ChemCost

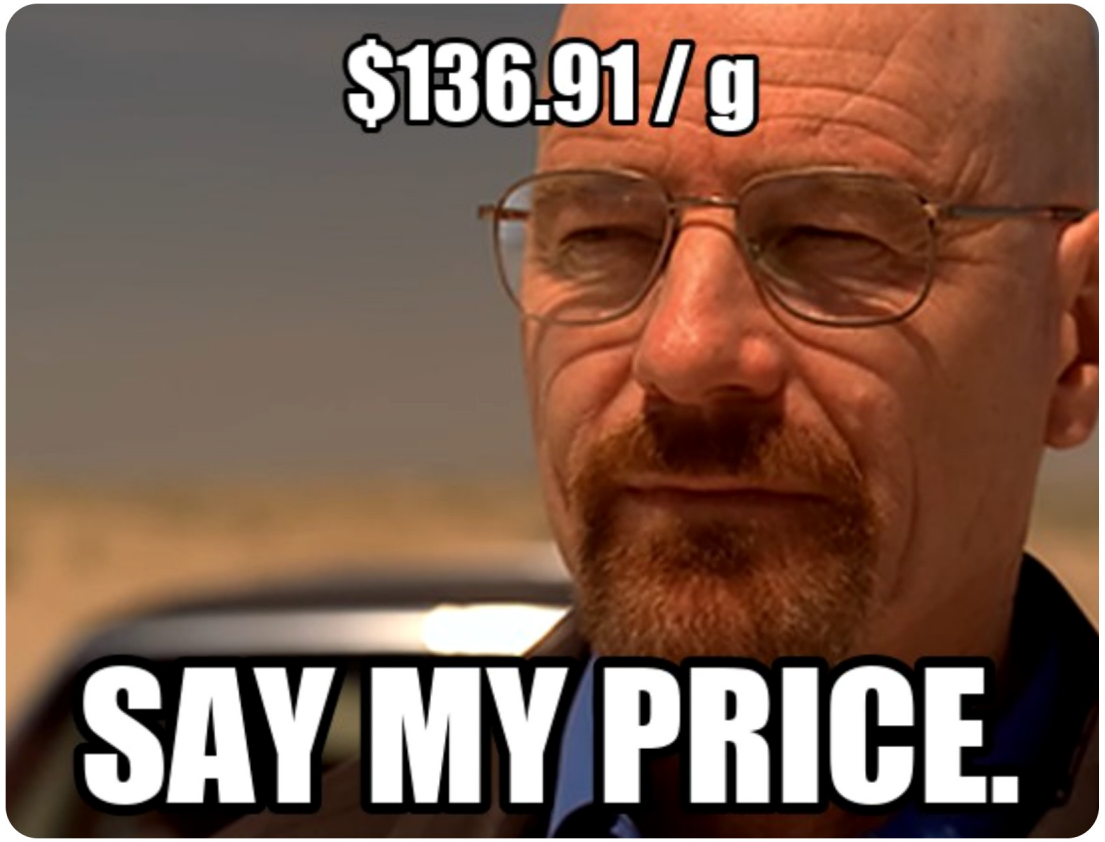
Can Agents Price a Reaction?

Evaluating LLMs on Chemical Cost Reasoning

Yuyang (Younger) Wu*, Yue Huang*, Shuaike Shen, Xujian Wang, Shuhao Zhang, Qiyao Xue, Weichen Liu, Runtian Gao, Jian Ma†, Xiangliang Zhang†, Olexandr Isayev†

CMU · Notre Dame · UNC Chapel Hill · Pittsburgh *Equal contribution †Corresponding

NeurIPS 2026 · Evaluations & Datasets Track · chemcost.org



Read a reaction.
Say its price.

- **Input:** a reaction as a chemist wrote it
- **Output:** $ per gram of product at 1 mmol scale
- **Labels:** exact, no judge

$136.91 / g

CHEMCOST-0001, an amide coupling

Every synthesis starts in the stockroom



- Before running a route, chemists price its materials
- Cost decides route choice, budget, suppliers and feasibility
- Today it is manual catalog work, one reagent at a time
- Can an agent do it from the reaction text alone?

