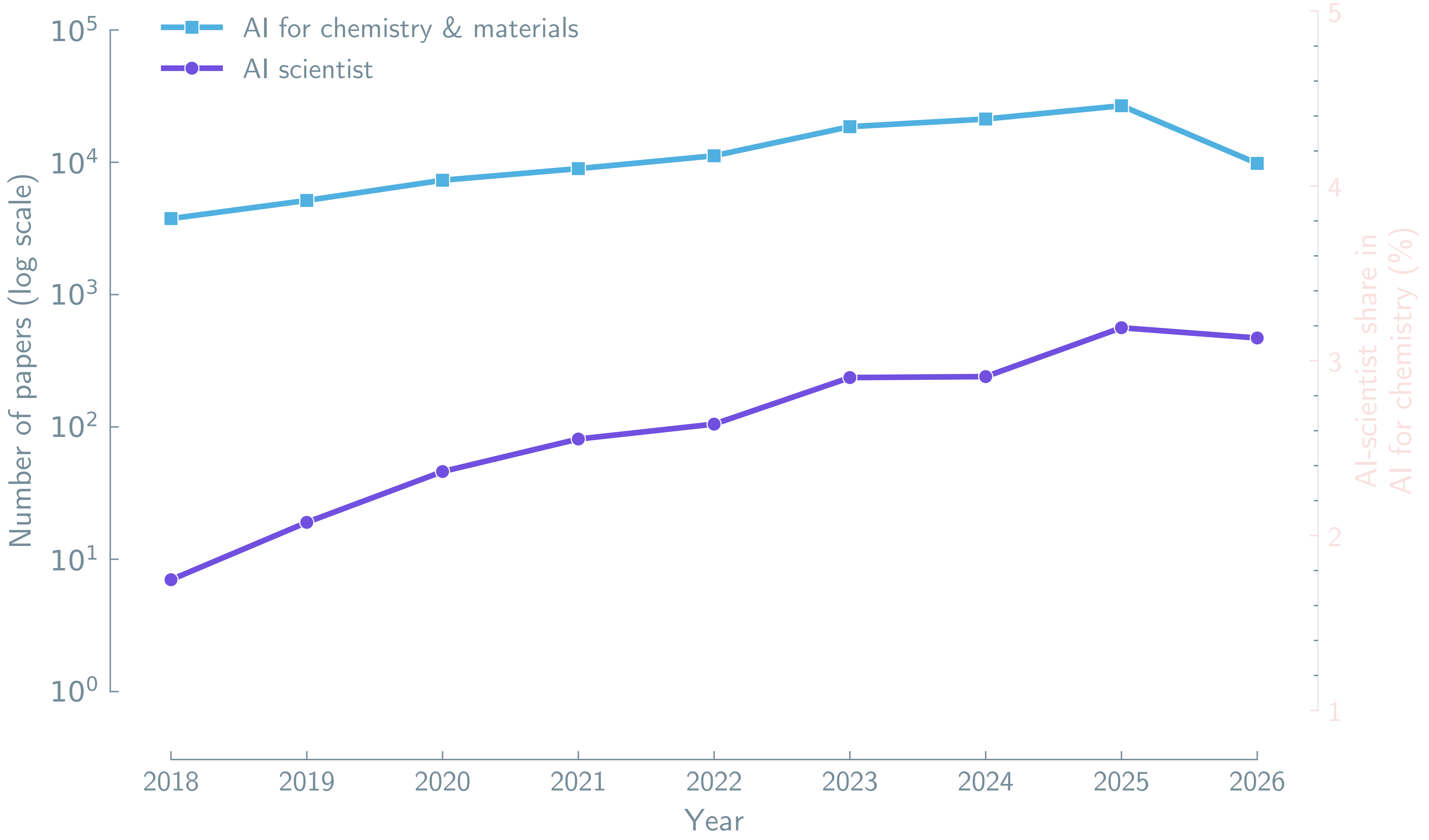
with large language
els

Boiko, Robert MacKnight, Ben Kline & Gabe Gomes ✉

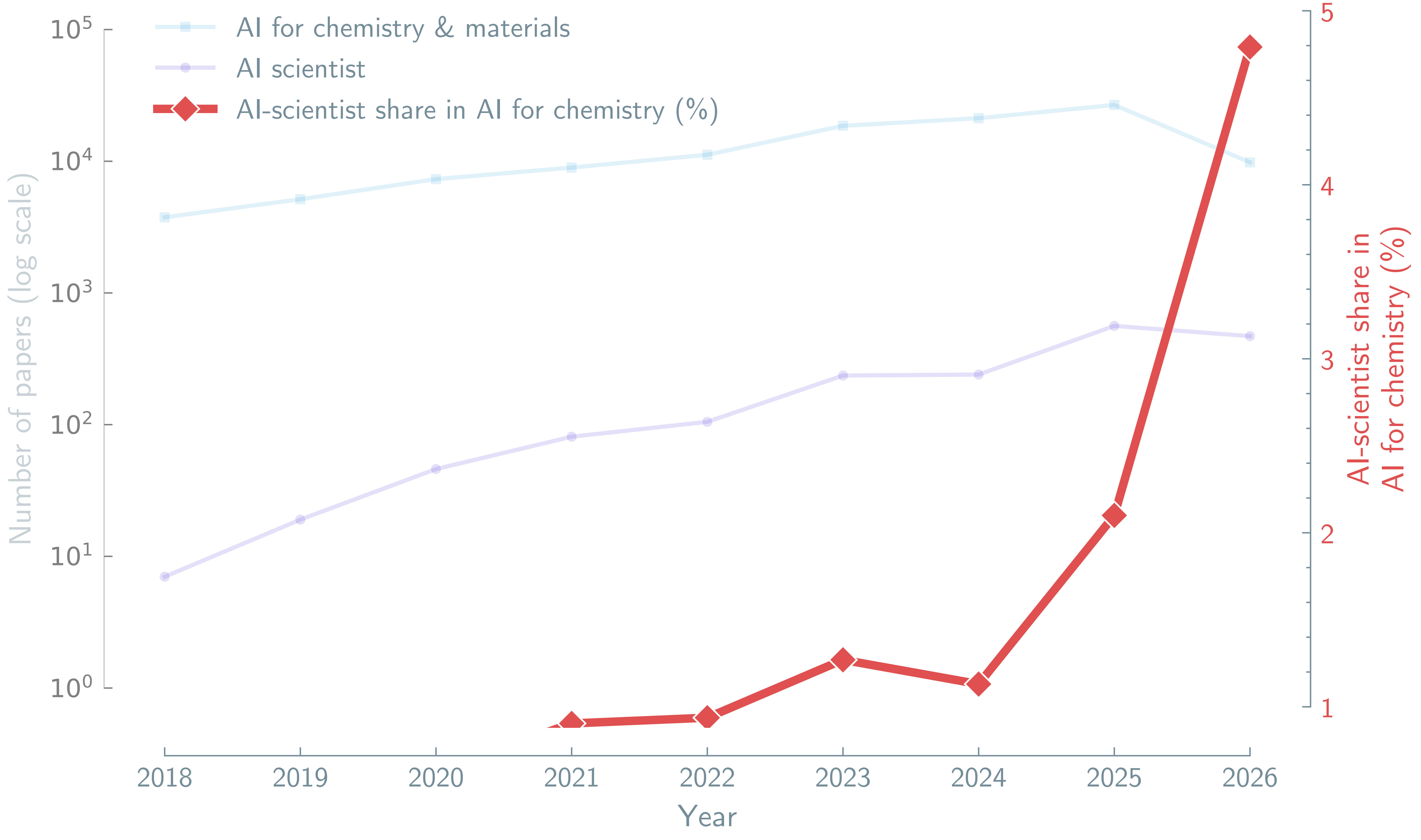
624, 570–578 (2023) | Cite this article

Accesses | **927** Citations | **974** Altmetric | Metrics

Rise of AI Scientists



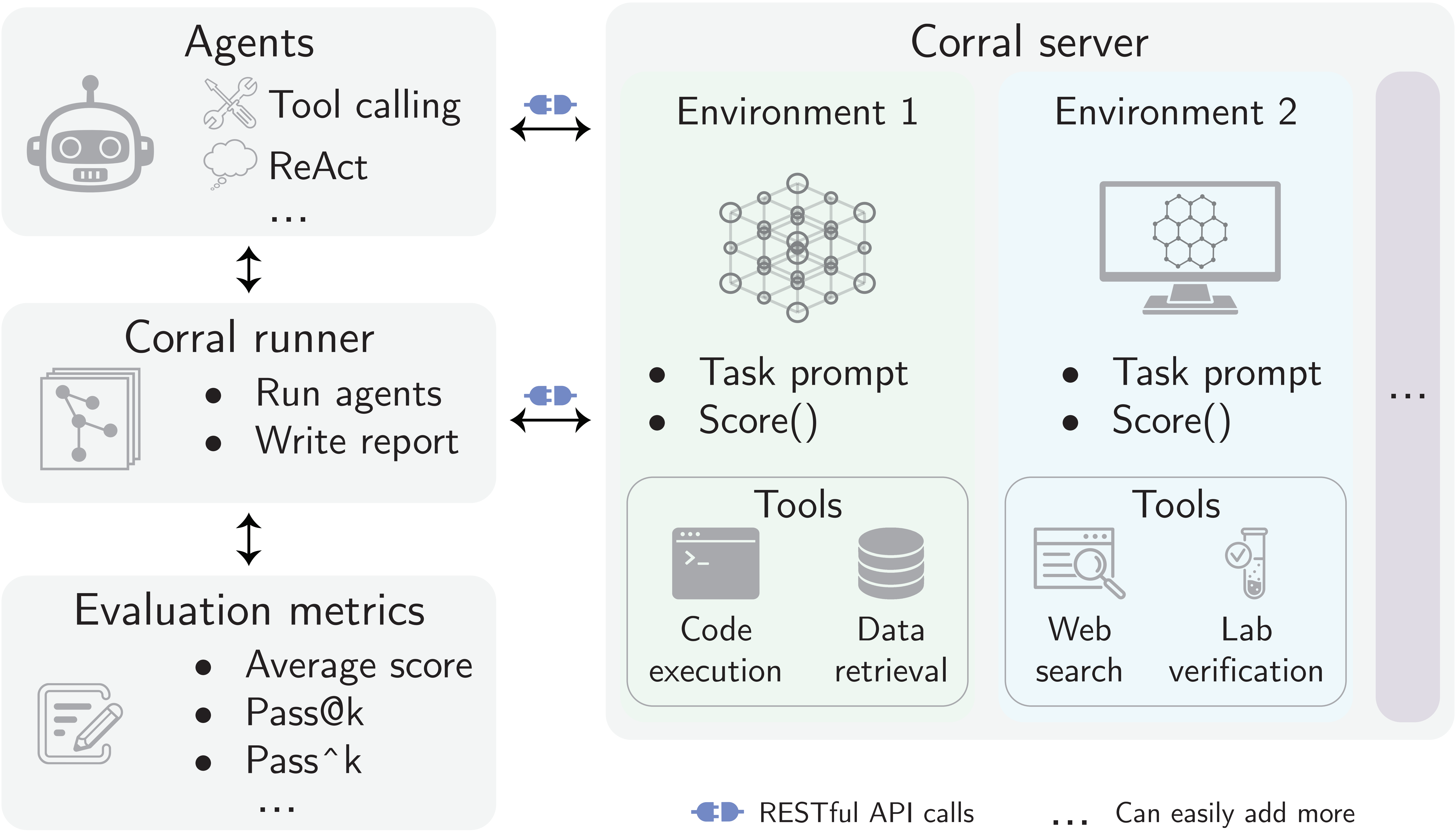
Rise of AI Scientists



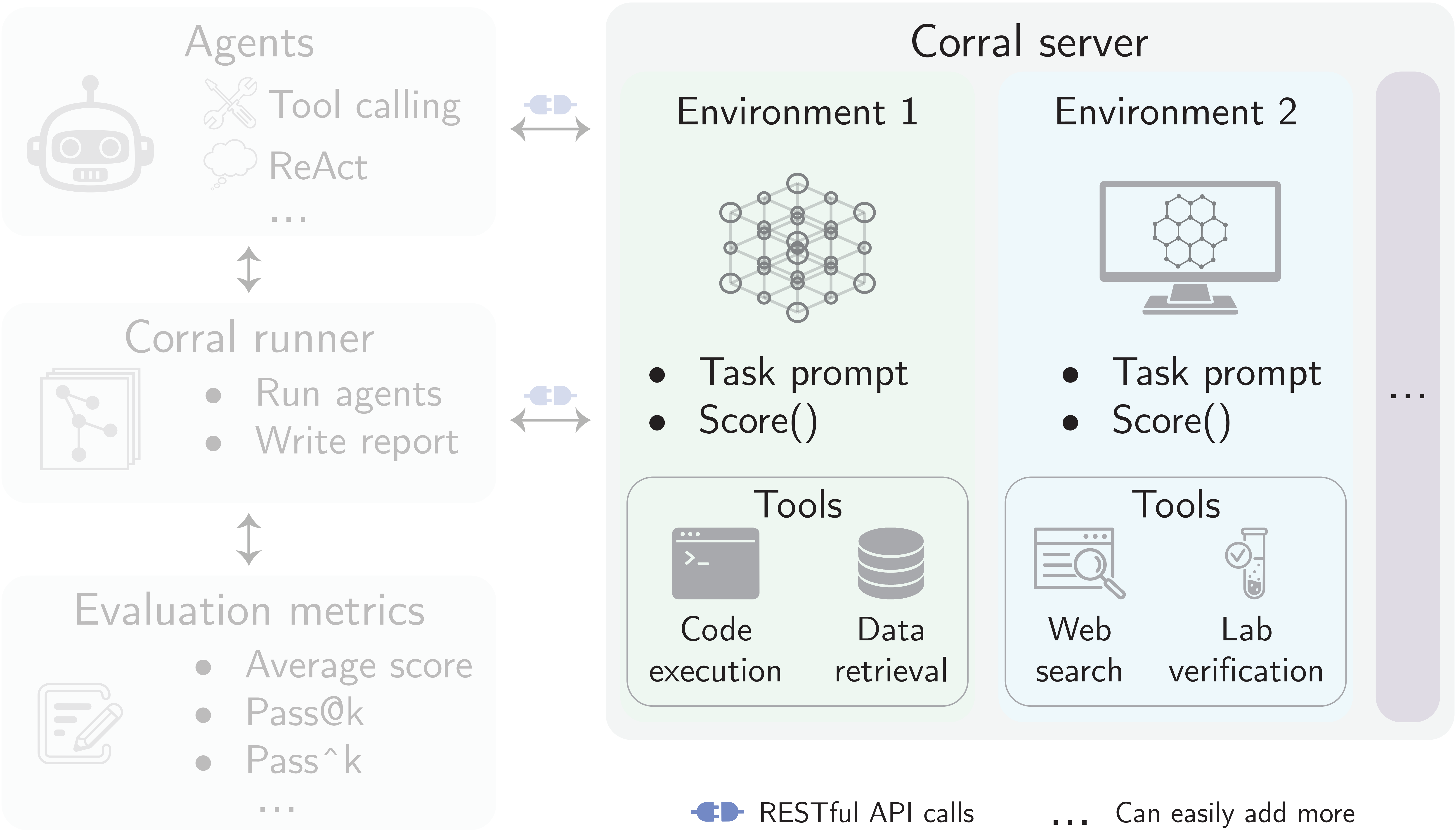
We are building scientific agents faster than we are learning how to evaluate them.

Corral

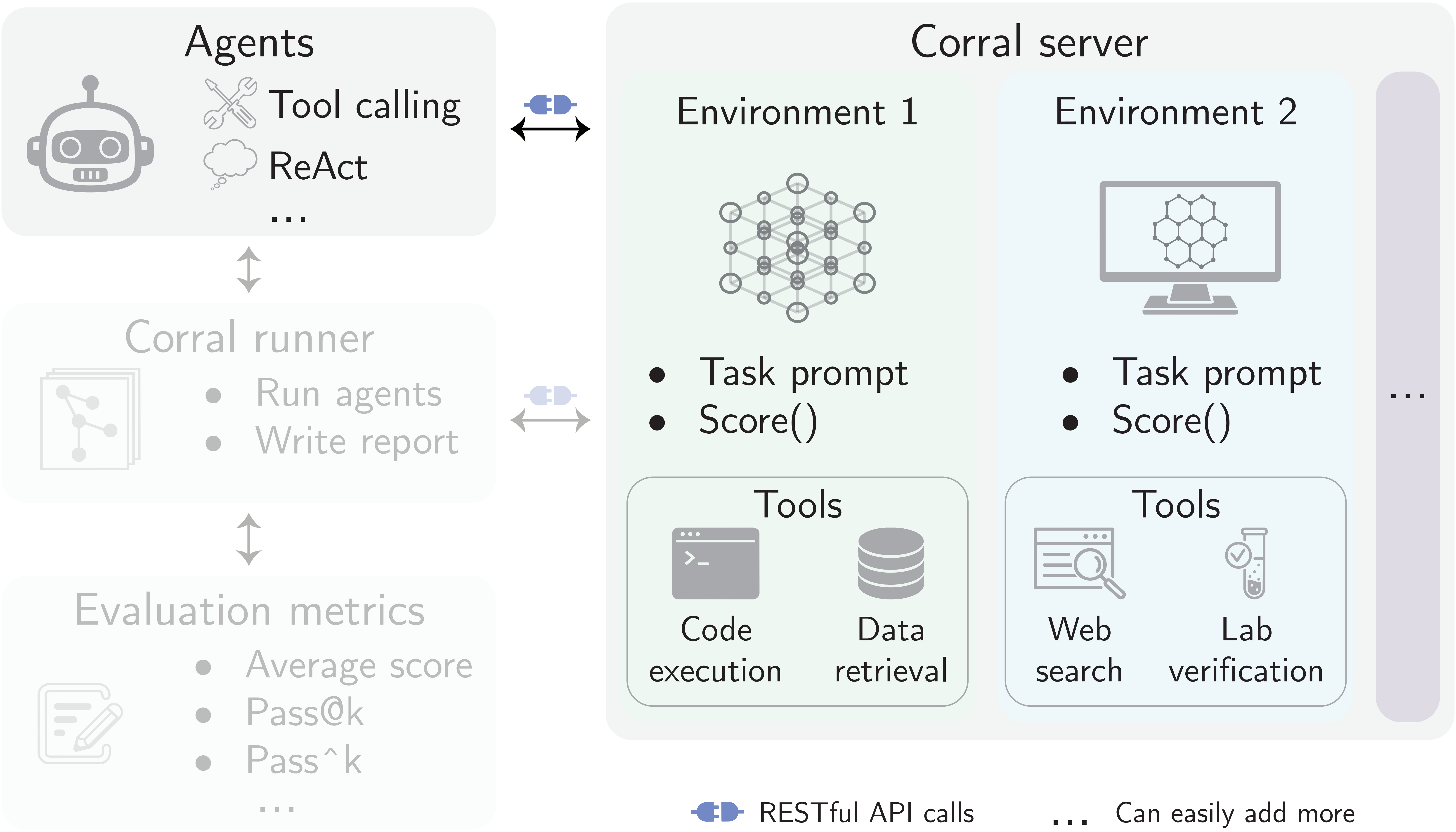
The Framework



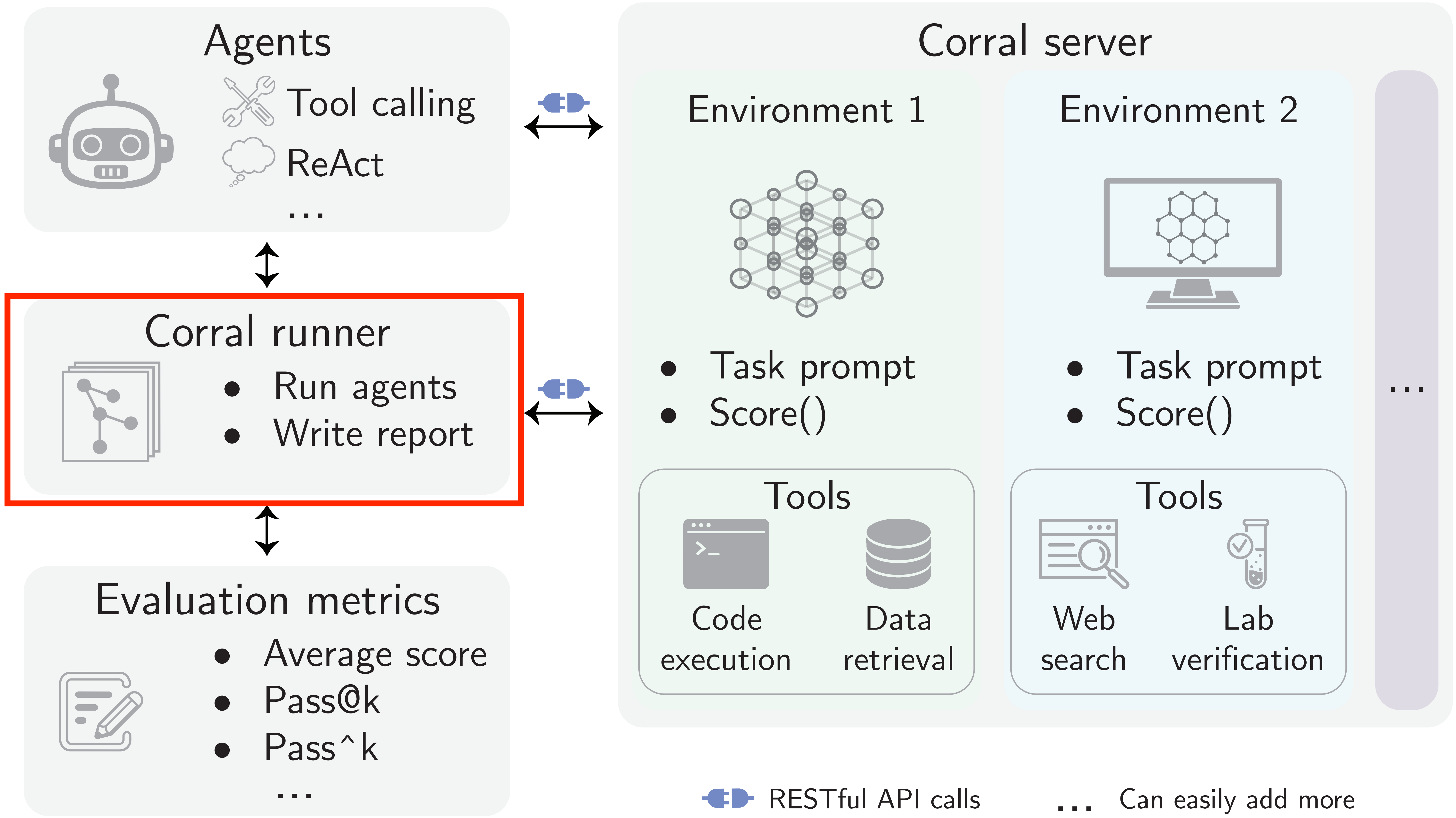
The Framework



The Framework



The Framework



The Environments

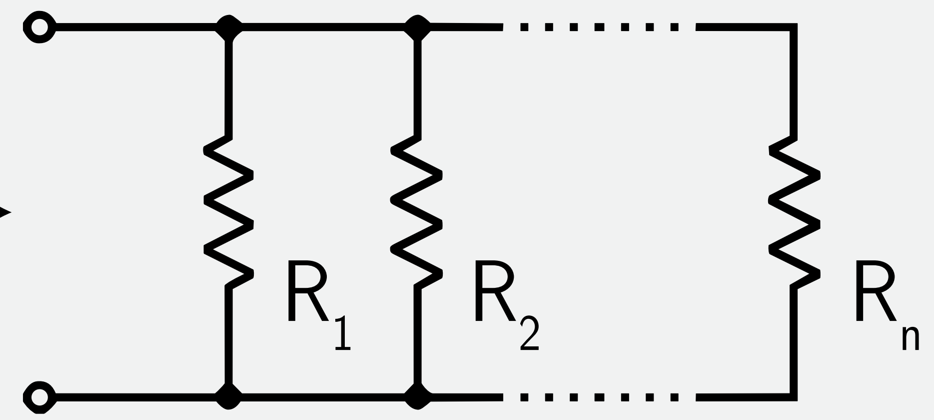
Hypothesis-driven Inquiry

Hypothesis-driven Inquiry

Circuit inference

Node 1	Node 2	Ω
A	X1	5
A	X3	6
...	...	...

