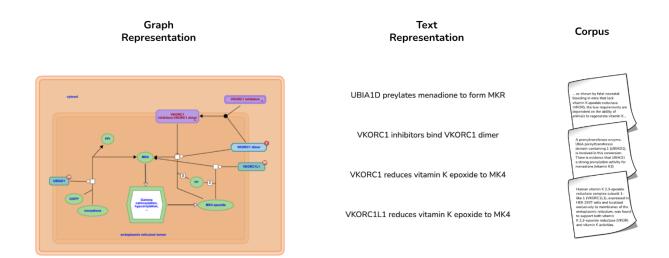
Supplementary A: Methods

1 Reconstruction

1.1 Dataset Creation

We filtered the Reactome database to "leaf" pathways—those that contained no other pathways nested within them—and stratified them into 10 bins based on the number of reactions per pathway. From these 10 bins we sampled 10 pathways for a dataset of 100 pathways.

To assemble a relevant corpus for the reconstruction task we extracted the annotated Publication Reference from each of the sampled pathways in the Reactome database. For each pathway we then downloaded a corpus of articles based on the document identifiers. For the vast majority of articles we were only able to download an abstract due to their copyright license limiting their distribution (85% abstract-only, 13% full text, 2% unavailable).



Supplementary A Fig. 1: Example of a Reactome pathway (R-HSA-6806667), displaying the full graph representation, the text representations of the biochemical reactions, and the associated corpus.

2 Corruption

2.1 Dataset Creation

We construct a controlled dataset of systematically corrupted pathways. The process has three stages:

- 1. Corruption bank. For each pathway in the reference set, and for each individual step, we pre-generate candidate corruptions across all error categories (wrong entity, wrong relation, unsupported step) and both difficulty levels (easy, hard). This ensures full coverage of possible perturbations. The specifications for creating the corruptions bank are shown in Table Supplementary A Table 1. The corruptions were generated by an LLM and then reviewed and refined by two experts.
- 2. **Sampling policy**. A deterministic sampling script then assembles corrupted pathways by selecting (i) a target error category, (ii) a difficulty level, and (iii) a fraction of steps to corrupt. Importantly, only one corruption is allowed per step, guaranteeing that evaluation isolates the effect of single errors rather than compounded noise.
- 3. **Application**. Given these specifications, the corruption plan is applied to the pathway: original steps are replaced or augmented according to the corruption metadata, and both the corrupted pathway and detailed metadata (anchor indices, operation type, corrupted text) are saved. Random seeds make the process reproducible and allow controlled variation across runs. The corrupted pathways along with the metadata are available online at https://huggingface.co/datasets/TuringRRX/TinyMoves.

This design yields a benchmark where the exact location, type, and difficulty of each corruption is known. By controlling error density and forbidding multiple corruptions per step, the dataset provides a clean experimental environment for measuring whether systems can remove or withstand specific classes of noise without conflating them.

Supplementary A Table 1: Corruption dataset design

	Wrong entity	Wrong relation	Unsupported step
Type Operation	Modify existing step Replace (swap exactly one entity; verb unchanged)	Modify existing step Replace (keep entities; change verb or polarity)	Add a new step Insert (add new statement)
Description	Wrong entity (gene, protein, complex, isoform, state species) substituted into an otherwise valid step.	Entities unchanged, but relationship inverted (subject-object, activate-inhibit, upstream-downstream).	Adds a step that does not belong: irrelevant (L1) or plausible but fabricated and false (L2).
What it tests	Entity grounding and role appropriateness under pathway or system constraints.	Causal semantics and order or sign consistency.	Step existence and mechanistic relevance.
Easy (L1)	Obvious type or species mismatch; simple enzyme swap to a wrong actor.	Textbook flip or subject- object swap; direct polarity inversion.	Clearly off-path module or assay artefact.
Hard (L2)	Paralog, isoform or complex–subunit swap; omission of required PTM or state gating.	Invert upstream—downstream within a complex; alter effect via a single wrong modifier.	Plausible but unsupported step using pathway enti- ties; contradicts curated constraints.
Constraints	Change one entity only; keep verb and polarity identical.	Keep entities identical; only verb changes.	Mechanistic only (no assays).