“We need to cook.” First, we need to price it.

One reagent, many packs, many prices



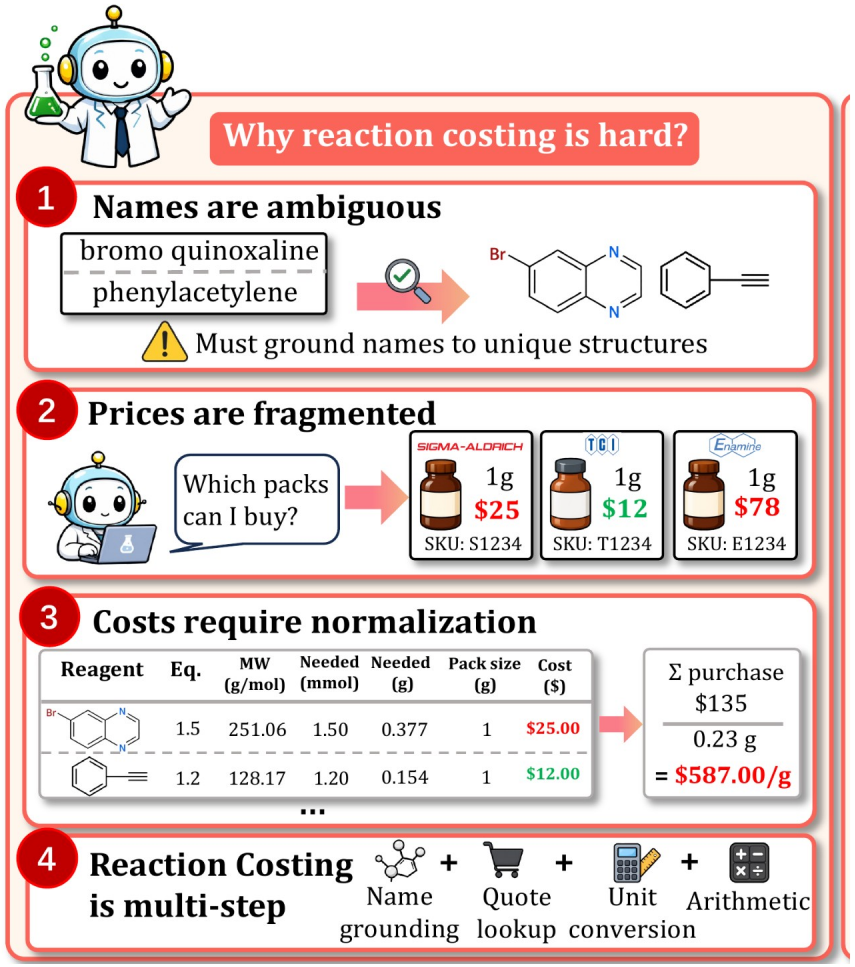




- Each bottle is one pack: supplier, catalog number, size, purity, price
- The cheapest valid pack depends on how much the reaction needs
- ChemCost freezes 230,775 pack-level quotes, so every label is exact

Know your packs, Jesse.