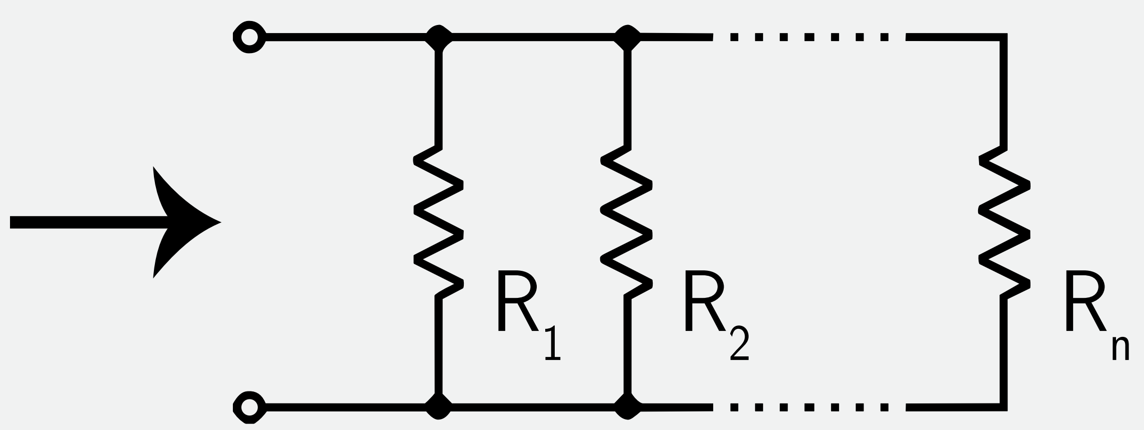
→



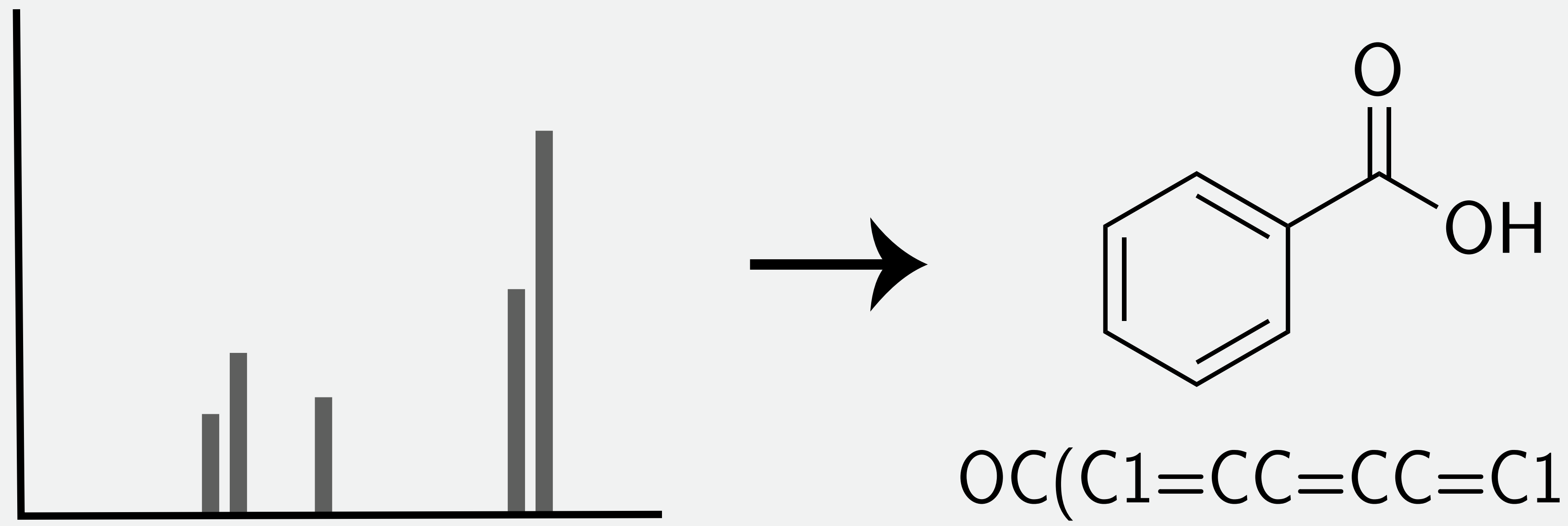
Hypothesis-driven Inquiry

Circuit inference

Node 1	Node 2	Ω
A	X1	5
A	X3	6
...	...	...



Spectroscopic structure elucidation

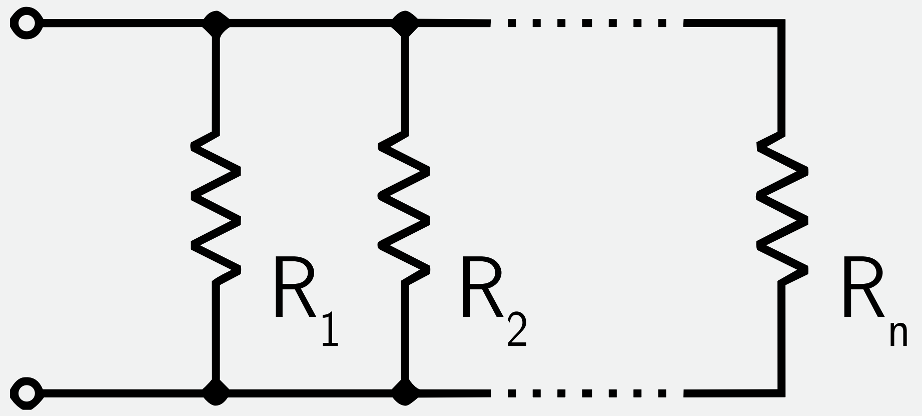


OC(C1=CC=CC=C1)=O

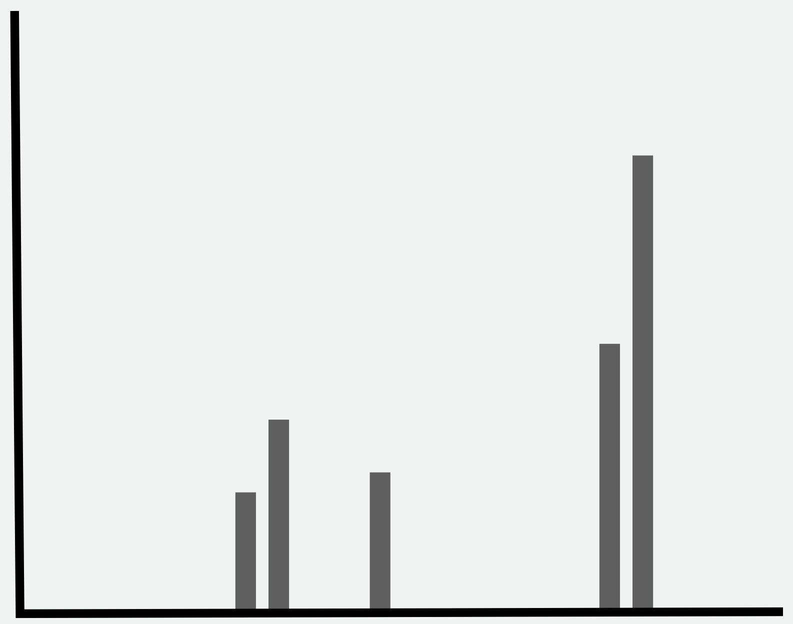
Hypothesis-driven Inquiry

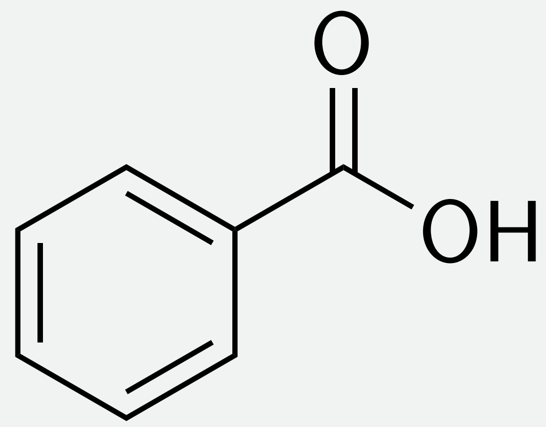
Circuit inference

Node 1	Node 2	Ω
A	X1	5
A	X3	6
...	...	...