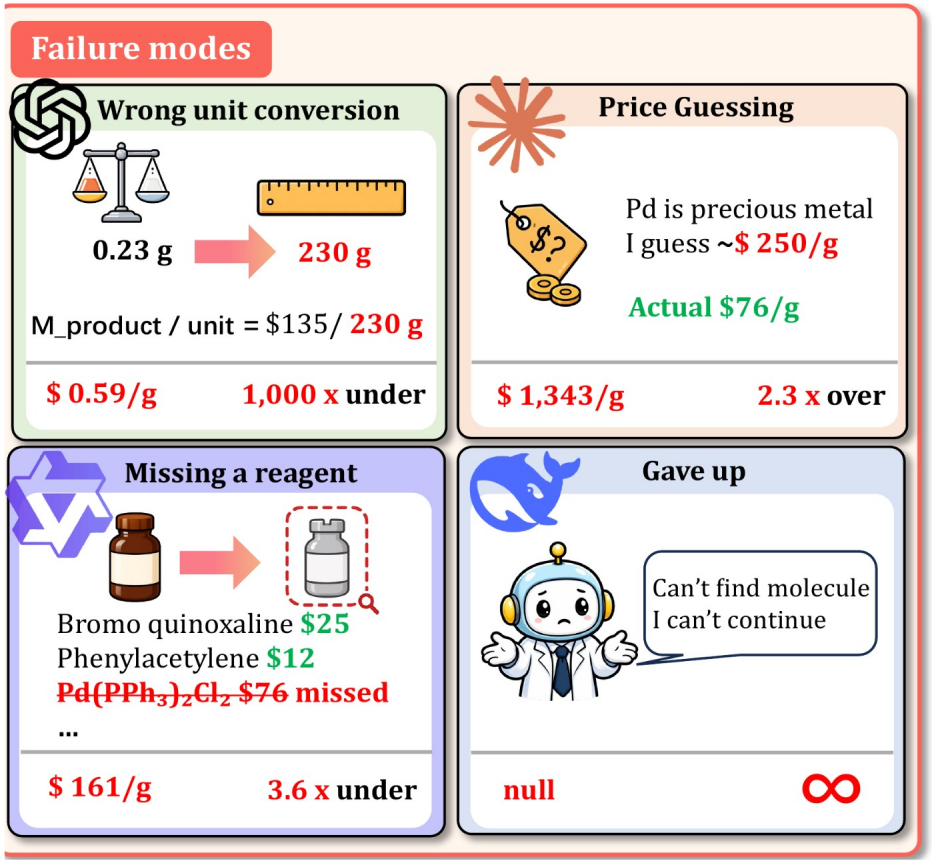
Why is pricing a reaction hard?



- **Names are ambiguous:** each surface form must map to one structure
- **Prices are fragmented:** quotes spread over suppliers, packs and purities
- **Costs need normalization:** equivalents → masses → packs → $/g
- **Costing is multi-step:** one wrong step ruins the number

"Chemistry is the study of change." So is pricing.

Frontier agents fail in four distinct ways



- **1,000× under:** wrong unit conversion, 0.23 g read as 230 g
- **2.3× over:** a price guessed from memory
- **3.6× under:** the Pd catalyst silently missed
- **null:** gave up on an unresolved name

"Jesse, we need to convert the units."

Scientific tool use lacks an exact, judge-free test

- **Tool-use benchmarks**

  SWE-bench, WebArena, ToolLLM: orchestration, little domain reasoning

- **Chemistry agents**

  ChemCrow, Coscientist: judged by demos, experts or LLM judges

- **ChemCost**

  Domain grounding plus multi-step tool use, scored against exact labels

No half measures: exact labels, every stage checked.