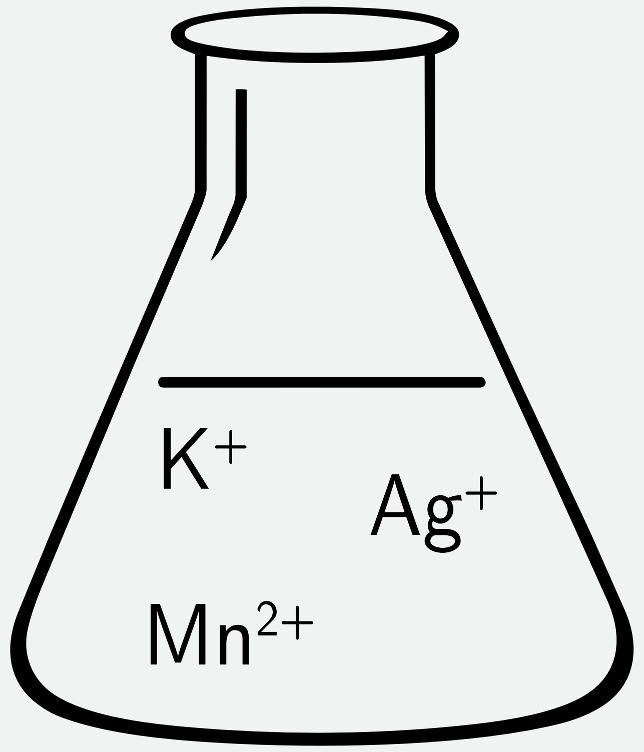
Spectroscopic structure elucidation





OC(C1=CC=CC=C1)=O

Inorganic qualitative analysis

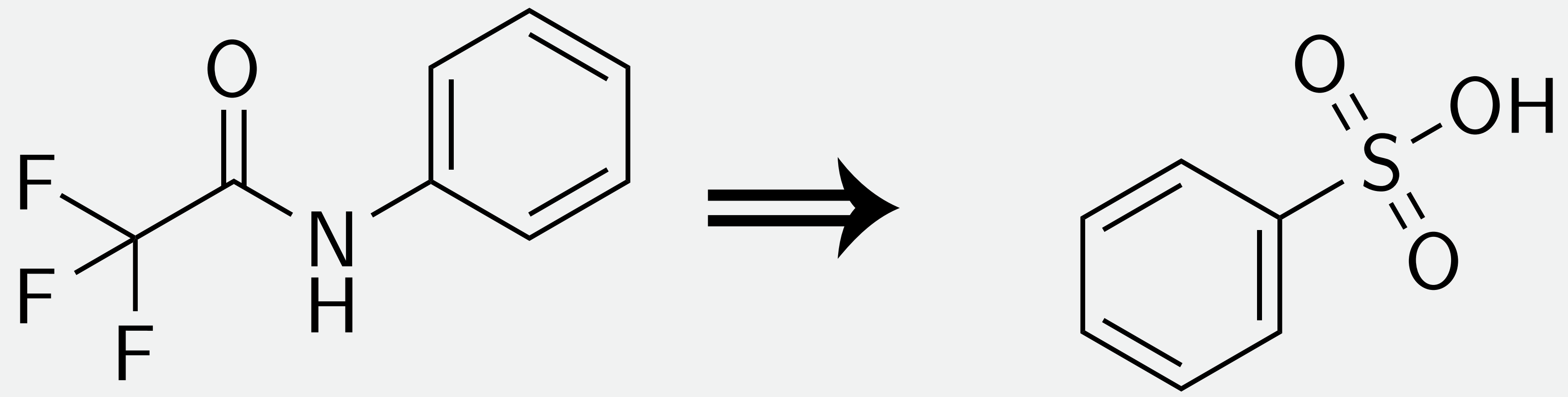


- K^+
- Ag^+
- Mn^{2+}

Strategic Reasoning

Strategic Reasoning

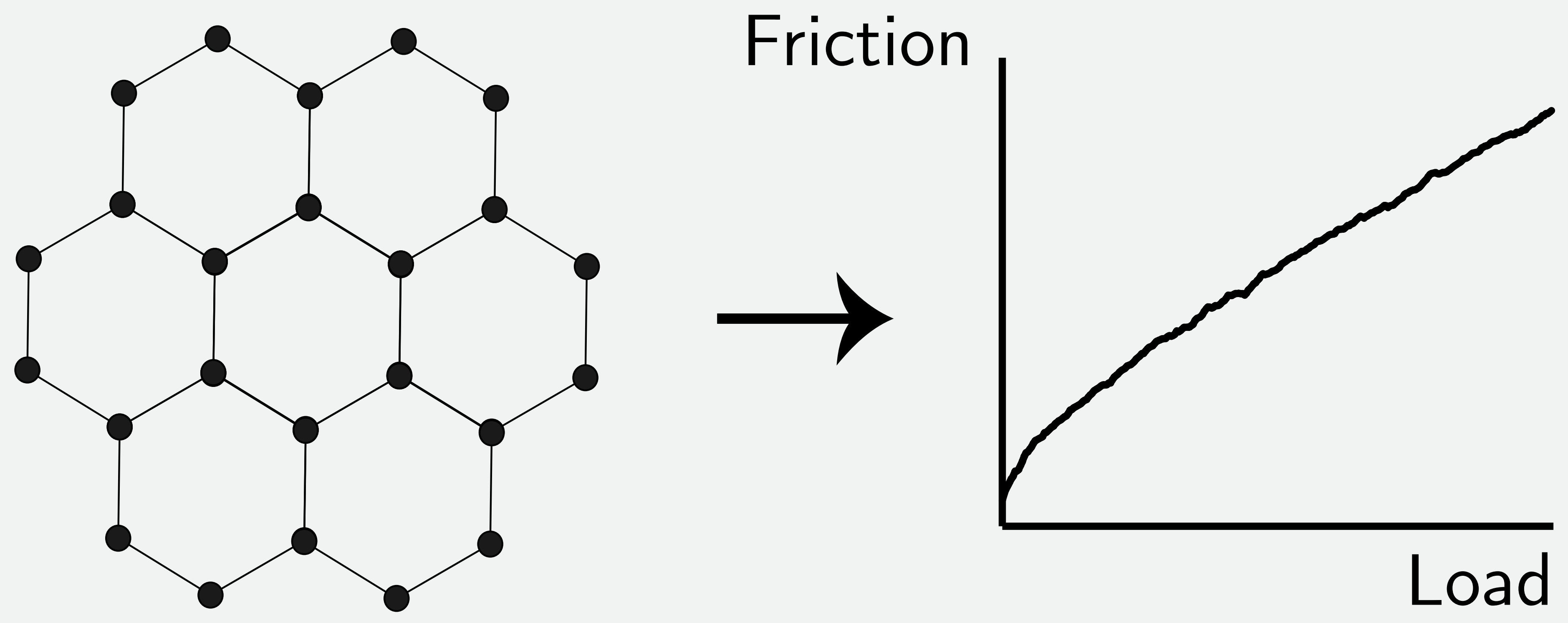
Retrosynthetic planning



Workflow Construction

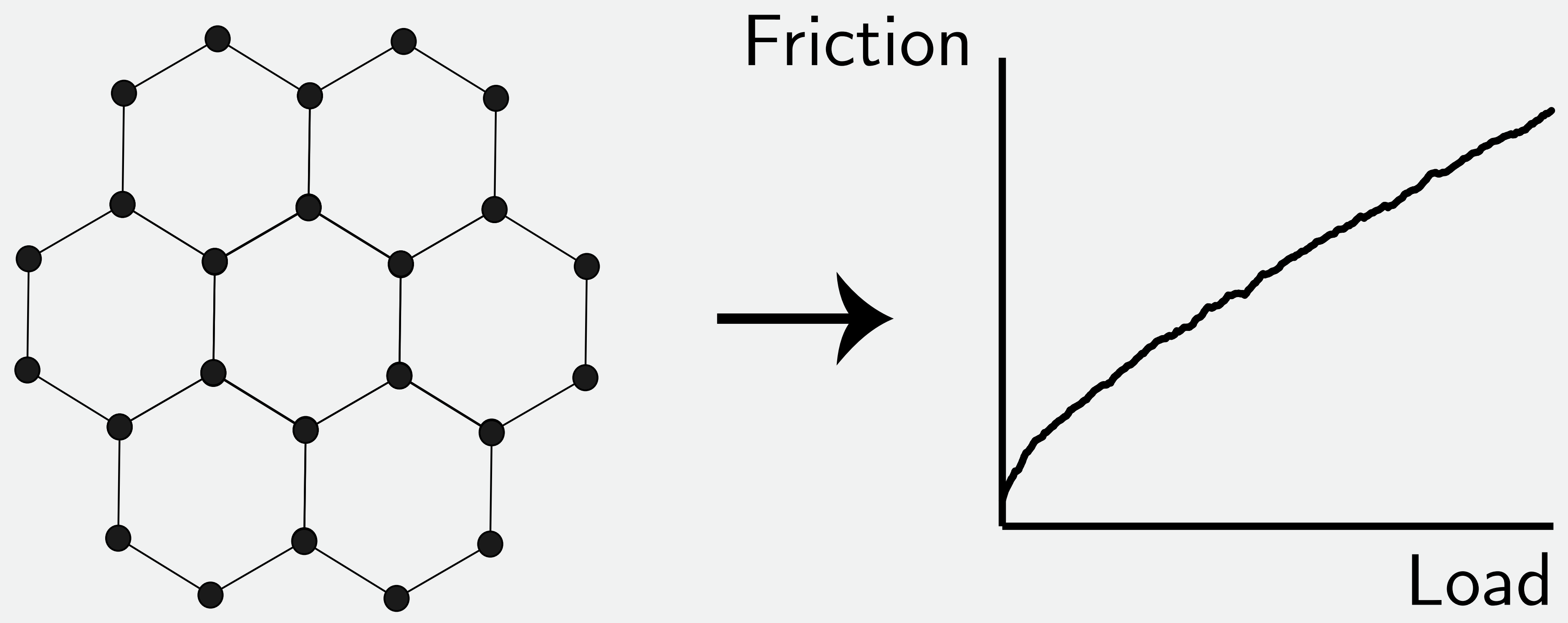
Workflow Construction

AFM experiment execution

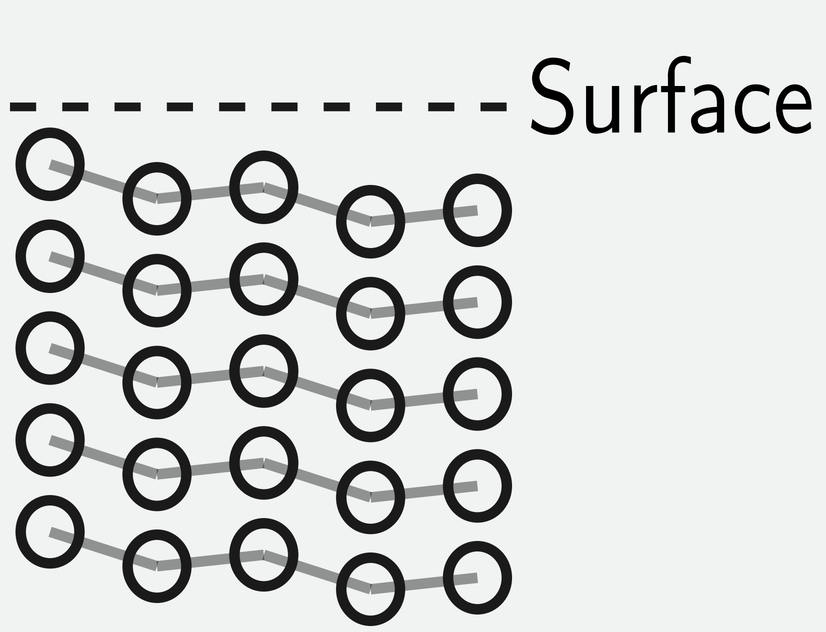


Workflow Construction

AFM experiment execution

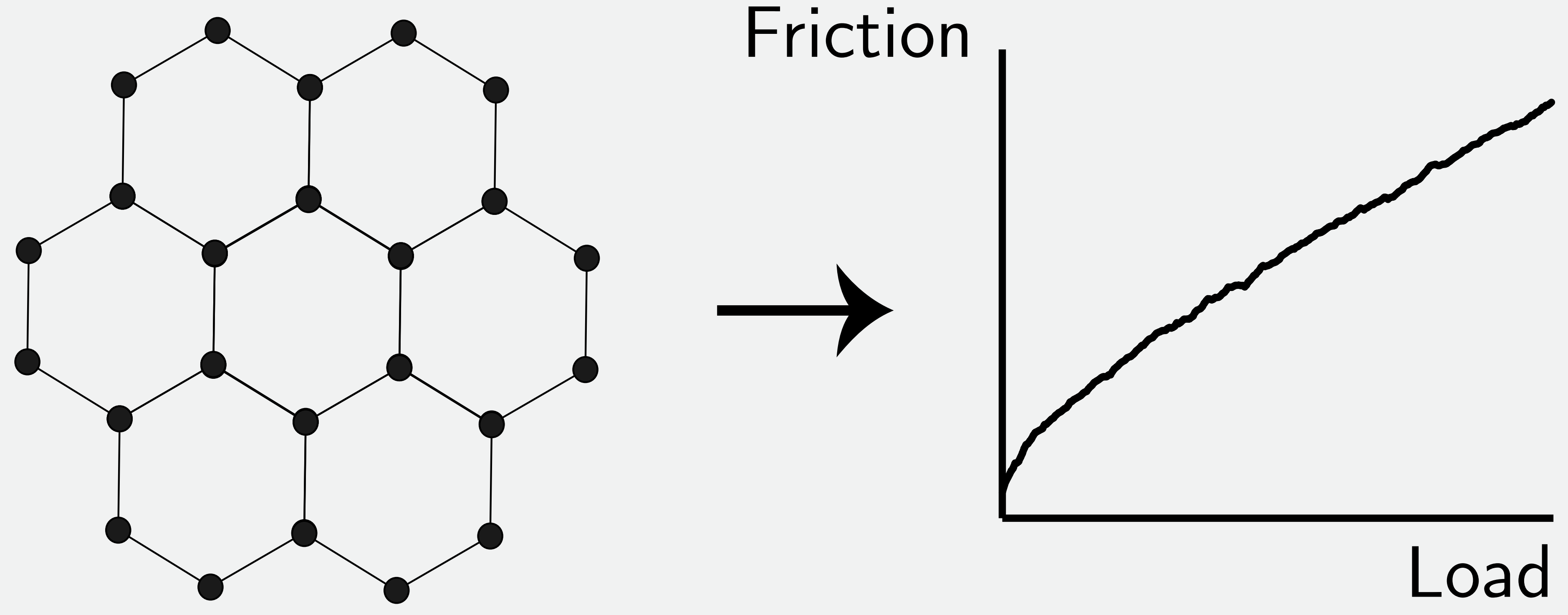


Adsorption surface construction

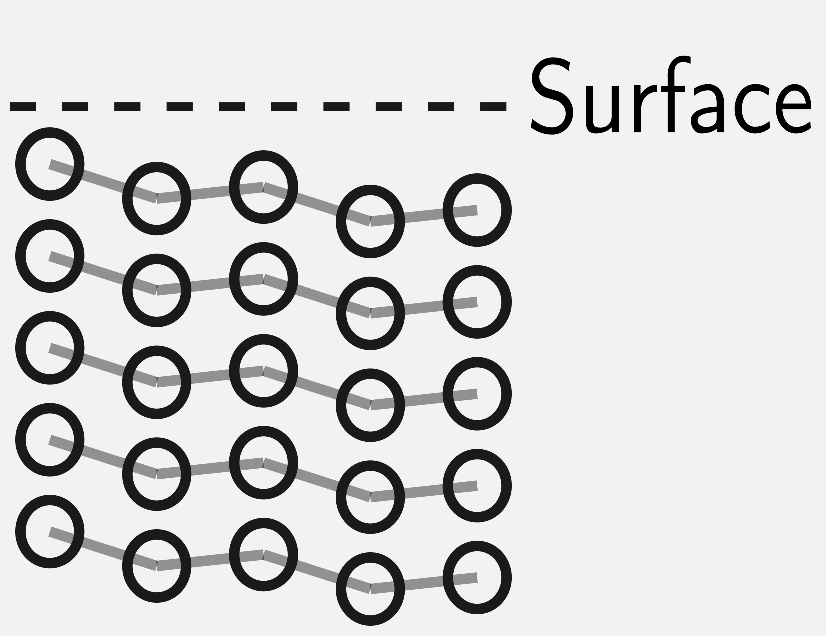


Workflow Construction

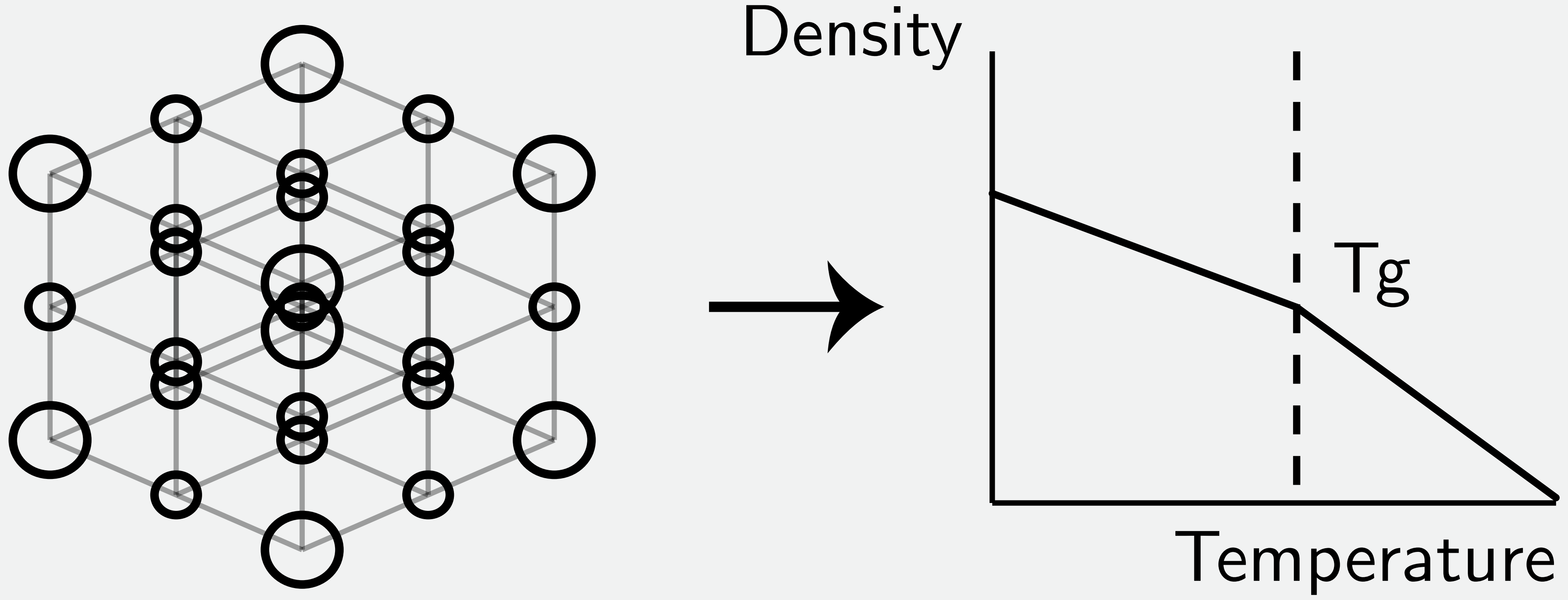
AFM experiment execution



Adsorption surface construction

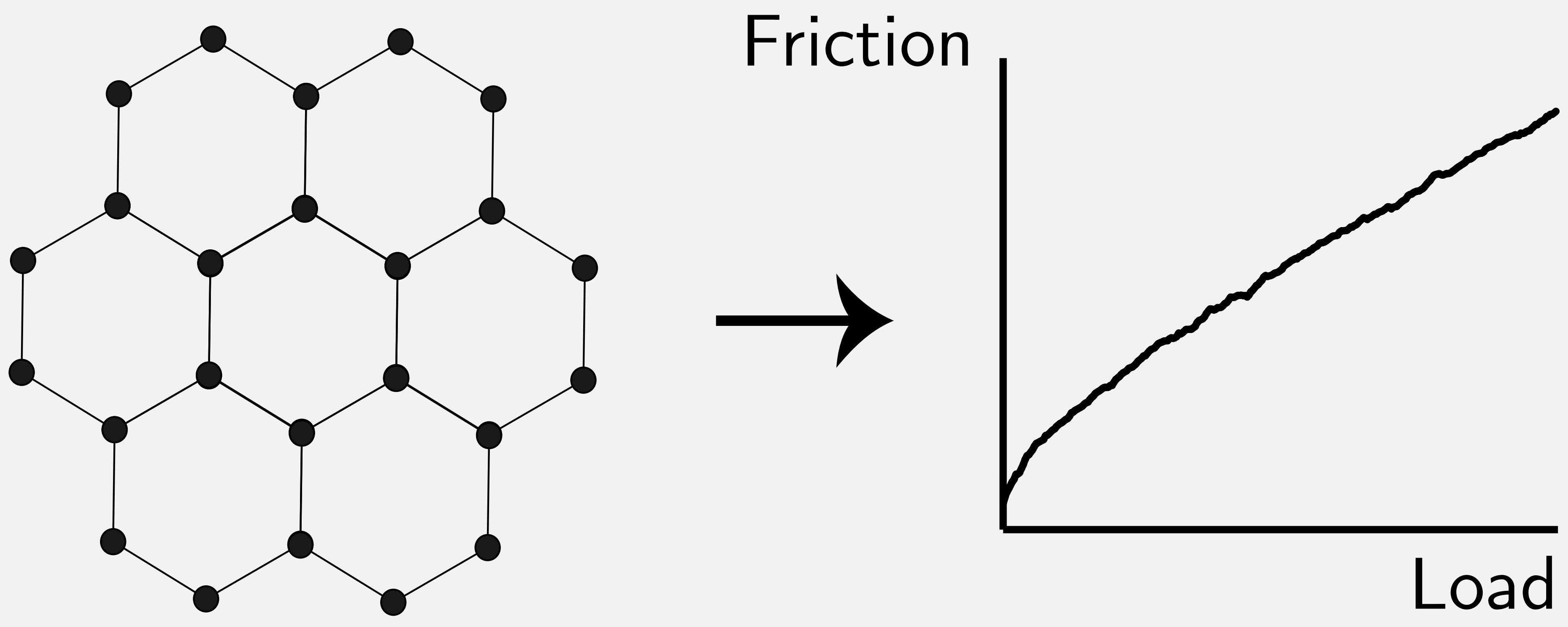


Molecular simulation

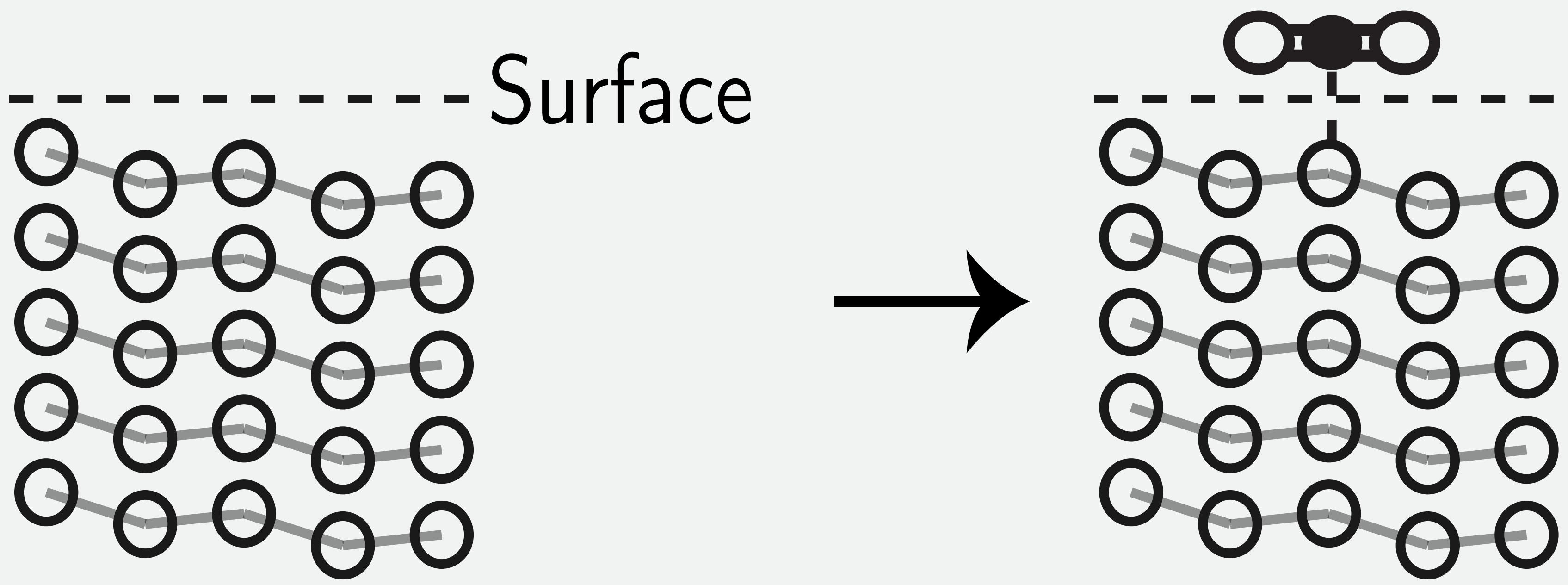


Workflow Construction

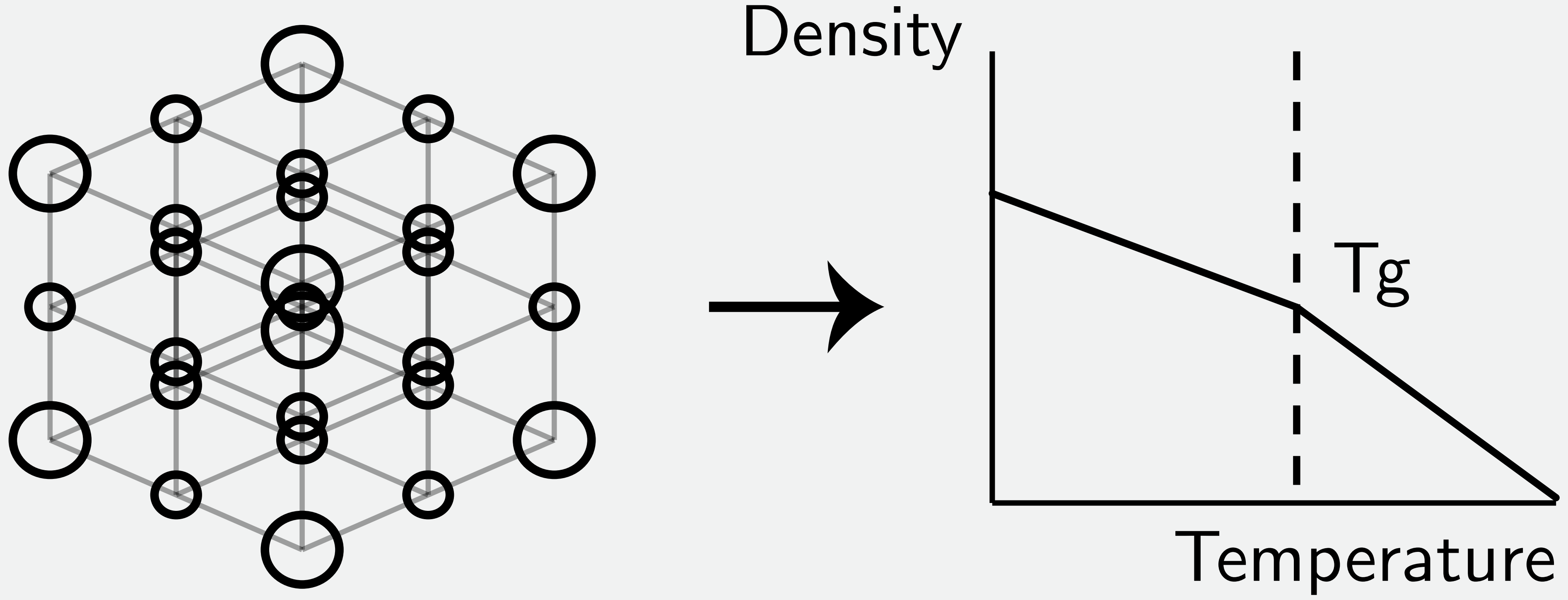
AFM experiment execution



Adsorption surface construction

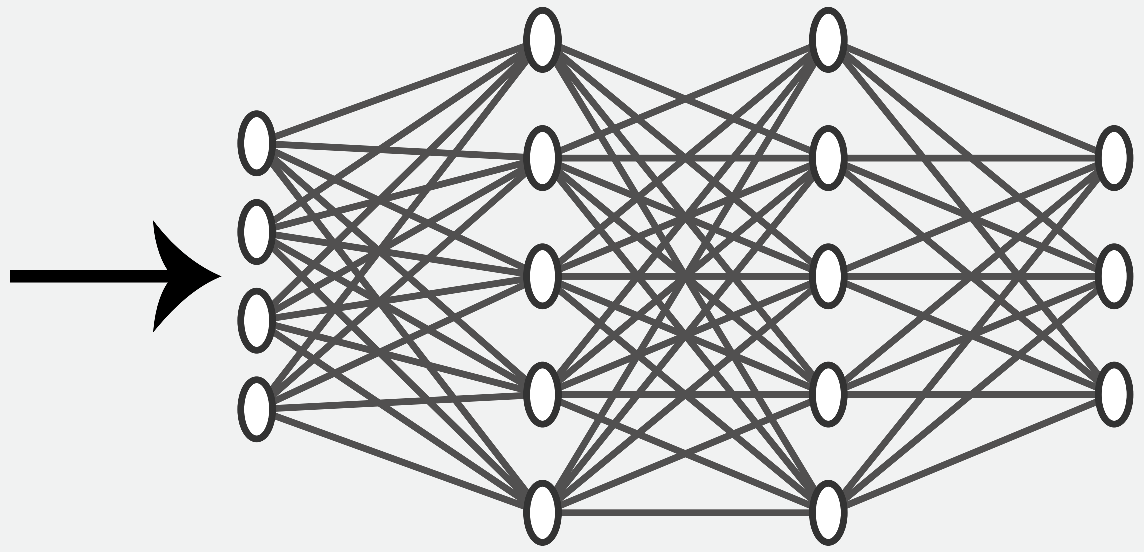


Molecular simulation



ML-based property prediction

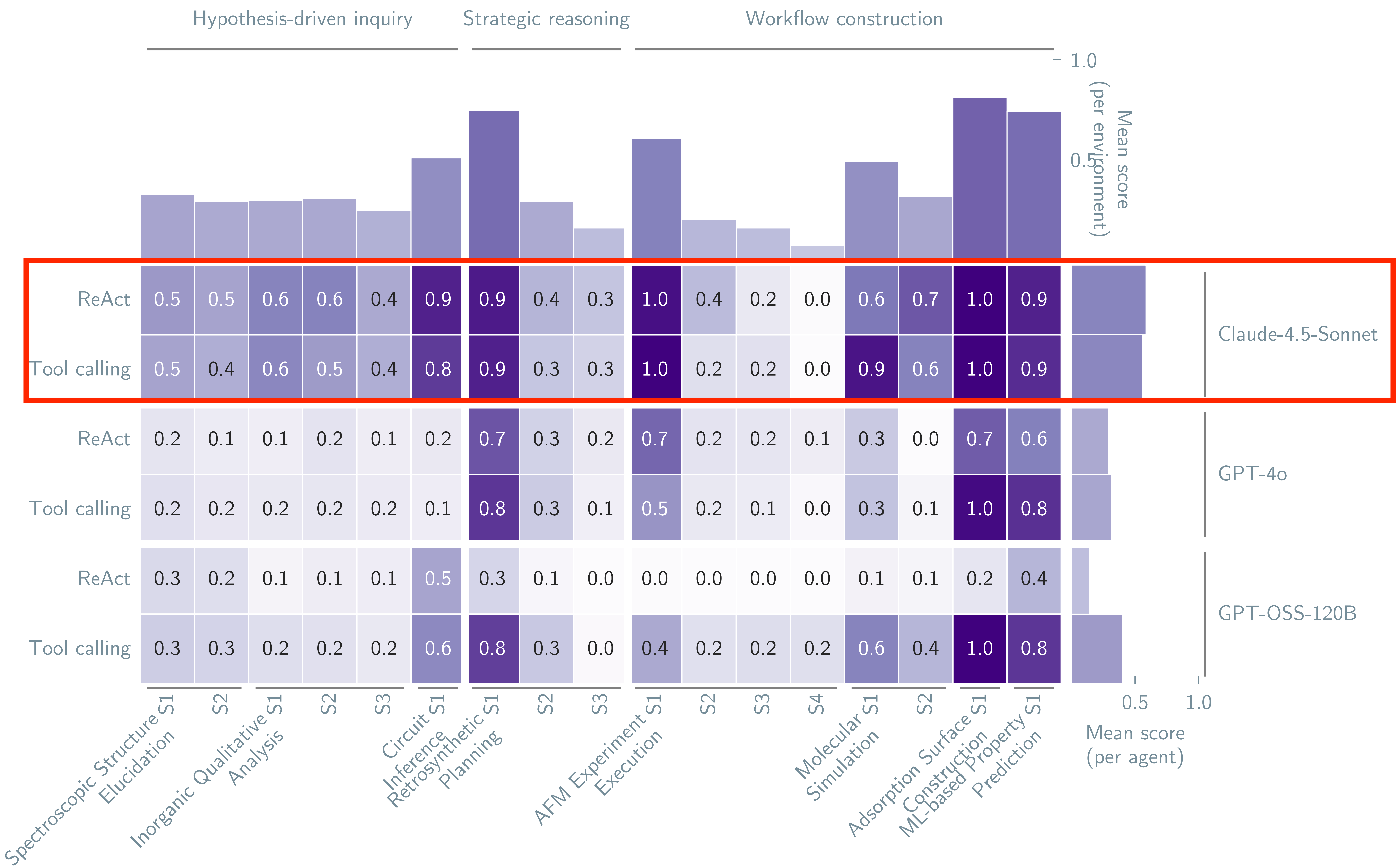
Polymorph	E^f
Hematite	-1.91
Rutile	-9.73
...	...



Base model determines performance

Base model determines performance

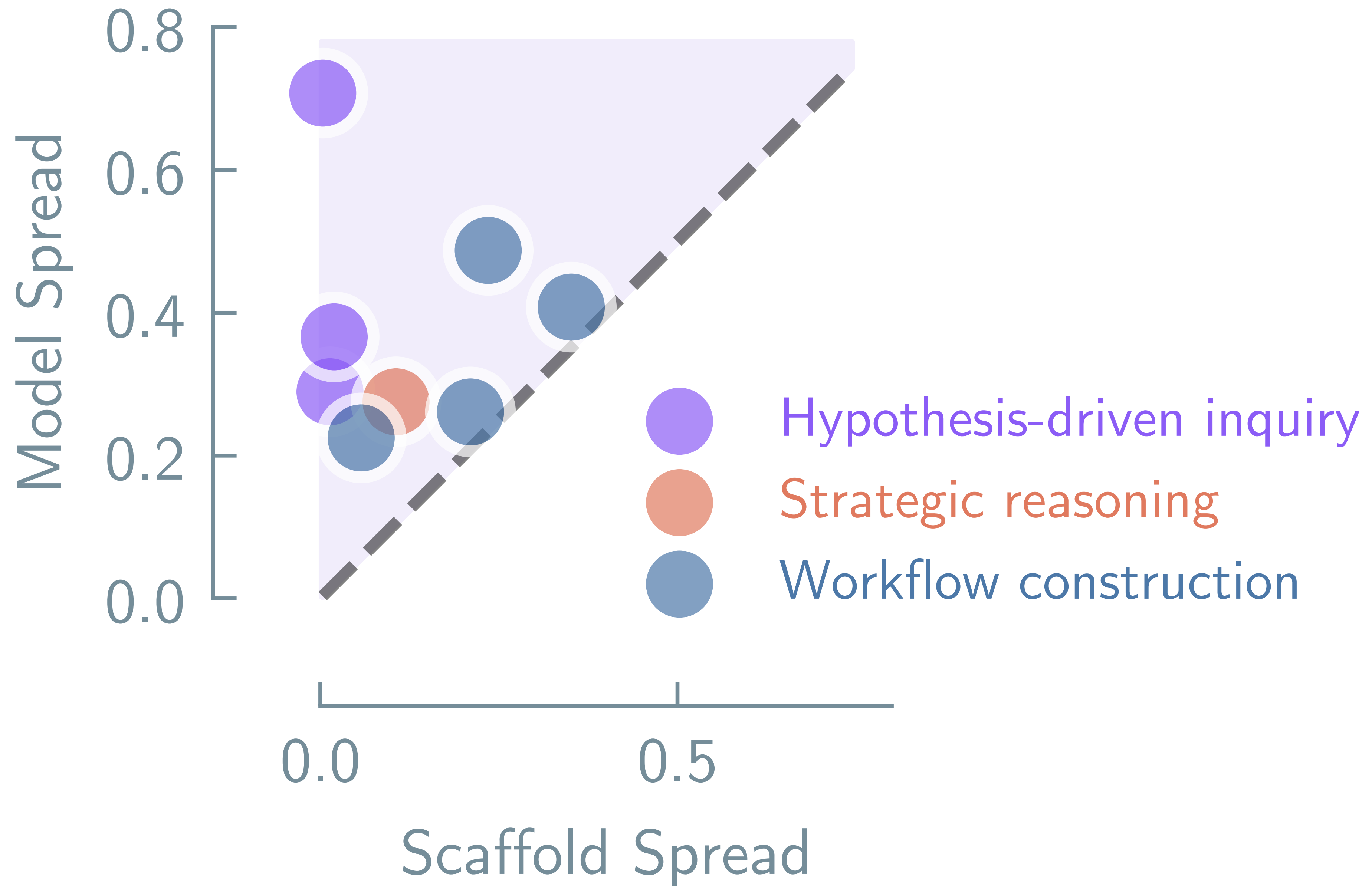
No differences between scaffolds



And for all models

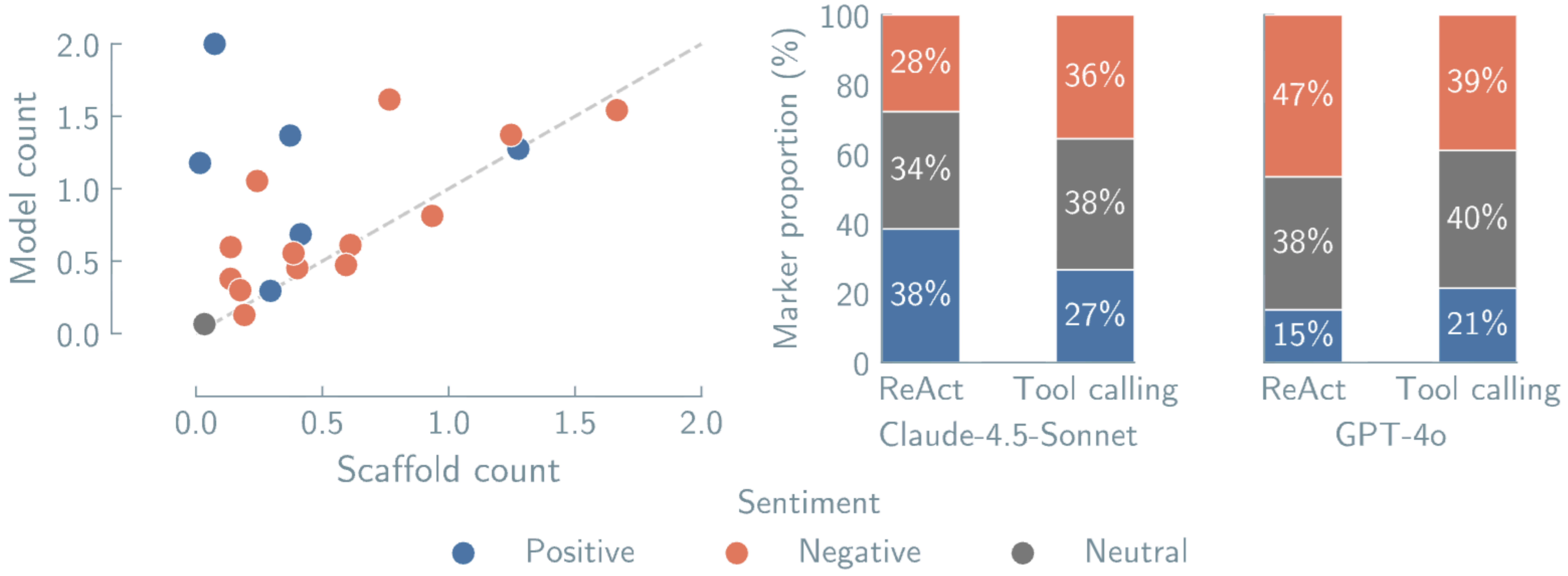
But big differences among tasks

Variance between models bigger than between scaffolds

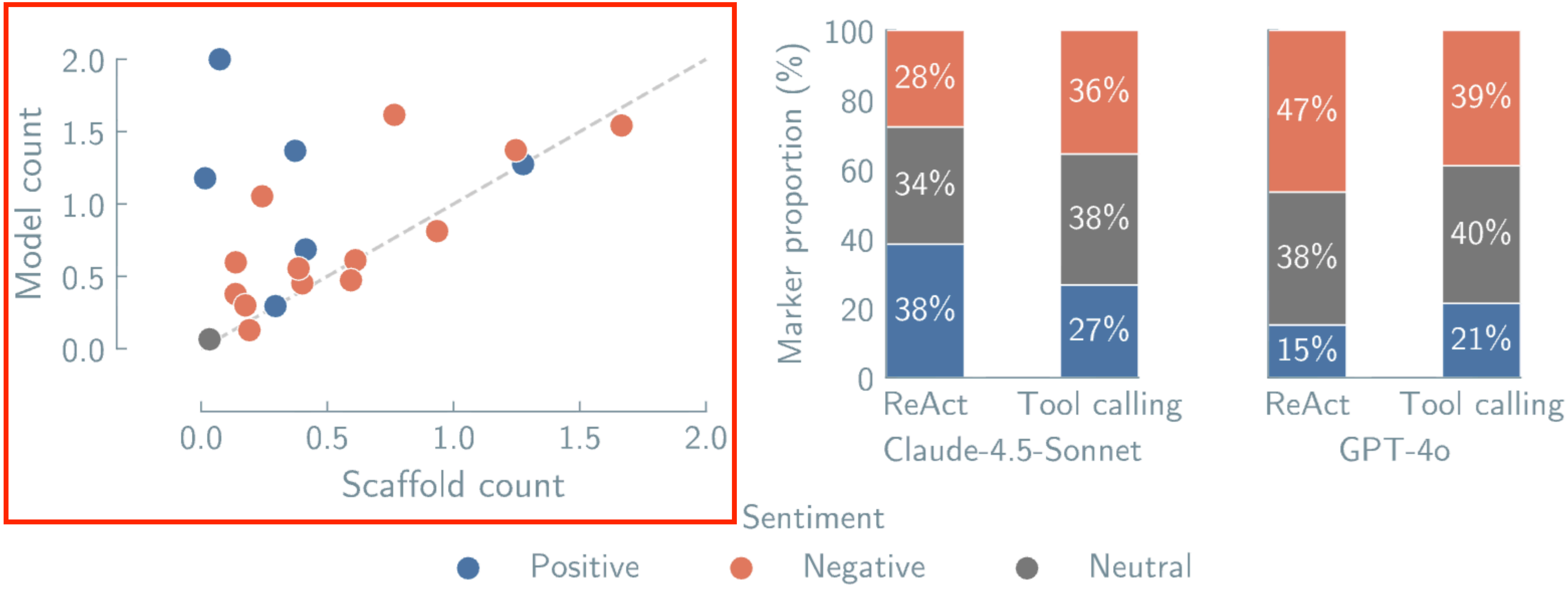


Manual annotation confirms the model influence

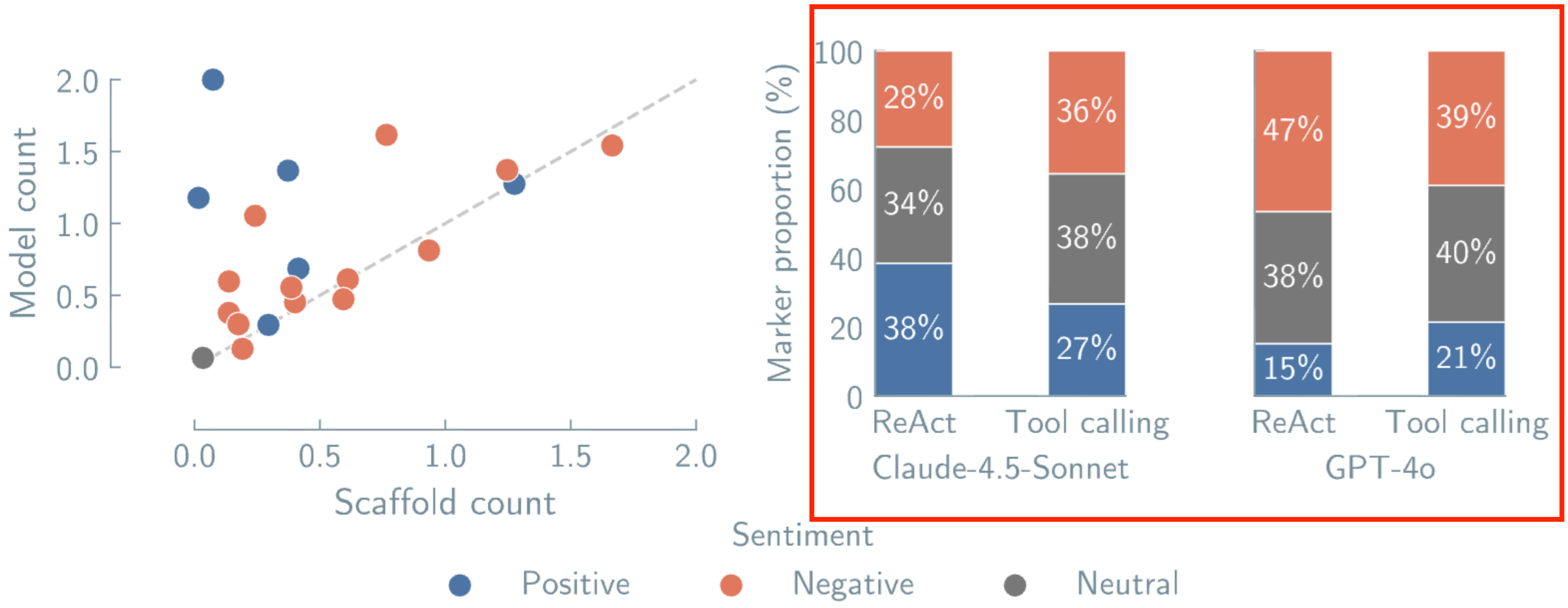
Manual annotation confirms the model influence