The task, made exact



- **Pipeline**

  ground → retrieve → select → aggregate

- **Scale**

  1 mmol limiting reagent, $m_i = e_i \cdot M_i \cdot 10^{-3}$ g

- **Pack rule**

  smallest pack with purity ≥ 95% covering m_i

- **Label**

  $c = \Sigma\, q_i \,/\, (M_p \cdot 10^{-3} \cdot y \,/\, 100)$

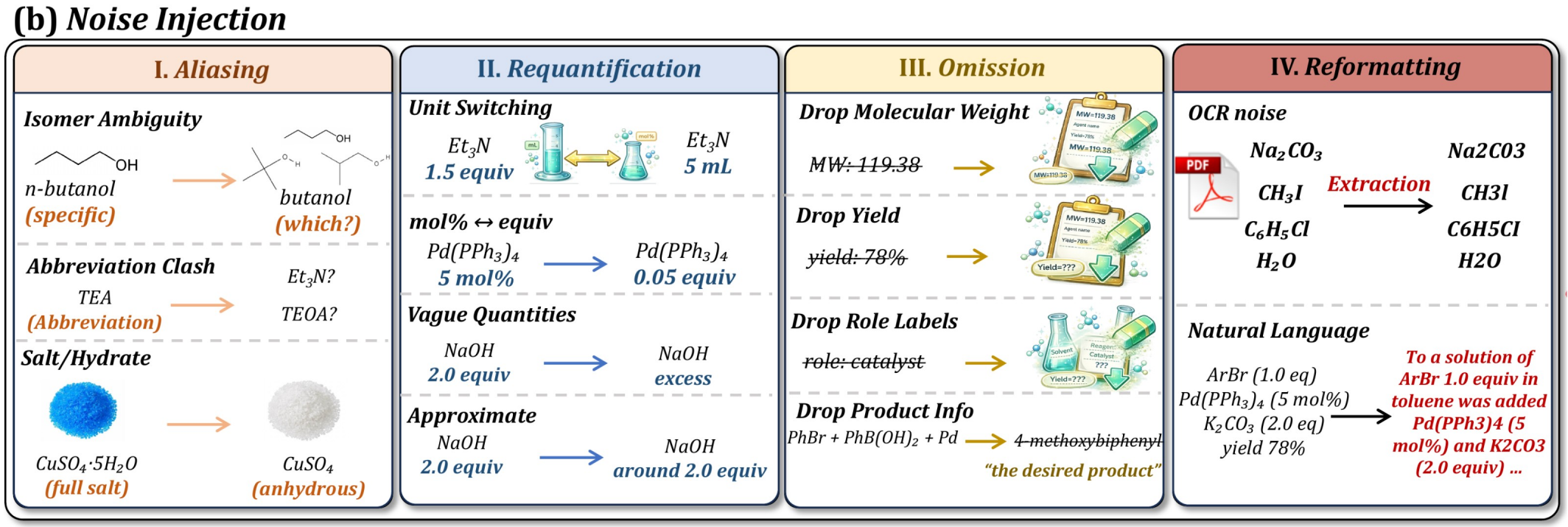
1,427 reactions, priced against a frozen market



- **1,427** evaluable reactions from five source families plus hand-crafted cases
- **2,261** chemicals grounded via PubChem and validated with RDKit
- **230,775** supplier quotes in one frozen snapshot, purity ≥ 95%
- **0** human or LLM judges: every label is deterministic

We froze the market. Respect the chemistry.

Four noise layers, same ground truth



- Aliasing, requantification, omission and reformatting
- Labels never change, so score drops measure robustness

Tread lightly: formatting breaks agents first.

Twelve tool-augmented ReAct agents

- **Closed-source**

  Qwen3.5-Plus, GPT-5, Claude Sonnet 4.6, Gemini 2.5 Flash, GPT-5.4 Mini

- **Open-weight**

  DeepSeek V4 Pro, Kimi K2.5, Qwen3-14B, Qwen3-235B-A22B

- **Chemistry-specialized**

  LlaSMol-7B, ChemDFM-v2.0-14B, ChemLLM-20B

- **Protocol**

  ReAct, 4 tools, 40 steps · CTA@k · 6 input views

Twelve cooks. One price.

Best agents: half the reactions within 25%

Closed-source	CTA@10	CTA@25
Human reference	57.0	66.9
Qwen3.5-Plus	34.6	50.6
GPT-5	24.7	38.3
Claude Sonnet 4.6	24.3	34.6
Gemini 2.5 Flash	21.8	32.2
GPT-5.4 Mini	5.8	11.5

Open & chemistry	CTA@10	CTA@25
DeepSeek V4 Pro	37.0	49.4
Kimi K2.5	24.7	37.0
Qwen3-14B	4.9	9.9
Qwen3-235B-A22B	3.7	6.2
LlaSMol-7B	1.2	1.2
ChemDFM-v2.0-14B	0.0	0.0
ChemLLM-20B	0.0	0.0

- Best CTA@25: **50.6%** (Qwen3.5-Plus); best CTA@10: **37.0%** (DeepSeek V4 Pro)
- Expert reference: **66.9%** CTA@25; chemistry-specialized models score ≈ 0%

Walt's product: 99.1% pure. Best agent: 50.6%.