Manual annotation confirms the model influence

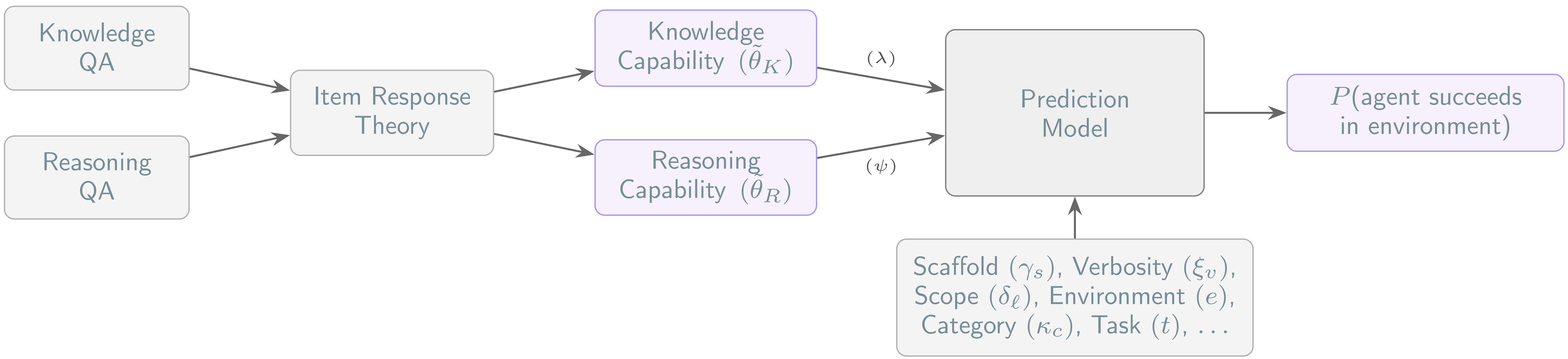


Manual annotation confirms the model influence

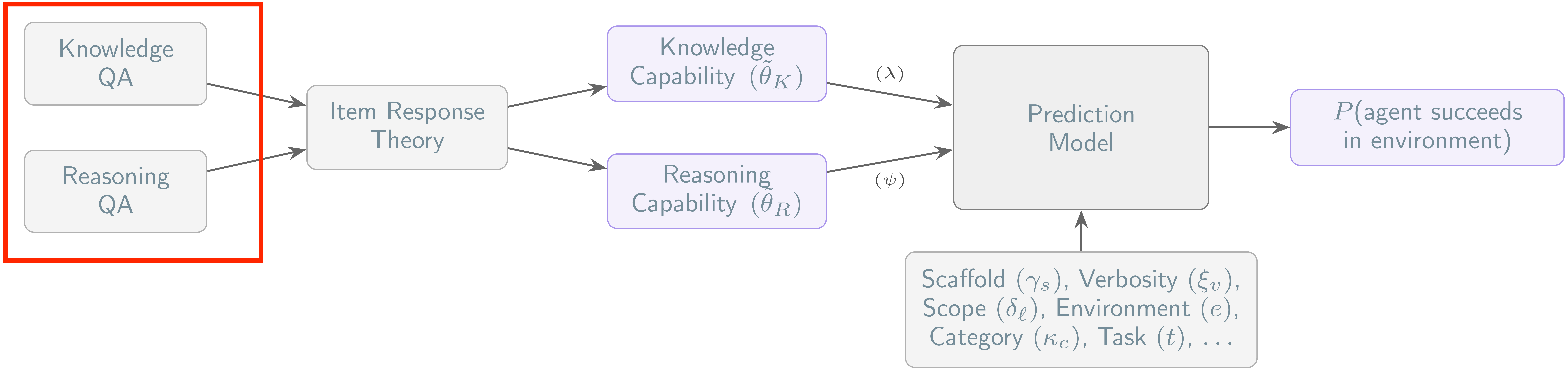


IRT for deeper analysis

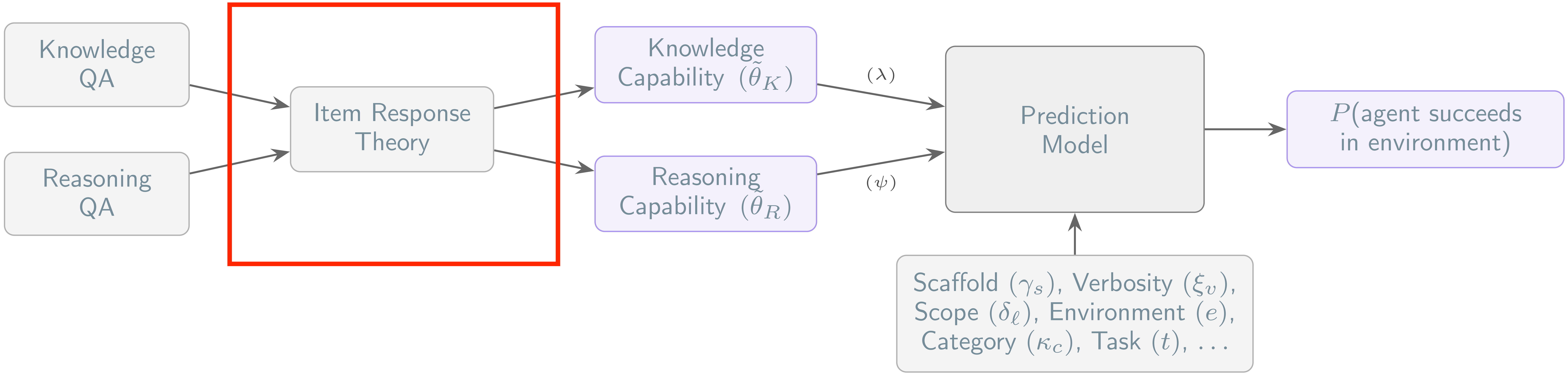
IRT for deeper analysis



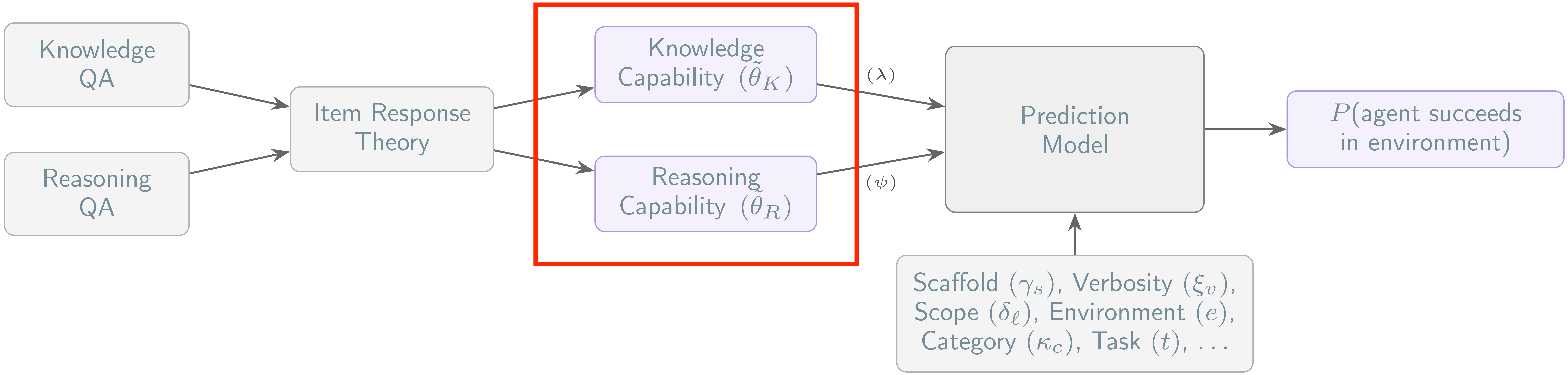
IRT for deeper analysis



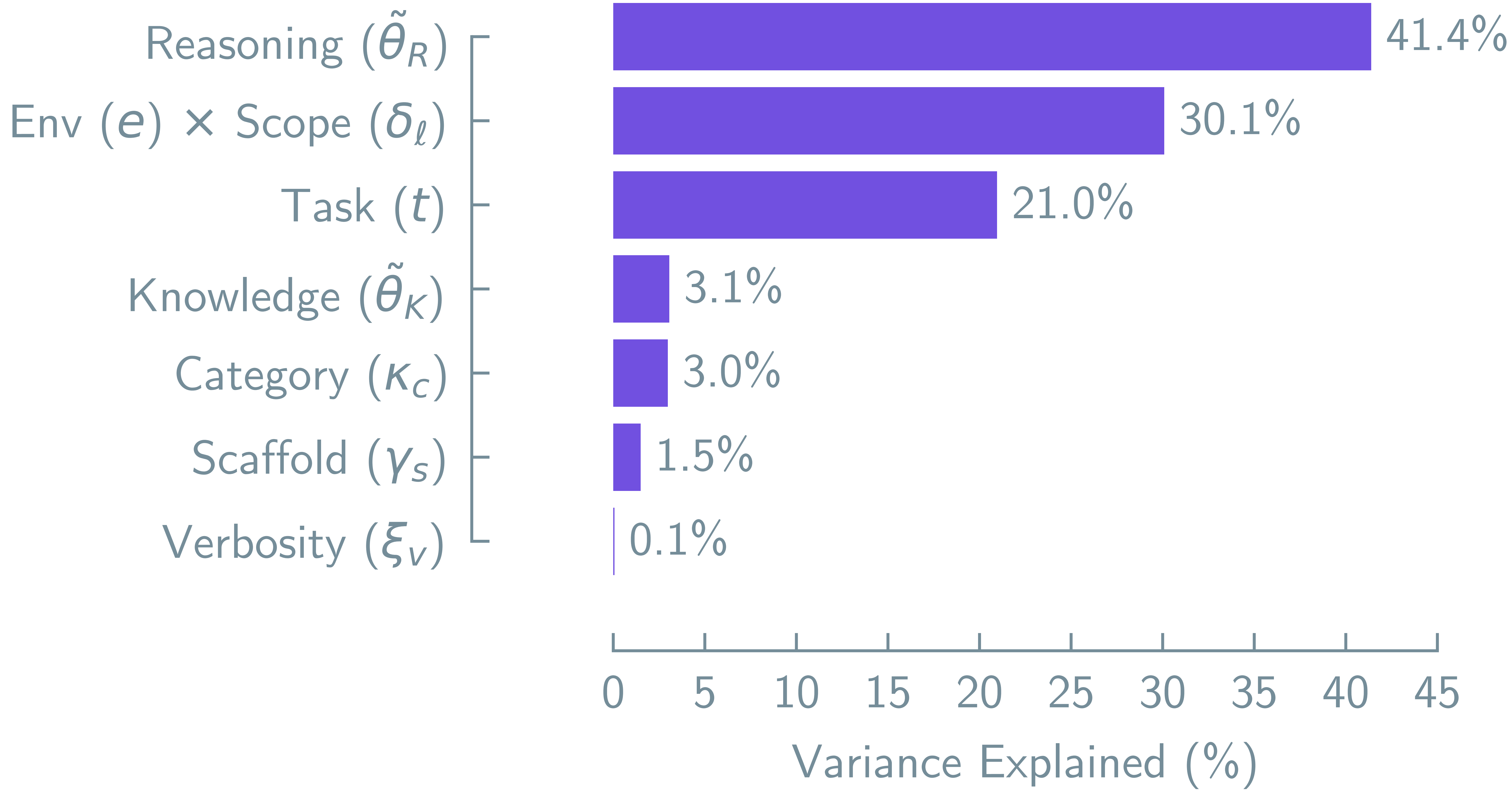
IRT for deeper analysis



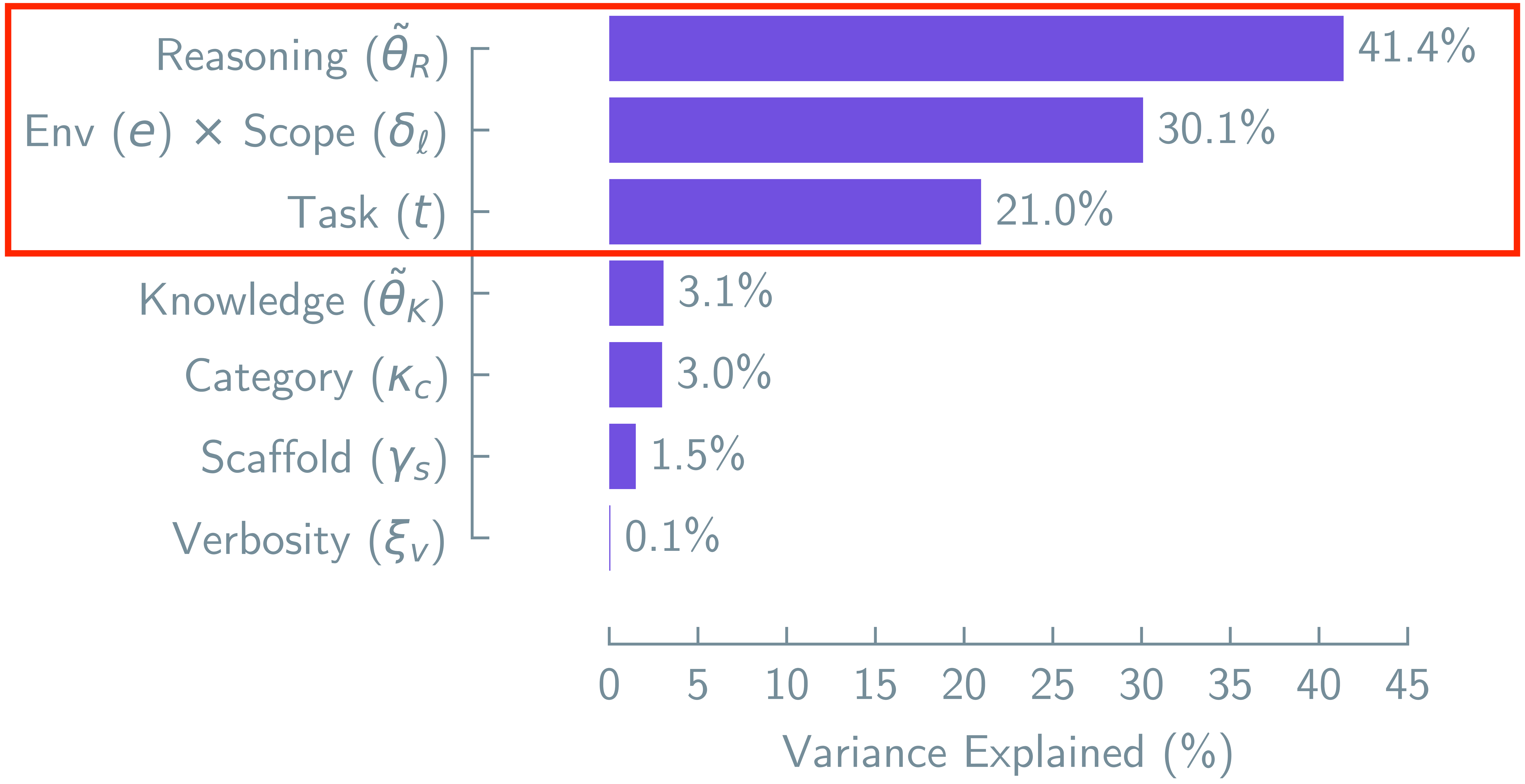
IRT for deeper analysis



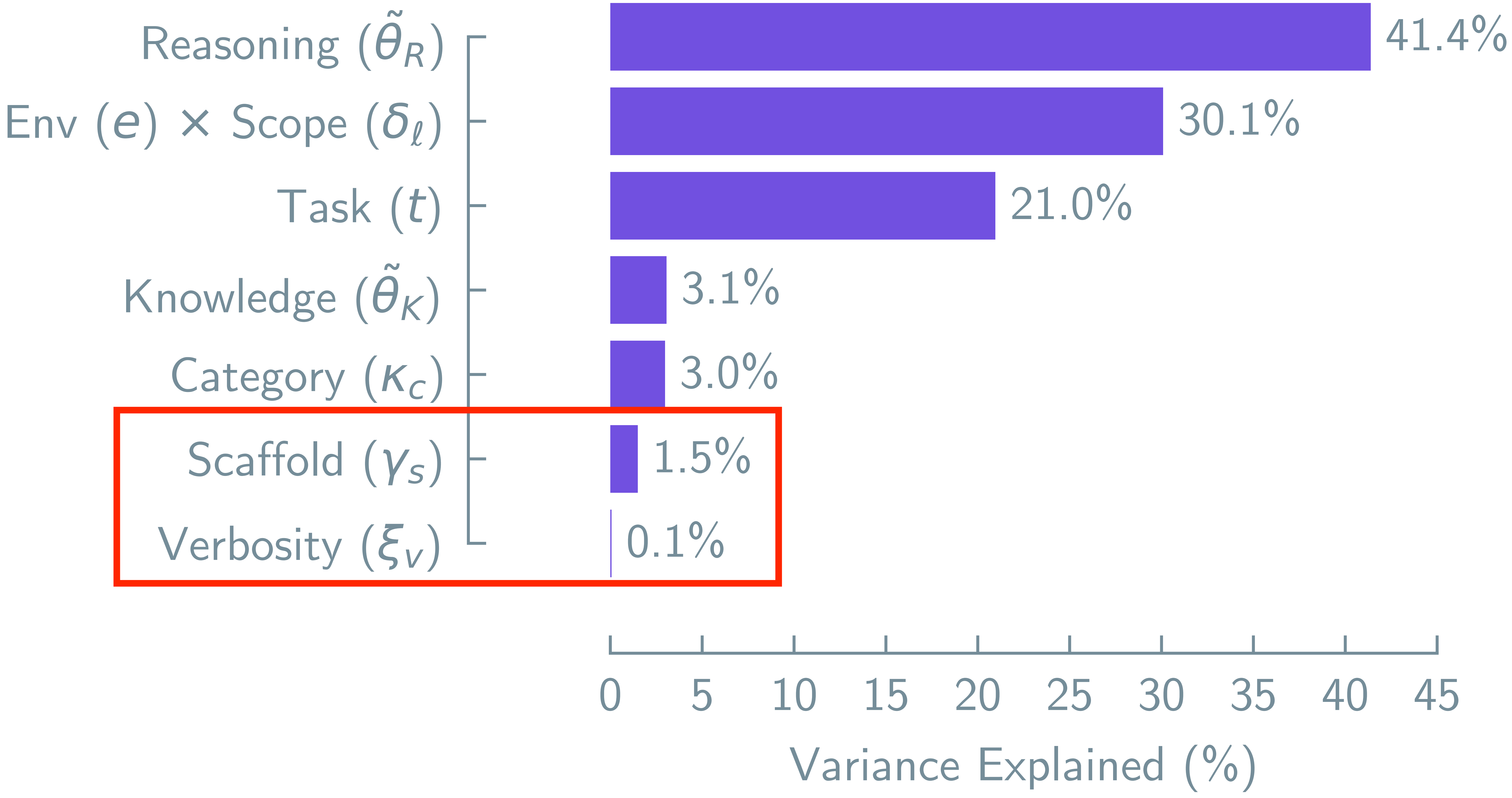
Reasoning ability of the model is the best indicator of performance



Reasoning ability of the model is the best indicator of performance

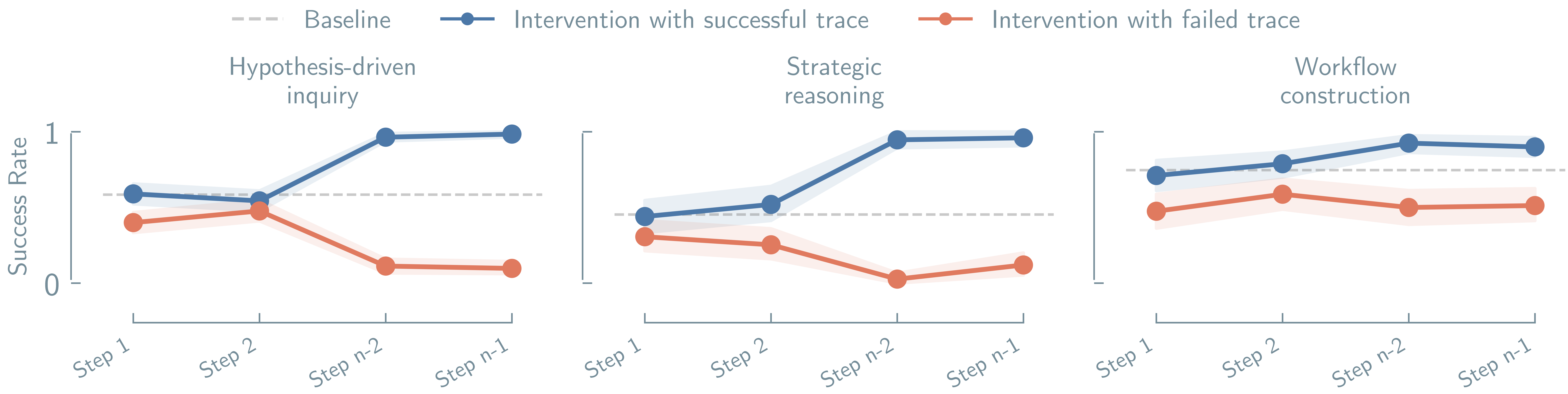


Scaffold and level of detail in tool descriptions do not influence scores

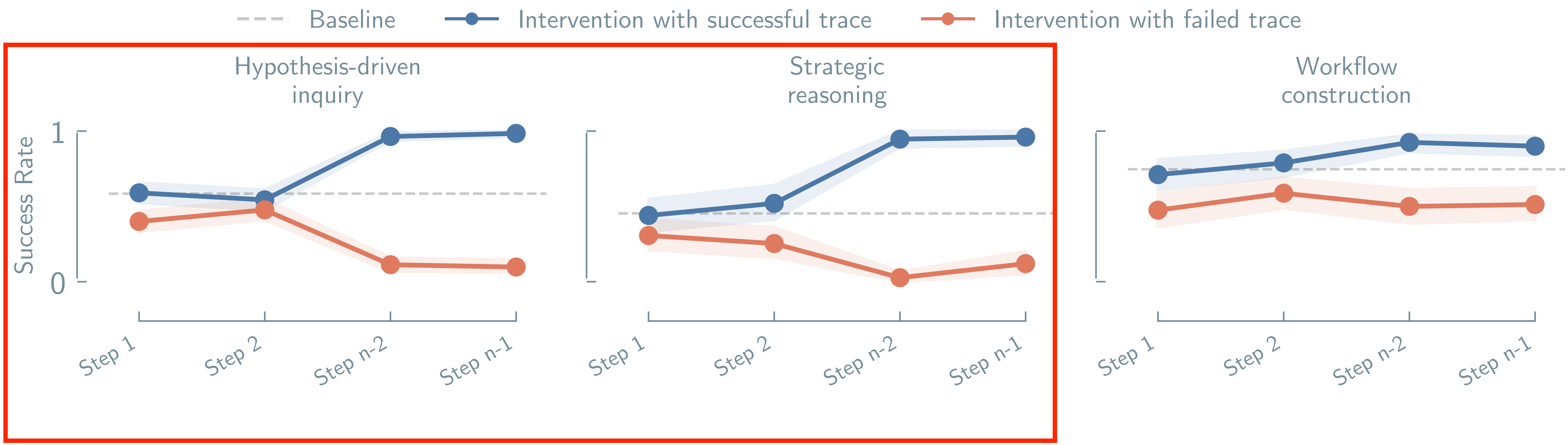


Intervention confirms that context does not matter a lot

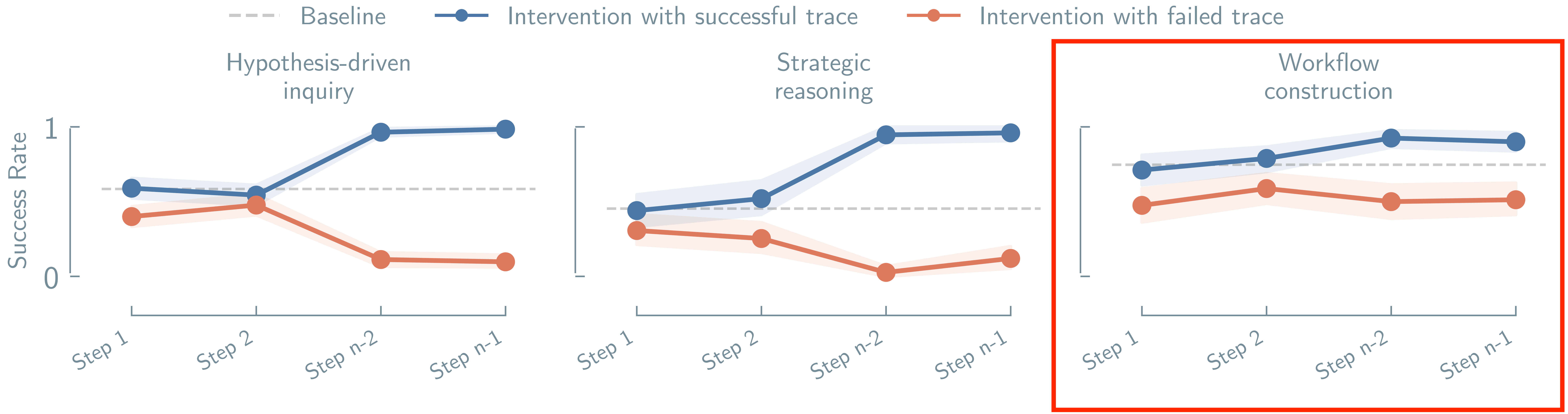
Intervention confirms that context does not matter a lot