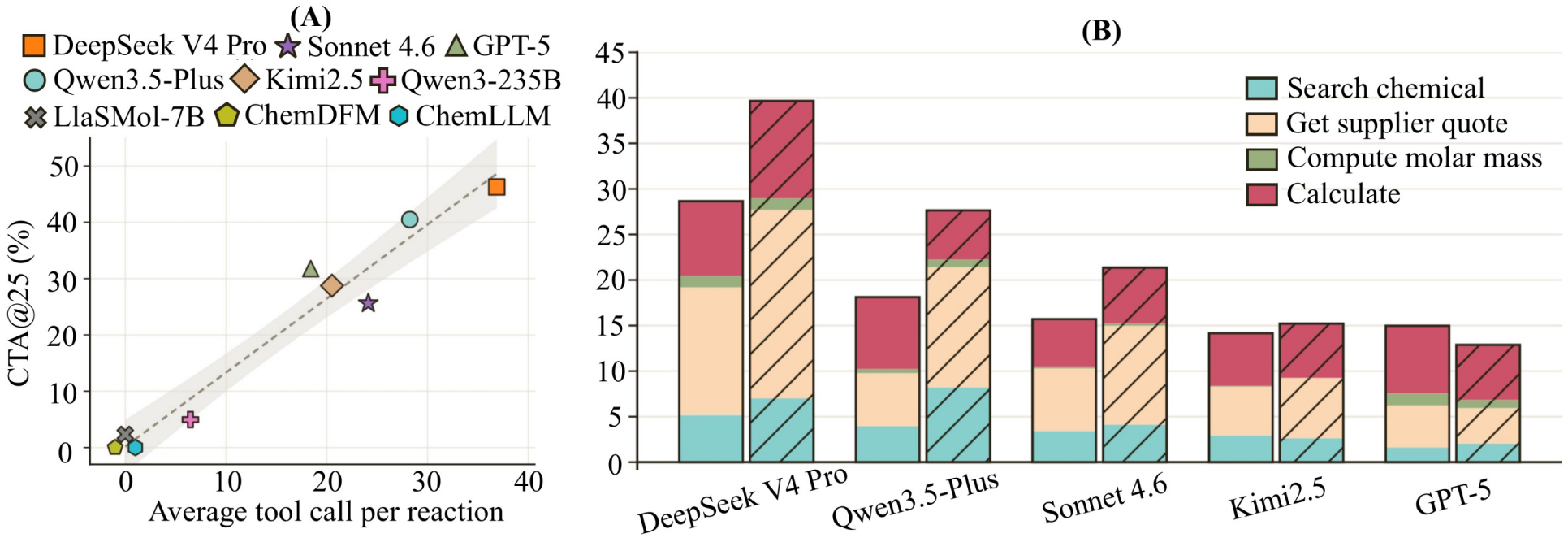
Without tools, no backbone gets past 7.4%

CTA@25 (%)	Sonnet 4.6	DeepSeek V4 Pro	Qwen3.5-Plus
ZeroShot	0.8	—	1.7
FewShot	4.1	2.5	7.4
Chain-of-Thought	2.5	4.1	4.1
ReAct + tools	34.6	49.4	50.6

- CoT lifts Sonnet's recall 8.3% → 52.9%; CTA@25 stays 2.5%
- Qwen ZeroShot: 98.4% recall, 1.7% CTA@25
- With tools: 34.6–50.6%

No tools? Then you're just guessing, Jesse.