Intervention confirms that context does not matter a lot

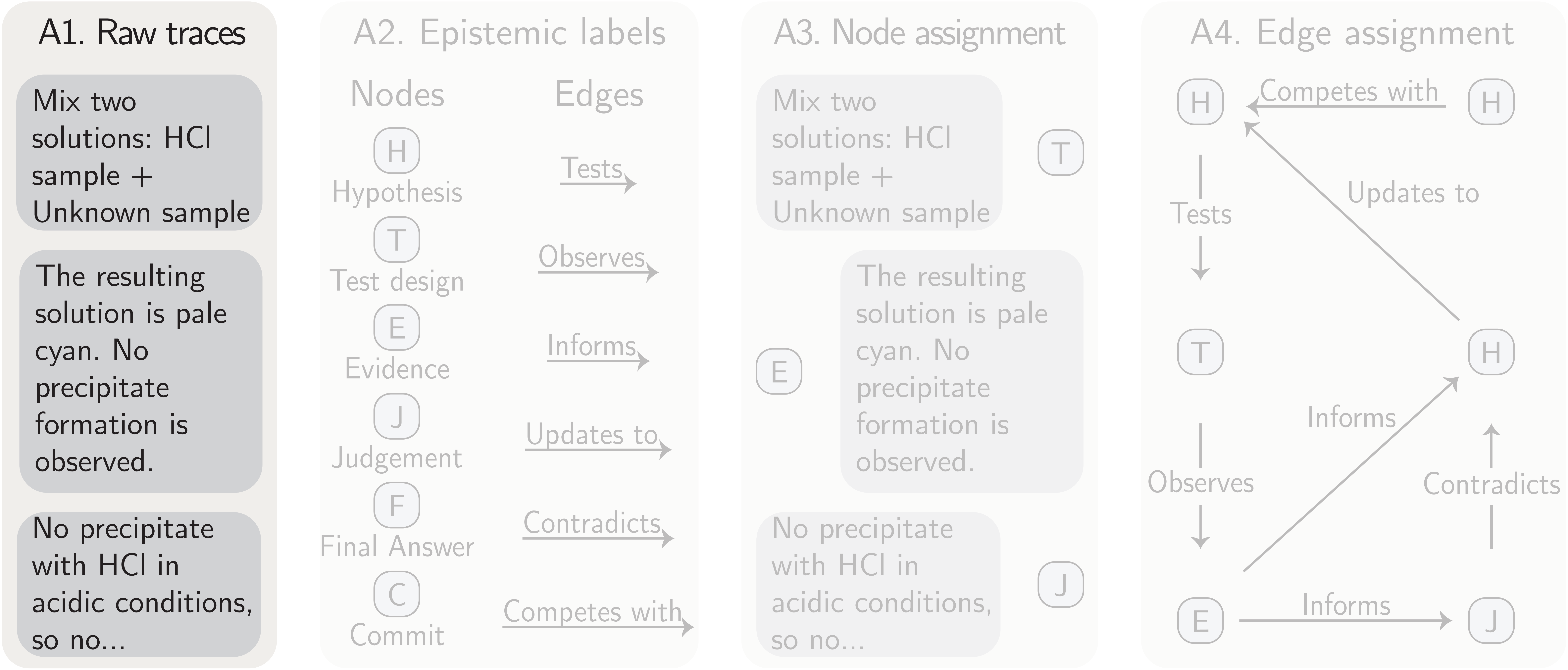


Intervention confirms that context does not matter a lot

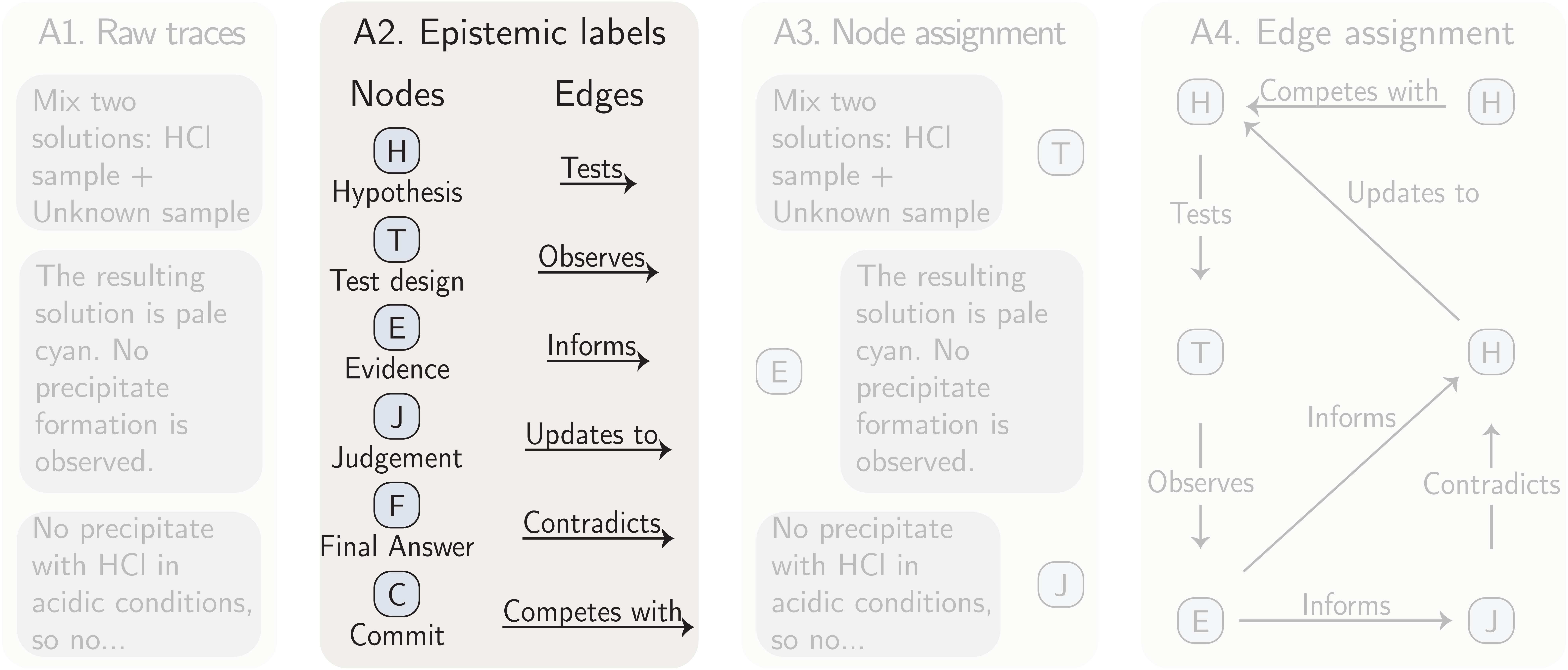


Epistemology

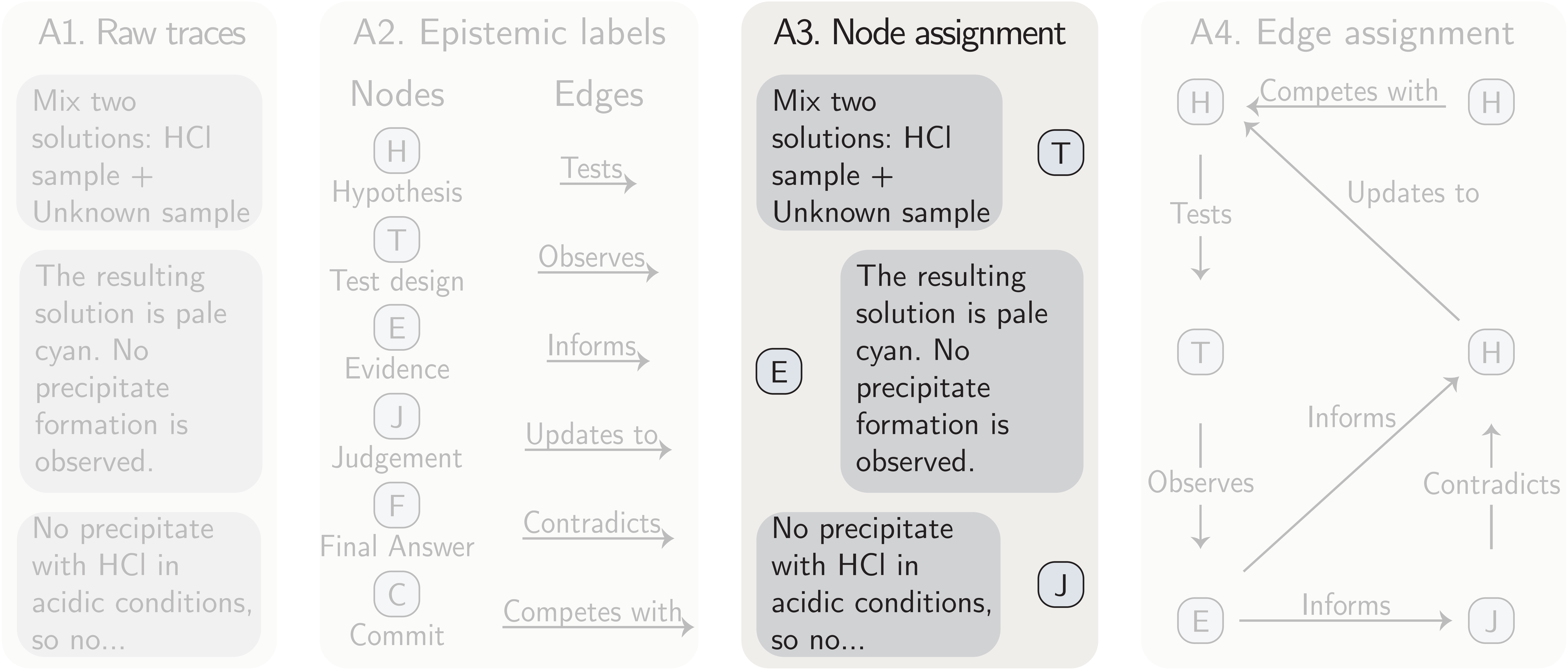
Epistemology



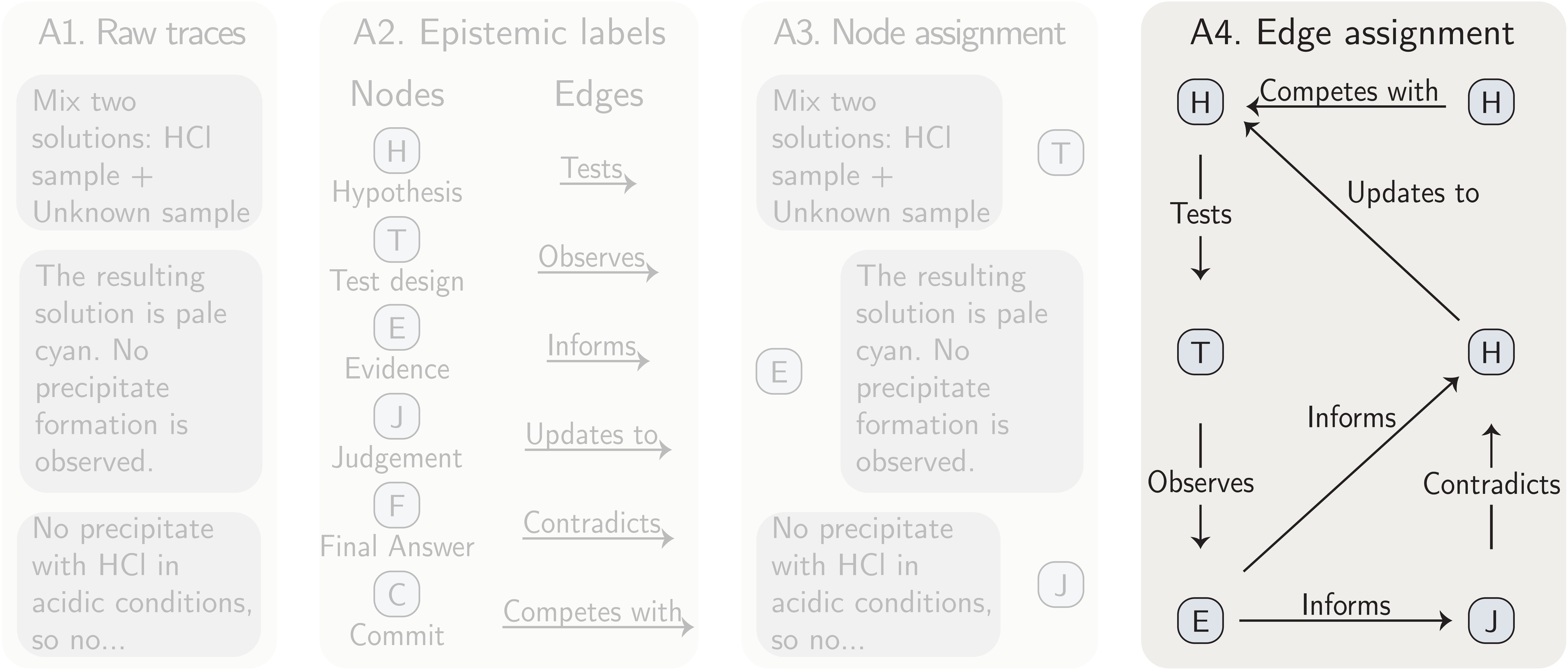
Epistemology



Epistemology

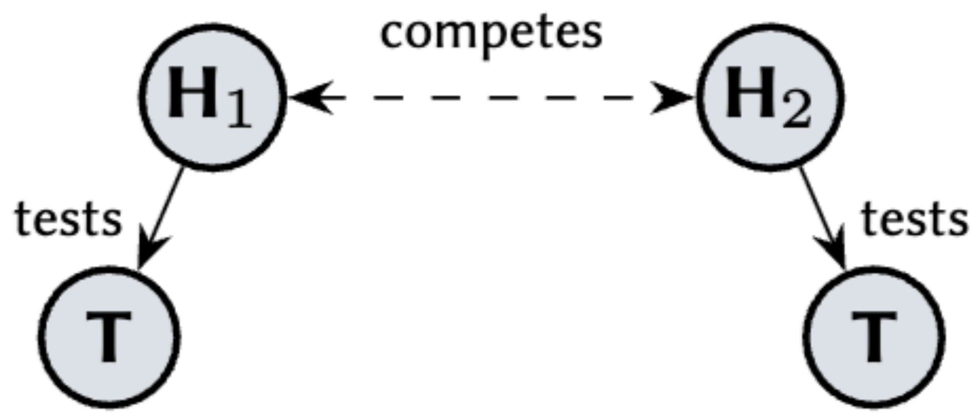


Epistemology

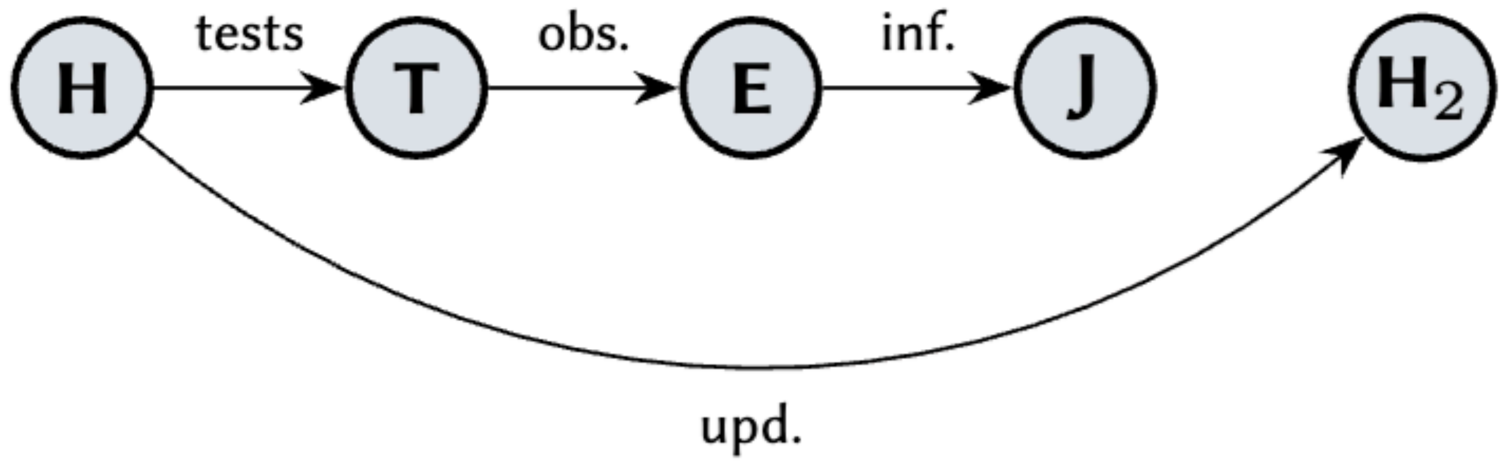


Productive Patterns

Evidence led hypothesis generation



Hypothesis reranking

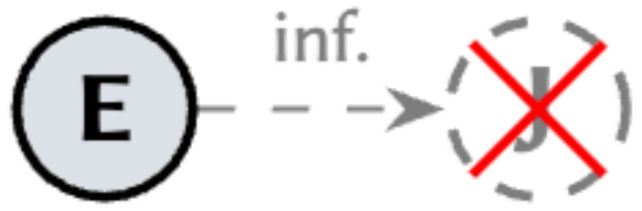


Refutation-driven belief revision

Explore then test transition

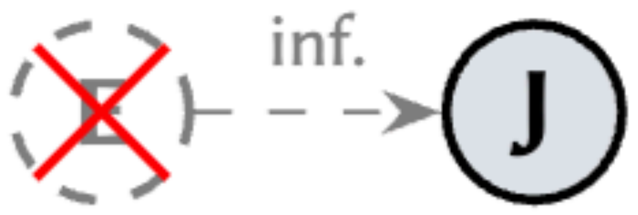
...

Reasoning Breakdowns

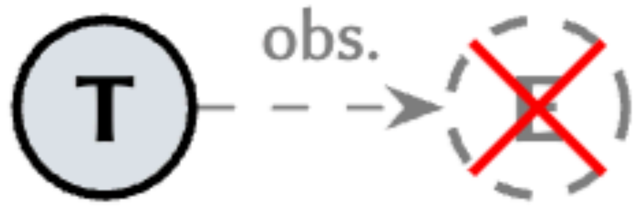


Evidence non-uptake

Disconnected evidence



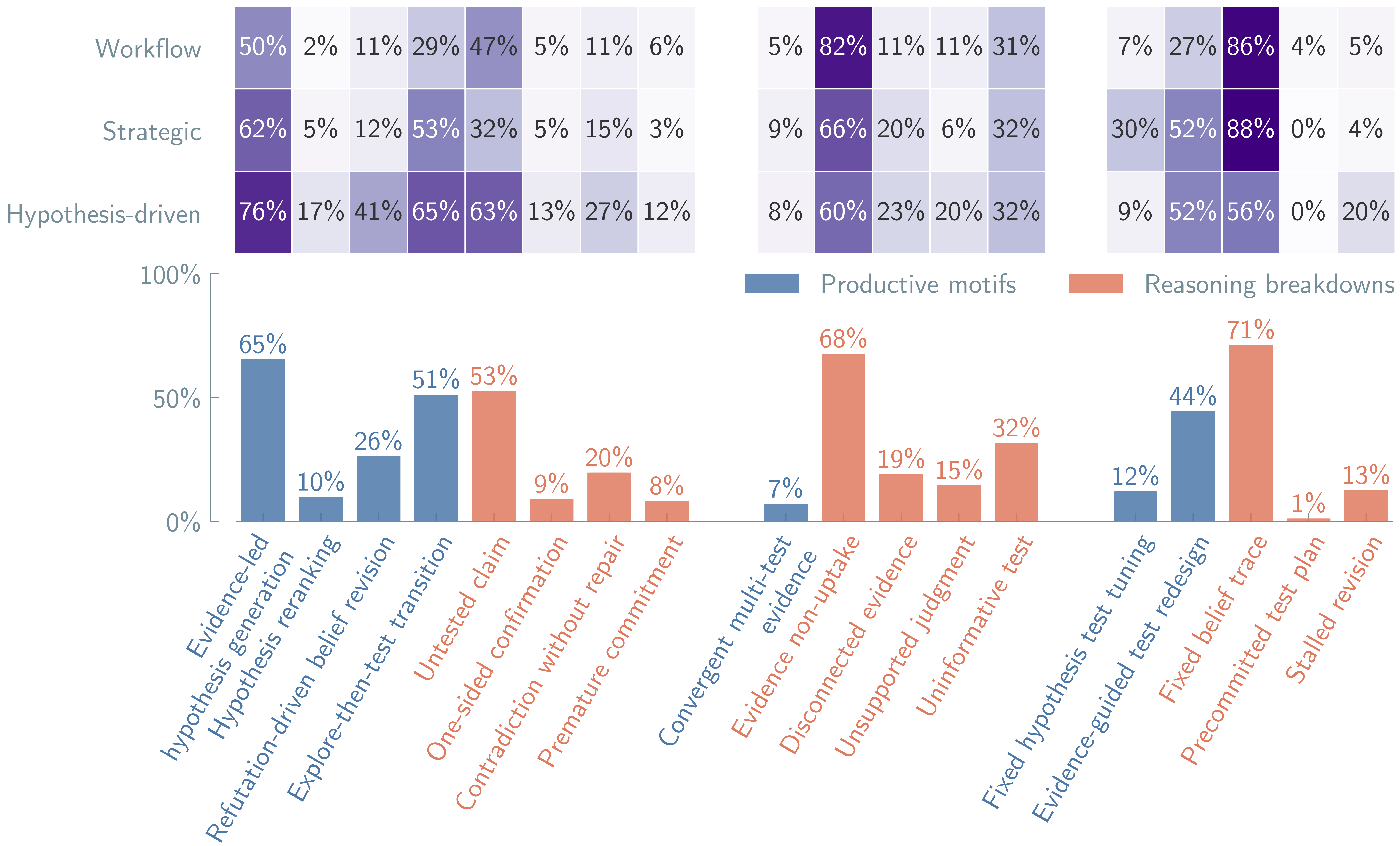
Unsupported judgment



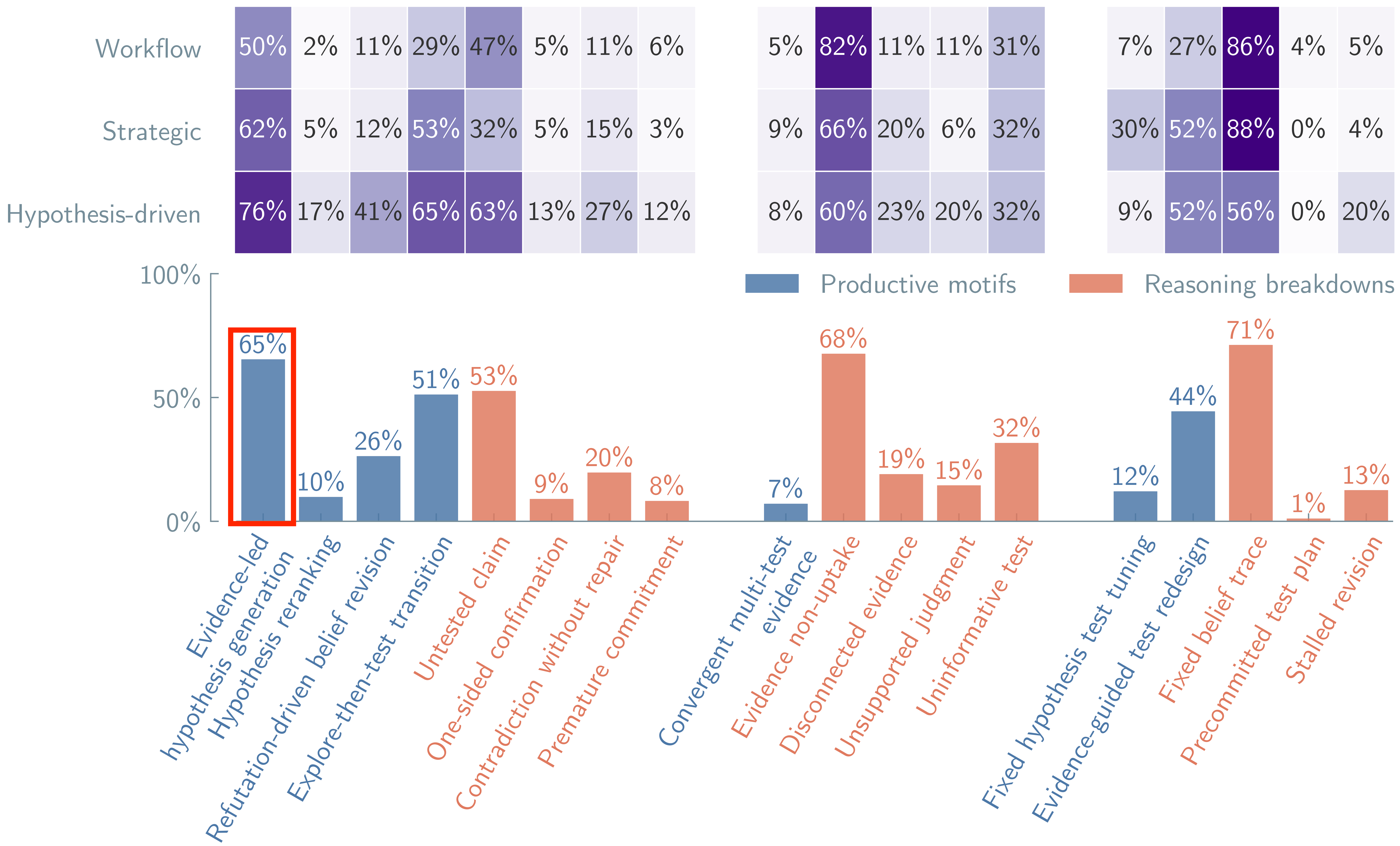
Uninformative test

...