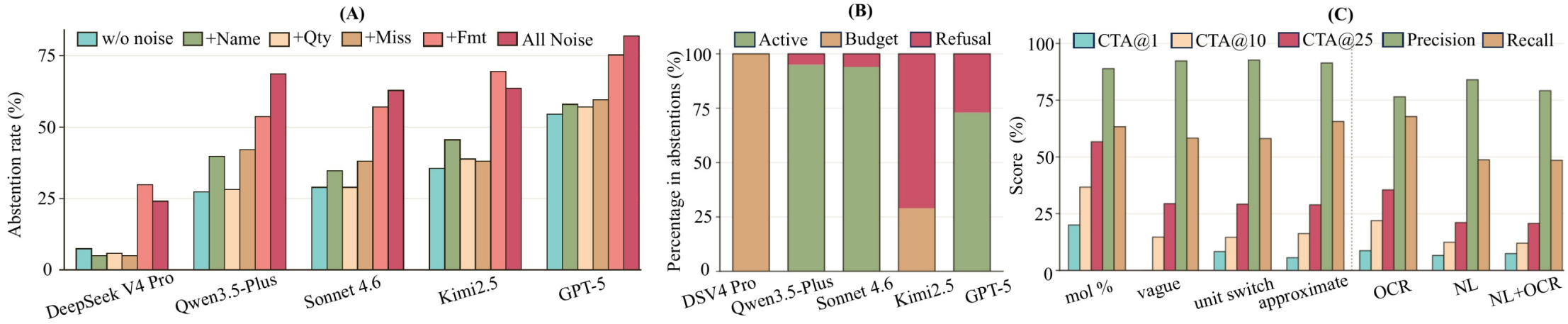
Tool engagement is necessary, not sufficient

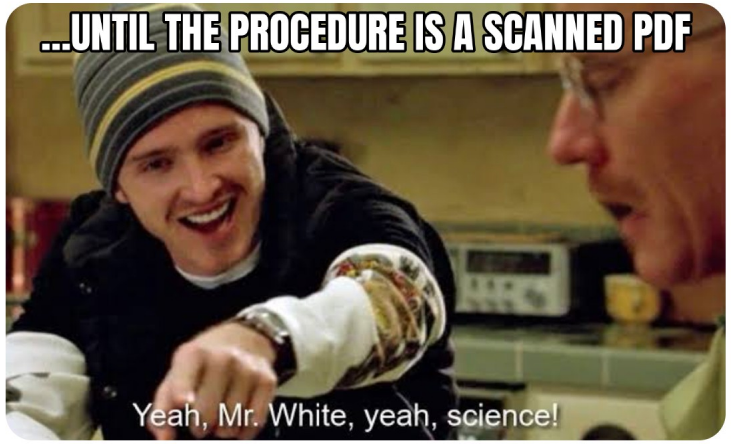


- Tool calls track CTA@25: r = 0.98, or 0.77 without near-zero callers
- Failed runs use 38–52% more calls, stuck in quote retrieval
- GPT-5 fails by stopping early, with ~14% fewer calls

More calls ≠ more product.

Formatting noise: +26.3 pp abstention

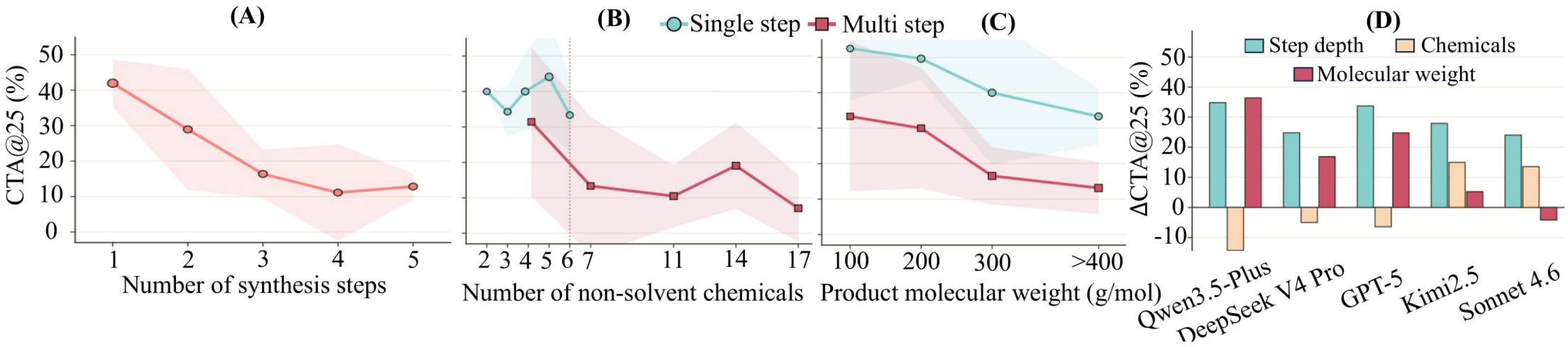




CTA@10	Clean	+Fmt	All
Qwen3.5-Plus	34.6	18.5	19.8
DeepSeek V4 Pro	37.0	22.2	19.8

- Same null, different cause: budget runs out, agents stall, or Kimi refuses

Route depth is the dominant difficulty axis



- CTA@25 by route length: 42% → 29% → 16% → 11% → 13% for 1–5 steps
- Every model loses about 24–25 points from shallow to deep routes
- Component count and product weight matter less, and differ by model

The deeper the route, the worse the cook.

Each agent breaks at a different stage

Pass rate (%)	Ground	Retrieve	Pack	Aggregate
Qwen3.5-Plus	49.6	76.3	18.4	33.8
GPT-5	15.6	67.0	1.8	12.1
Claude Sonnet 4.6	47.2	72.9	58.6	3.1
DeepSeek V4 Pro	85.1	54.3	66.2	21.6
Kimi K2.5	35.0	78.7	48.3	16.9

- DeepSeek grounds 85% of components but often skips retrieval
- Qwen retrieves reliably, then misses the pack rule (18.4%)
- Sonnet picks packs (58.6%) but rarely aggregates right (3.1%)
- GPT-5 has the weakest grounding (15.6%) and stops early

Better call... a calculator.

What ChemCost tells us

- **Tools are necessary, not sufficient**

  Best agent 50.6% CTA@25 vs. 66.9% expert reference

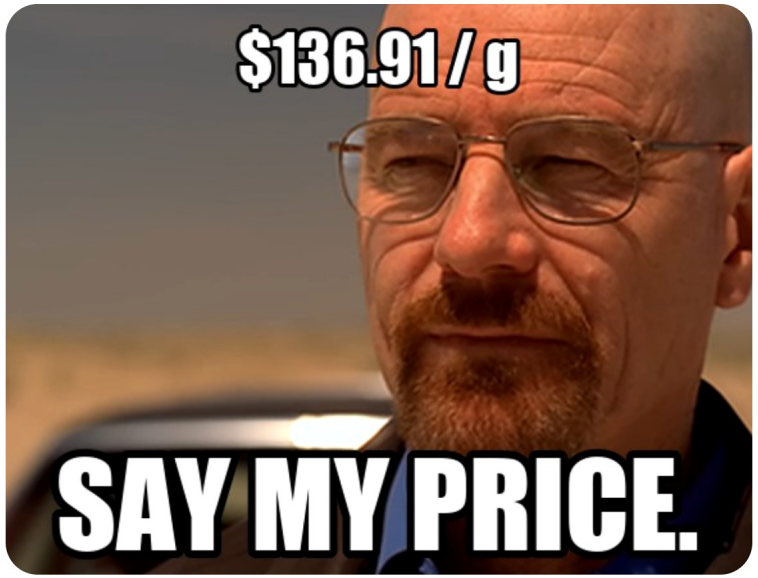
- **Robustness breaks at the input**

  Formatting noise and deep routes hurt every model

- **Fix the weakest stage**

  Chemical grounding, state tracking, calibrated abstention

We're done when the label says we're done.



Thank you

chemcost.org

SCAN ME

Project

chemcost.org

SCAN ME

Paper

arXiv 2605.07251

SCAN ME

Code

GitHub

SCAN ME

Dataset

Hugging Face

Photos via Wikimedia Commons: reagent cabinet by A. L. Carvalho & L. M. Lourenço (CC BY-SA 3.0); cabinets by Reconrabbit (CC0); bottles by D. A. Reid (public domain).