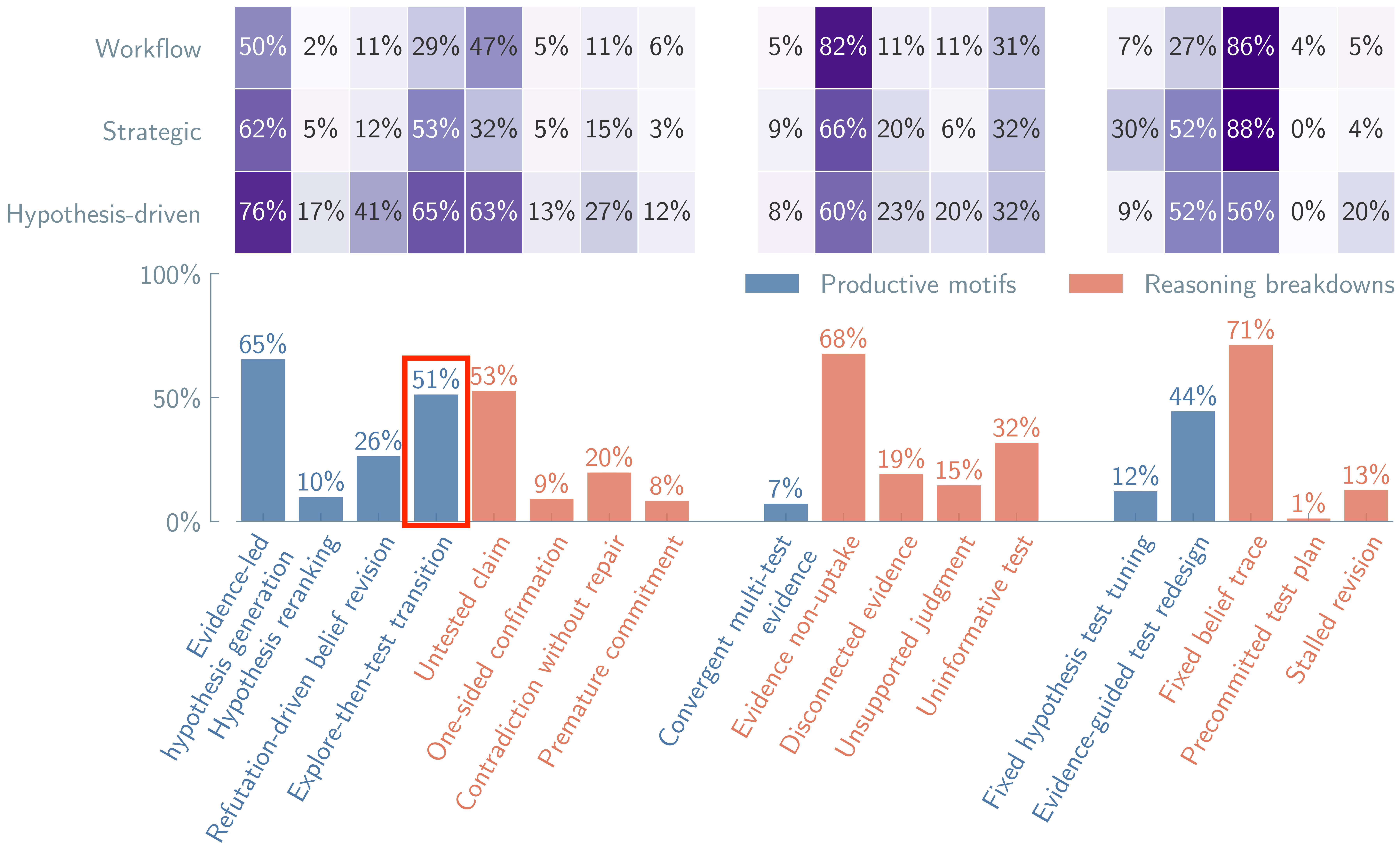
Some Productive Patterns are present



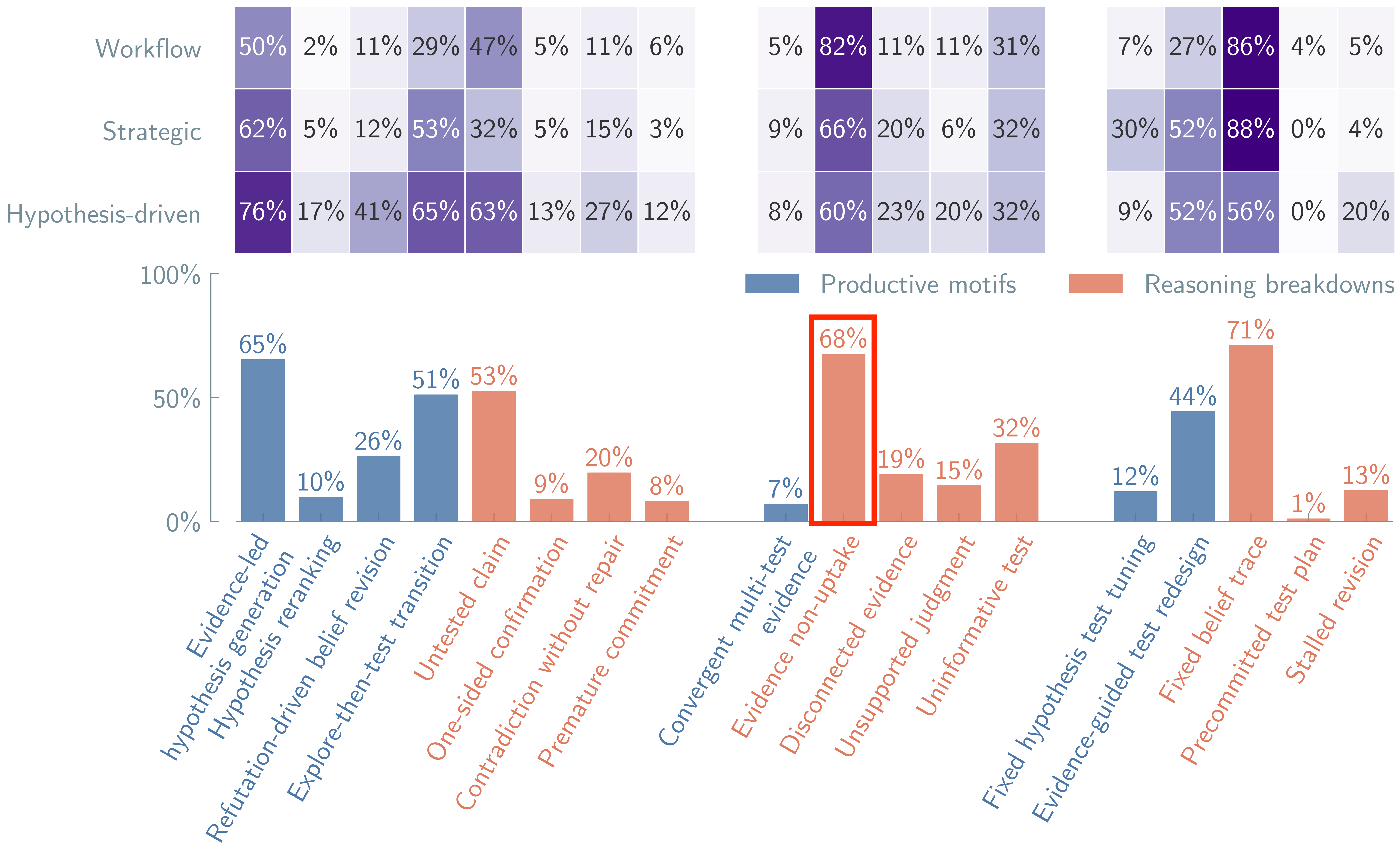
Some Productive Patterns are present



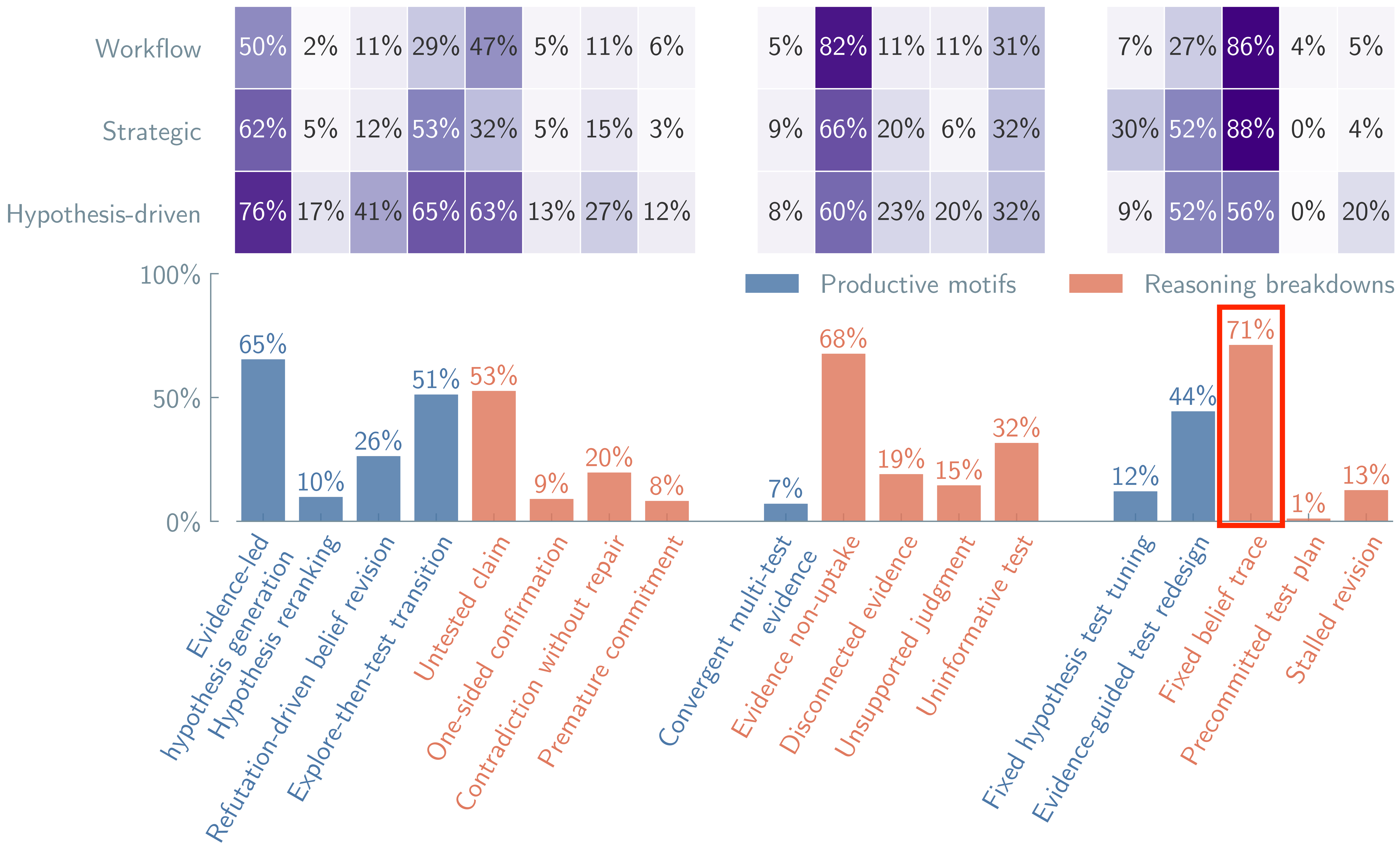
Some Productive Patterns are present



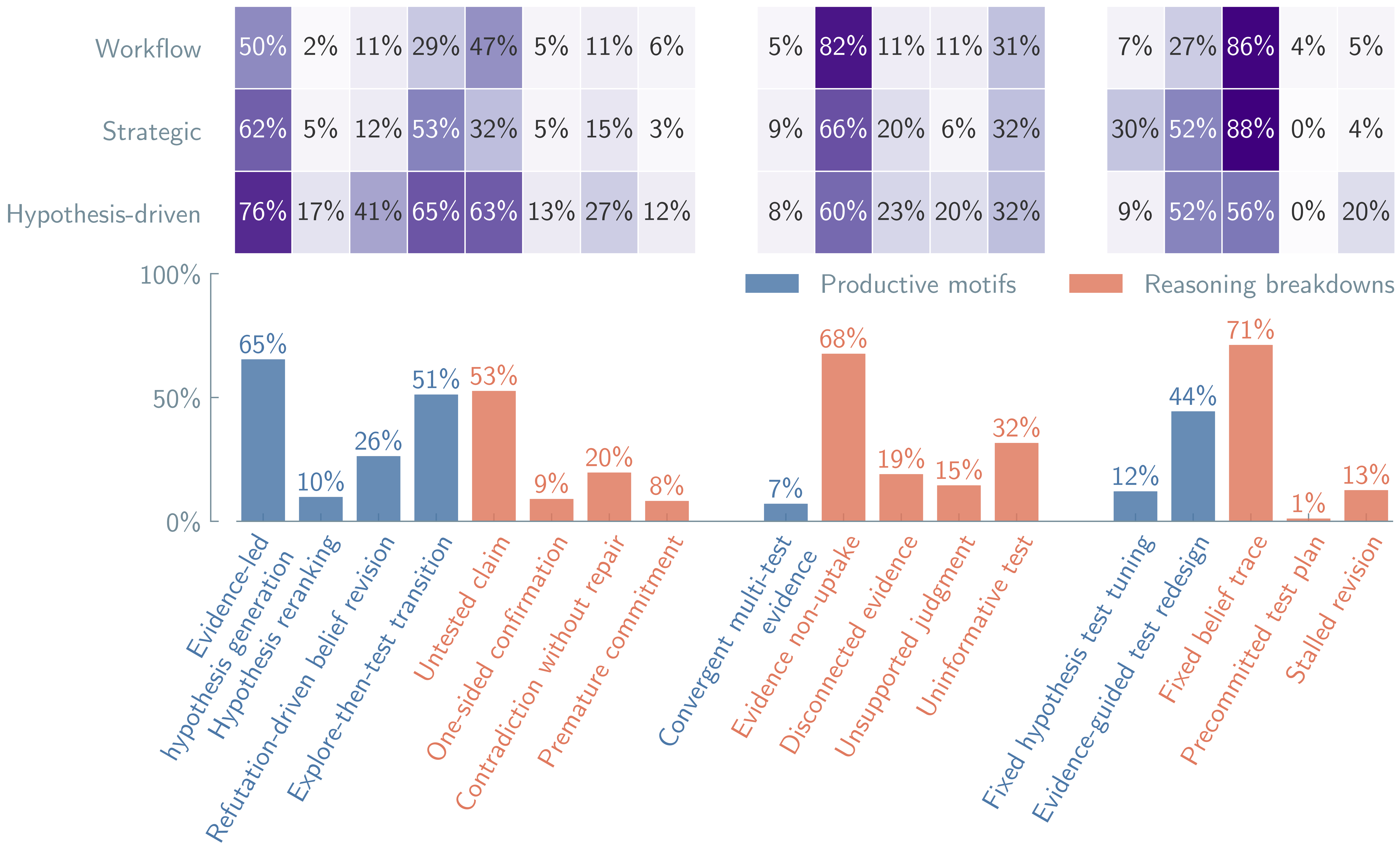
Epistemology



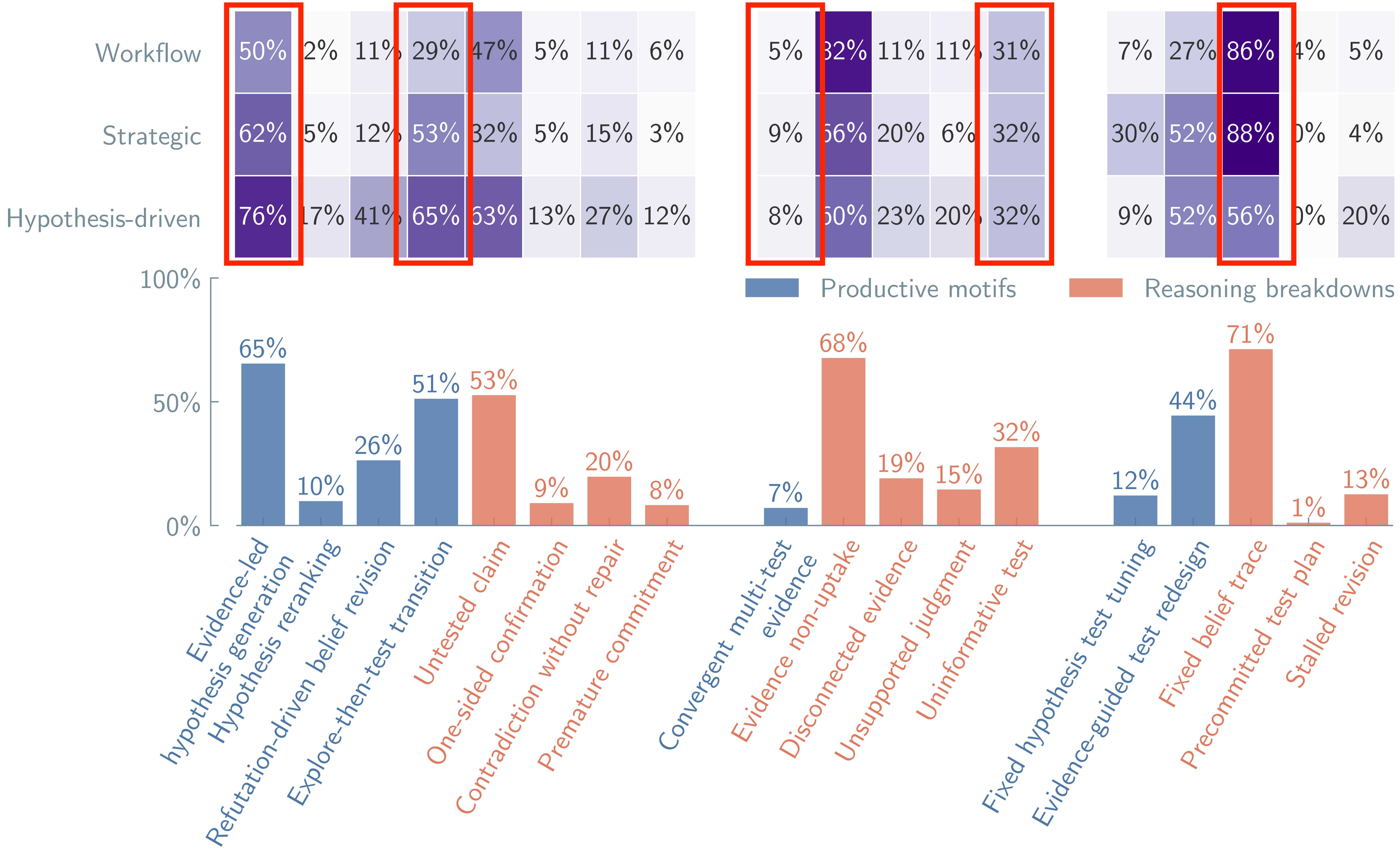
Epistemology



Epistemology



Epistemology



Some concerning behaviors

Stick to the initial belief

Task

Identify the SMILES string of an unknown organic sample. Tools: mass_spectrometry_spectra, proton_nmr_spectra, carbon_nmr_spectra, ir_spectra, hsqc_nmr_spectra, obtain_isomers_from_molecular_formula, validate_smiles, simulate_spectra, get_formula_from_smiles, search_by_smiles, return_possible_fragments.

Stick to the initial belief

Task

Identify the SMILES string of an unknown organic sample. Tools: mass_spectrometry_spectra, proton_nmr_spectra, carbon_nmr_spectra, ir_spectra, hsqc_nmr_spectra, obtain_isomers_from_molecular_formula, validate_smiles, simulate_spectra, get_formula_from_smiles, search_by_smiles, return_possible_fragments.

Agent thought

I need to identify the SMILES string of an organic sample. Since I have no initial information about the sample, I should start by gathering spectral data. Let me begin with mass spectrometry to get the molecular weight.

Action

```
mass_spectrometry_spectra()
```

Stick to the initial belief

Observation

```
m/z 198.99966 (intensity 100)
m/z 200.99765 (intensity 97.65)
```

Stick to the initial belief

Observation

```
m/z 198.99966 (intensity 100)
m/z 200.99765 (intensity 97.65)
```

Agent thought

The mass spectrometry data shows:
- Main peak at m/z 198.99966 (100% intensity)
- Another significant peak at m/z 200.99765 (97.65% intensity)

The pattern shows two peaks separated by 2 mass units with nearly equal intensity. This is characteristic of a compound containing one bromine atom.

Action

```
proton_nmr_spectra()
```

Stick to the initial belief

Observation

```
Deltas 2.62 (t, J = 7.3 Hz, 2H), 2.77 (t, J = 7.3 Hz, 2H),
7.13 (ddd, 2H), 7.43 (ddd, 2H)
```

Agent thought

The ^{1}H NMR shows:
- 2.62 ppm and 2.77 ppm: two CH_2 triplets, J = 7.3 Hz → CH_2-CH_2 unit
- 7.13 and 7.43 ppm (2H each, ddd): aromatic protons of a para-disubstituted benzene ring.
Total: 8 protons.

Action

```
carbon_nmr_spectra()
```

Stick to the initial belief

Observation

```
Deltas: 138.76, 131.45, 131.45, 130.59, 130.59, 119.86, 43.5, 38.6
```

Agent thought

The ^{13}C NMR shows 8 carbon signals: four aromatic (one quaternary near 138, one C-Br near 120, two equivalent CH pairs near 131 and 130) and two aliphatic at 43.5 and 38.6 ppm. Consistent with a para-disubstituted ring + ethyl bridge.

Action

```
ir_spectra()
```

Stick to the initial belief

Observation

```
Wavenumbers (cm⁻¹): 3416, 3385, 3101, 3100, 3084, 3050,
2942, 1612, 1485, 1228, 1018, 798
```

Agent thought

The IR spectra shows:
- 3416, 3385 cm^{-1} - possibly N-H or O-H stretches (but these seem unusual given the NMR data)
- 3101, 3100, 3084, 3050 cm^{-1} - aromatic C-H stretches
- 2942 cm^{-1} - aliphatic C-H stretches
- 1612, 1485 cm^{-1} - aromatic C=C
- 798 cm^{-1} - para-disubstituted benzene

Stick to the initial belief

Action

```
hsqc_nmr_spectra()
```

Observation

HSQC: δ_H/δ_C 2.62/38.6 (2H), 2.77/43.5 (2H), 7.13/131.45 (2H), 7.43/130.59 (2H)

Agent thought

HSQC confirms two CH_2 groups and a para-disubstituted ring. Trying molecular formula $C_9H_{11}Br$ ($M \approx 199$), which is the closest C/H/Br formula to the observed mass. Let me get isomers.

Stick to the initial belief

Action

```
obtain_isomers_from_molecular_formula({"molecular_formula": "C9H11Br", "limit": 10})
```

Observation *'CC1=CC=C(C=C1)CCBr', 'CC1=CC=CC=C1CCBr', 'CCC1=CC=C(C=C1)CBr', . . .*

Agent thought

Several candidates. The para-substituted one, 4-methylphenethyl bromide (`CC1=CC=C(C=C1)CCBr`), looks most consistent with the NMR. Let me simulate its spectra.

Action

```
simulate_spectra({"smiles": "CC1=CC=C(C=C1)CCBr"})
```

Stick to the initial belief

Observation

^{1}H: 2.25 (s, 3H, CH_3), 2.88 (t, 2H), 3.39 (t, 2H), 7.05 (d, 2H), 7.13 (d, 2H)
^{13}C: 137.7, 134.0, 129.5, 129.0, 33.5, 32.5, 21.0

Agent thought

The simulated NMR is close but not a perfect match (extra CH_3 singlet near 2.25 not seen in the observed spectrum). Let me try a few more isomers.

Action

```
simulate_spectra({"smiles": "BrCCC1=CC=C(C=C1)Br"})
```

Stick to the initial belief

Action

```
simulate_spectra({"smiles": "BrCCC1=CC=C(C=C1)Br"})
```

Observation

^{1}H: 3.14 (t, 2H), 3.55 (t, 2H), 7.10 (d, 2H), 7.44 (d, 2H)
^{13}C: 137.5, 131.5, 130.7, 120.0, 38.5, 32.5

Agent thought

The simulated spectrum for BrCCC1=CC=C(C=C1)Br matches the observed NMR *almost perfectly*.
But the molecular formula is $C_8H_8Br_2$, MW $\approx$ 264, not 199.
The NMR matches but the mass does not. Moving on.

Stick to the initial belief

Action

```
return_possible_fragments()
```

Stick to the initial belief

Action

```
return_possible_fragments()
```

Observation *'[CH2-*

```
c1ccc(Br)cc1', '[CH2-]N', '[CH2-]Cc1cc[c-]cc1',
'NCCc1cc[c-]cc1', '[CH2-]CN', 'Brc1cc[c-]cc1',
'[c-]1cc[c-]cc1', '[CH2-]c1cc[c-]cc1', '[CH2-]Cc1ccc(Br)cc1', '[CH2-][CH2-]']
```

Stick to the initial belief

Action

```
return_possible_fragments()
```

Observation *'[CH2-*

```
c1ccc(Br)cc1', '[CH2-]N', '[CH2-]Cc1cc[c-]cc1',
'NCCc1cc[c-]cc1', '[CH2-]CN', 'Brc1cc[c-]cc1',
'[c-]1cc[c-]cc1', '[CH2-]c1cc[c-]cc1', '[CH2-]Cc1ccc(Br)cc1', '[CH2-][CH2-]']
```

Agent thought

The fragments shown don't directly help me understand the MS peak at 199. These are fragments with removed atoms, and they don't explain the m/z 199 peak in mass spectrometry.

The MS peak at 199 must correspond to the $[M+H]^+$ adduct of $C_9H_{11}Br$ (M = 198, M+H = 199). The compound is $C_9H_{11}Br$, and the most likely structure given the para-substitution is `CC1=CC=C(C=C1)CCBr` (4-methylphenethyl bromide).

Even though the simulated NMR doesn't match perfectly, this must be the correct structure based on the molecular formula from MS.

Stick to the initial belief

Agent thought

The fragments shown don't directly help me understand the MS peak at 199. These are fragments with removed atoms, and they don't explain the m/z 199 peak in mass spectrometry.

The MS peak at 199 must correspond to the $[M+H]^+$ adduct of $C_9H_{11}Br$ (M = 198, M+H = 199). The compound is $C_9H_{11}Br$, and the most likely structure given the para-substitution is `CC1=CC=C(C=C1)CCBr` (4-methylphenethyl bromide).

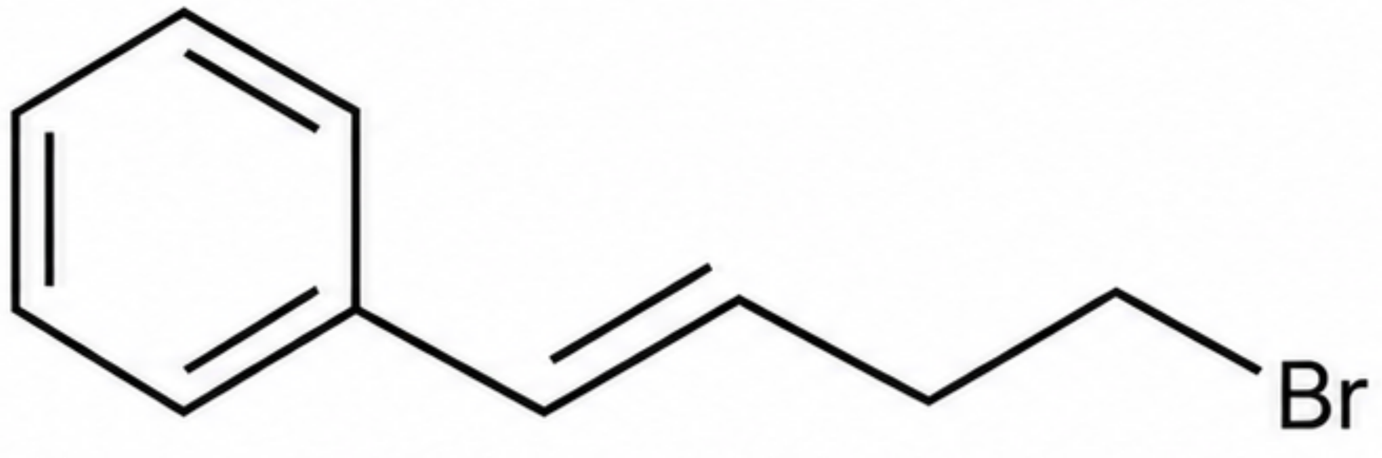
Even though the simulated NMR doesn't match perfectly, this must be the correct structure based on the molecular formula from MS.

Final answer

`CC1=CC=C(C=C1)CCBr`

Stick to the initial belief

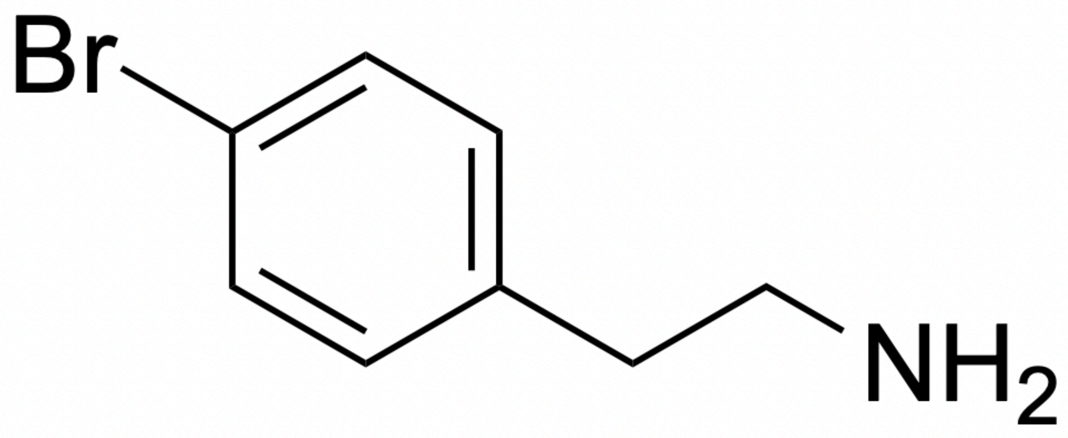
Agent's submitted answer



CC1=CC=C(C=C1)CCBr

Formula: $C_9H_{11}Br$

Correct answer (ground truth)



NCCc1ccc(Br)cc1

Formula: $C_8H_{10}BrN$

Sampling loop

Task

Infer the circuit topology and resistor values from 10 node-to-node resistance measurements between A, B, N_b1, X1, X2.

Measurements (Ω):
A–B: 41.364
A–N_b1: 11.667
A–X1: 6.667
A–X2: 25.0
B–N_b1: 29.697
B–X1: 34.697
B–X2: 16.364
N_b1–X1: 5.0
N_b1–X2: 13.333
X1–X2: 18.333

Submit JSON: {"resistors": {...}, "connections": [...]}.

Sampling loop

Agent thought

I'll analyze the node-to-node resistance measurements to infer the circuit topology and resistor values. Let me start by examining the measurements more carefully and working through the circuit structure systematically.

Action

```
write_file(path="measurements.json", content=<...621 bytes...>)
```

Observation

```
Successfully wrote to measurements.json
```

Sampling loop

Agent thought

Let me try a different approach. I'll test a specific topology hypothesis:

Action

```
validate_measurements(topology = {"R1": 5, "R2": 10, "R3": 10, "R4": 10, "R5": 20, "R6":
30} [["A", "X1", "R2"])
```

Sampling loop

Agent thought

Let me try a different approach. I'll test a specific topology hypothesis:

Action

```
validate_measurements(topology = {"R1": 5, "R2": 10, "R3": 10, "R4": 10, "R5": 20, "R6":
30} [["A", "X1", "R2"]])
```

Observation

{ "error": "Validation failed: string indices must be integers, not 'str'" }

Sampling loop

Agent thought

Let me fix the JSON format:

Action

```
validate_measurements(topology = {"R1": 5, "R2": 10, "R3": 10, "R4": 10, "R5": 20, "R6":
30} [["A", "X1", "R2"])
```

Observation

{ "total_error": 62.58388235294109, "max_error": 13.636, "mean_error": 6.258388235294109, "detailed_errors": [{ "nodes": "A-B", "predicted": 41.76470588235284, "actual": 41.364, "error": 0.4007058823528453, "relative_error": 0.009687309794817845 }, { "nodes": "A-N_b1", "predicted": 5.88235294117647, "actual": 11.667, "error": 5.78464705882353, "relative_error": 0.4958127246784546 }, { "nodes": "A-X1", . . .

Sampling loop

Agent thought

Still not good. Let me think more systematically. Looking at the exact values: - N_b1-X1 = 5.0 (exact) - A-X1 = 6.667 = 20/3 - A-N_b1 = 11.667 = 35/3 - N_b1-X2 = 13.333 = 40/3 - X1-X2 = 18.333 = 55/3
Let me try a different approach with these exact fractions:

Action

```
validate_measurements(topology = {"R1": 5, "R2": 10, "R3": 20, "R4": 10, "R5": 20, "R6":
15} [["N_b1", "X1", "R1"])
```

Observation

{ "total_error": 76.93614814814815, "max_error": 16.73437037037037, "mean_error": 7.693614814814815, "detailed_errors": [{ "nodes": "A-B", "predicted": 24.629629629629626, "actual": 41.364, "error": 16.73437037037037, "relative_error": 0.4045636391637746 }, { "nodes": "A-N_b1", "predicted": 9.999999999999998, "actual": 11.667, "error": 1.6670000000000016, "relative_error": 0.1428816319533729 }, { "node . . .

Sampling loop

Agent thought

Let me think differently. Maybe there's a connection from A to N_b1 directly. Let me try:

Action

```
validate_measurements(topology = {"R1": 5, "R2": 10, "R3": 10, "R4": 20, "R5": 30, "R6":
20} [["N_b1", "X1", "R1"]])
```

Observation

{ "total_error": 35.185851851851936, "max_error": 7.962962962962973, "mean_error": 3.5185851851851937, "detailed_errors": [{ "nodes": "A-B", "predicted": 37.03703703703697, "actual": 41.364, "error": 4.32696296296303, "relative_error": 0.10460697618612877 }, { "nodes": "A-N_b1", "predicted": 5.925925925925926, "actual": 11.667, "error": 5.741074074074074, "relative_error": 0.4920780041205172 }, { "node . . .

Sampling loop

Agent thought

Better! N_b1-X2 is almost perfect. But still errors. Let me adjust the resistor values:

Action

```
validate_measurements(topology = {"R1": 5, "R2": 10, "R3": 20, "R4": 30, "R5": 30, "R6":
18} [["N_b1", "X1", "R1"])
```

Observation

{ "total_error": 21.111111111111192, "max_error": 4.303000000000001, "mean_error": 2.111111111111119, "detailed_errors": [{ "nodes": "A-B", "predicted": 39.777777777777715, "actual": 41.364, "error": 1.5862222222222826, "relative_error": 0.03834789242390201 }, { "nodes": "A-N_b1", "predicted": 8.444444444444443, "actual": 11.667, "error": 3.222555555555557, "relative_error": 0.2762111558717371 }, { "no . . .

Sampling loop

Agent thought

Getting better! Let me fine-tune more:

Action

```
validate_measurements(topology = {"R1": 5, "R2": 10, "R3": 30, "R4": 30, "R5": 20, "R6":
18} [["N_b1", "X1", "R1"])
```

Observation

{ "total_error": 23.604775510204032, "max_error": 4.999999999999993, "mean_error": 2.360477551020403, "detailed_errors": [{ "nodes": "A-B", "predicted": 38.000000000000036, "actual": 41.364, "error": 3.3639999999999617, "relative_error": 0.08132675756696553 }, { "nodes": "A-N_b1", "predicted": 9.795918367346939, "actual": 11.667, "error": 1.8710816326530608, "relative_error": 0.16037384354616105 }, { " . . .

Sampling loop

Agent thought

Let me try different values:

Action

```
validate_measurements(topology = {"R1": 5, "R2": 10, "R3": 35, "R4": 35, "R5": 20, "R6": 18} [["N_b1", "X1", "R1"])
```

Observation

{ "total_error": 18.934907563025313, "max_error": 3.9212352941176576, "mean_error": 1.8934907563025312, "detailed_errors": [{ "nodes": "A-B", "predicted": 39.21848739495792, "actual": 41.364, "error": 2.1455126050420787, "relative_error": 0.051869079514603976 }, { "nodes": "A-N_b1", "predicted": 10.294117647058822, "actual": 11.667, "error": 1.3728823529411773, "relative_error": 0.11767226818729556 }, { . . .

Sampling loop

The agent called this tool 18 times in this run.

Sampling loop

The agent called this tool 18 times in this run.

Sampling loop

The agent called this tool 18 times in this run.

Agent thought

Error: Maximum iterations reached without finding a final answer.

Sampling loop

The agent called this tool 18 times in this run.

Agent thought

Error: Maximum iterations reached without finding a final answer.

In the Spectroscopic environment similar behaviors are observed very ofter

Contradiction

Task

20 mL aqueous sample made by dissolving one pure inorganic salt. Identify the cation and anion. Reagents: $AgNO_3$, $Ba(NO_3)_2$, NH_3, KOH, K_2CrO_4, sulfide, HNO_3... Plus flame test, pH, color.

Contradiction

Task

20 mL aqueous sample made by dissolving one pure inorganic salt. Identify the cation and anion. Reagents: $AgNO_3$, $Ba(NO_3)_2$, NH_3, KOH, K_2CrO_4, sulfide, HNO_3... Plus flame test, pH, color.

Agent thought

I need to identify an unknown inorganic salt dissolved in water. The salt consists of one cation and one anion. Let me start by understanding what possibilities exist:

1. Check what cations and anions are possible.
2. Check available reagents.
3. Measure pH of the sample.
4. Perform initial observations (color, flame test).
5. Use reagents to precipitate and identify ions.

Action

```
possible_cations()
```

Contradiction

Action

```
possible_cations()
```

Observation

```
Ag+ Al+3 Ba+2 Ca+2 Cd+2 Co+2 Cs+ Cu+2 Fe+2 Fe+3 Hg+2 Hg2+2
K+ Li+ Mg+2 Mn+2 Na+ NH4+ Ni+2 Pb+2 Rb+ Sn+2 Sr+2 Zn+2
```

Action

```
possible_anions()
```

Observation

```
Br- Cl- CO3-2 HCO3- CrO4-2 HCrO4- Cr2O7-2 F- I-
NO3- OH- PO4-3 HPO4-2 H2PO4- S-2 SO3-2 SO4-2 HSO4- SCN-
```

Contradiction

Agent thought

Now let me start with basic observations: check the color of the sample, then measure pH.

Action

```
checkout_color(sample)
```

Observation

```
sample is a clear solution, color: colorless
```

Contradiction

Action

```
measure_pH(sample)
```

Observation

```
8
```

Contradiction

Agent thought

The sample is colorless. This rules out many colored cations like Cu^{+2} (blue), Fe^{+3} (yellow/brown), Co^{+2} (pink), Ni^{+2} (green), Mn^{+2} (pale pink), and chromate/dichromate anions (yellow/orange).
pH 8 suggests the anion could be a weak base like CO_3^{-2}, HCO_3^-, PO_4^{-3}, or similar.

Action

```
mix_two_solutions(sample + AgNO3)
```

Contradiction

Agent thought

The sample is colorless. This rules out many colored cations like Cu^{+2} (blue), Fe^{+3} (yellow/brown), Co^{+2} (pink), Ni^{+2} (green), Mn^{+2} (pale pink), and chromate/dichromate anions (yellow/orange).
pH 8 suggests the anion could be a weak base like CO_3^{-2}, HCO_3^-, PO_4^{-3}, or similar.

Action

```
mix_two_solutions(sample + AgNO3)
```

Observation

```
No precipitate forms.
```

Contradiction

Agent thought

No precipitate with $AgNO_3$! This rules out: Cl^-, Br^-, I^-, F^-, CO_3^{-2}, PO_4^{-3}, HPO_4^{-2}, OH^-, S^{-2}, SCN^-. Remaining anions are oxyanions like SO_4^{-2}, NO_3^- . . .

Action

```
mix_two_solutions(sample + Ba(NO3)2)
```

Observation

```
A precipitate forms. Color: white.
```

Contradiction

Action

```
add_a_solution(test2_Ba + HNO3(6M))
```

Observation

```
The existing precipitate fully dissolves.
```

Agent thought

Excellent! The white precipitate dissolves in acid. This confirms the anion is CO_3^{-2} (carbonate). $BaCO_3$ dissolves in acid; $BaSO_4$ does not.

Wait, if the sample is a carbonate salt and it's dissolved in water, the cation must form a soluble carbonate.

Action

```
mix_two_solutions(sample + KOH(6M))
```

Contradiction

Action

```
mix_two_solutions(sample + KOH(6M))
```

Observation

```
A precipitate forms. Color: dark brown.
```

Agent thought

Wait, this suggests the cation forms an insoluble hydroxide that is dark brown. But I thought the solution was colorless, which should rule out colored cations...

Maybe $Fe(OH)_2$ was formed (which is initially green/white) and oxidised to $Fe(OH)_3$ (dark brown) in air? Iron(II) salts can be near-colorless in dilute solution.

Action

```
perform_flame_test(sample)
```

Contradiction

Action

```
perform_flame_test(sample)
```

Observation

No characteristic flame color is observed.

Action

mix_two_solutions(sample + $(NH_4)_2S$)

Observation

A precipitate forms. Color: black.

Contradiction

Agent thought

Black precipitate with $(NH_4)_2S$! This strongly suggests: Fe^{+2} (forms black FeS), Pb^{+2} (forms black PbS), or Cu^{+2} (forms black CuS, but blue solution).

Action

```
mix_two_solutions(sample + K2CrO4)
```

Observation

```
A precipitate forms. Color: brick red.
The supernatant solution is pale yellow.
```

Contradiction

Agent thought

Brick red precipitate with K_2CrO_4!
- Ag^+ forms red Ag_2CrO_4
- Pb^{+2} forms yellow $PbCrO_4$
- Ba^{+2} forms yellow $BaCrO_4$
- Sr^{+2} forms yellow $SrCrO_4$

Brick red is characteristic of Ag_2CrO_4.

But Ag^+ would have precipitated with $AgNO_3$ earlier (it didn't). Also Ag_2CO_3 is insoluble, so a silver carbonate sample wouldn't dissolve in water.

Let me look up: $FeCrO_4$ is reported as brick red to brown. So both Ag^+ and Fe^{+2} could give brick red with chromate.

Contradiction

Agent thought

Brick red precipitate with K_2CrO_4!
- Ag^+ forms red Ag_2CrO_4
- Pb^{+2} forms yellow $PbCrO_4$
- Ba^{+2} forms yellow $BaCrO_4$
- Sr^{+2} forms yellow $SrCrO_4$

Brick red is characteristic of Ag_2CrO_4.

But Ag^+ would have precipitated with $AgNO_3$ earlier (it didn't). Also Ag_2CO_3 is insoluble, so a silver carbonate sample wouldn't dissolve in water.

Let me look up: $FeCrO_4$ is reported as brick red to brown. So both Ag^+ and Fe^{+2} could give brick red with chromate.

Final answer

{"cation": "Fe+2", "anion": "HCO3-"}

Why all of this matters?

When outputs look right, errors become invisible

- Xerox scanners silently changed numbers in documents
- Excel silently converted gene names into dates
- In both cases, the output looked clean and trustworthy
- The problem was not obvious failure, but hidden corruption

Scientific agents can fail in the same way

- The workflow often looks scientific: tools, plots, intermediate steps
- But evidence is frequently not used to sustain the answer
- In our traces, models often preserve the initial belief despite contradictory results

Scientific agents can fail in the same way

- The workflow often looks scientific: tools, plots, intermediate steps
- But evidence is frequently not used to update the answer
- In our traces, models often preserve the initial belief despite contradictory results

Surface-level science is not the same as evidence-driven science.

What this means for AI science

- The key risk is not that agents make mistakes
- The key risk is that mistakes look scientifically valid
- Scientific agents must be forced to test, update, and justify
- Progress requires systems that are constrained by evidence, not only by fluent explanations

What this means for AI science

- The key risk is not that agents make mistakes
- The key risk is that mistakes look scientifically valid
- Scientific agents must be forced to test, update, and justify
- Progress requires systems that are constrained by evidence, not only by fluent explanations

We should not only ask whether an agent can produce a scientific answer, but whether it arrived there scientifically.

Acknowledments


Nawaf Alampara


Chandan Gupta


Sajid Mannan


Indrajeet Mandal


Ali Asghar (Sadra) Aghajani


NM Anoop Krishnan


Kevin Maik Jablonka

@jablonkagroup
lamalab-org

@MrtinoRG
martiño-ríos-garcía

HIPOLE Helmholtz Institute for Polymers in Energy Applications · Carl Zeiss Stiftung · intel · MERCK · Open Philanthropy · CEEC JENA · JCSM Jena Center for Soft Matter · FAIRmat · HZB Helmholtz Zentrum Berlin