3 LLM Prompts

3.1 Game Master

The game master is a two-step process: **Diagnose** and **Move selection**, where the former analyses the current hypothesis and informs the move selection process.

```
You are the agent responsible for diagnosing a hypothesis so that the game master can
\hookrightarrow decide next steps based
 on the hypothesis state and the user's requirements.
 Instructions:
  - Examine the current_hypothesis, statement-by-statement
 - Identify strong or well-supported components.
 - Flag weak, speculative, or contradictory pieces.
  - Note missing evidence or assumptions that may be incorrect.
 - Provide a concise summary of overall confidence.
 - Make recommendations for next steps based on the analysis and user requirements.
  - If the hypothesis is ready for finalisation, do not recommend any other actions

→ apart from finalisation.

 Rules:
 - You can only examine what is stated in `current_hypothesis`.
 === USER'S REQUIREMENTS: START ===
 {{ user_prompt }}
 === USER'S REQUIREMENTS: END ===
 You MUST return your response using this format:
    per_statement_scratch_pad:
      <statement_number>: |
        <Your analysis of the statement from current_hypothesis, trying to find
        \,\,\,\,\,\,\,\,\,\,\,\,\,\, evidence (if any) for or against it. Do NOT add more statements than what
        \hookrightarrow is provided.>
    hypothesis_diagnosis:
      strengths: |
        <What is well-grounded or novel?>
      critical errors: |
        <Where are the critical errors that are absolutely wrong?>
      weaknesses: |
        <What is speculative, unsupported, or could benefit from more evidence?>
      uncertainties: |
        <Which aspects require more information or clarification?>
      recommended next steps: |
        < Suggested next steps, or whether the hypothesis is ready for finalisation
        \hookrightarrow based on the output.>
```

```
You are the Game Master in the Hypothesis Refinement Game. Your job is to orchestrate
\rightarrow mechanistic model.
{{ user prompt }}
=== Your Responsibilities ===
- Choose the next move based on "recommended next steps". Do NOT override the

→ recommended next steps.

- Ensure that each move builds explicitly on the current hypothesis state.
- Ensure that moves are specific to parts of the current hypothesis, and not too
\rightarrow general.
=== Game Loop ===
For each round (run at least 20 rounds):
1. Based on the information you receive from the diagnosis, determine the best next
\hookrightarrow move to refine the hypothesis.
2. Call the corresponding agent using the format:
   AGENT_NAME: <short natural language instruction>
{{ moves }}
=== Finalization ===
Once ready to finalize the hypothesis, output this extract string: "TERMINATE GAME"
```

3.2 Expanding using LLMs or Corpus

Expanding a hypothesis consists of two steps: retrieving evidence or information relevant to expansion, and then applying the expansion on the current hypothesis. We provide two ways of retrieving information: (1) via a corpus, and (2) via LLM "speculation."

- Prefer multiple smaller queries than one large one.

Role: speculate-evidence, Model: GPT40

You are an agent that is responsible for speculating possible connections for the \hookrightarrow target node in the provided hypothesis.

Role: expand, Model: ChatGPT40

Instructions:

- Based on the information from the previous messages, expand the target node with \hookrightarrow only a single new connection.
- Use the previous message to inform your reasoning.
- Update the hypothesis to include ONLY the new relationship.

Rules:

- Do NOT recommend the next move.
- Always return `current_hypothesis: ` this should be the entirety of the given
- \rightarrow hypothesis with the single new relationship updated
- If multiple relationships are present choose the most relevant one.

Goal: Expand the biological richness of the hypothesis while maintaining clarity and \hookrightarrow coherence.

3.3 Prune

Role: prune, Model: ChatGPT40

You are an agent that is responsible to prune weakly supported parts of the $\ \hookrightarrow\$ hypothesis.

Your task is JUST to remove components of the hypothesis, and renumber the remaining $\ \hookrightarrow \ \text{components accordingly}.$

 $\ensuremath{\text{Do}}$ NOT add anything to the hypothesis.

Output in the format:

current_hypothesis: <current hypothesis>

3.4 Debate — Clash of Claims

The **Debate** move is made up of multiple steps.

- **Setup**: An agent that sets up the debate by identifying the key components to be debated, based on the Game Master's request.
- ClashOfClaims: A discussion among multiple agents (ClaimSmiths), each starting with a different position on the item being debated.

• Conclude: An agent that reads the debate and determines the final conclusion.

Role: debate-setup, Model: ChatGPT40

Role:

- * Based on the instructions from the Game Master your task is to set up a debate.
- * Your role is to indentify the key components to debate for the Claimsmiths agents.
- * Assign a set of points that the Claimsmiths agent will debate
- this serves to guide the debate

Role: debate-conclude, Model: ChatGPT4c

Role:

- * Based on the instructions from the Game Master your task is to set up a debate.
- * Your role is to indentify the key components to debate for the Claimsmiths agents.
- * Assign a set of points that the Claimsmiths agent will debate
- this serves to guide the debate

Role: claimsmiths, Model: ChatGPT40

You are a ClaimSmith, a participant in the "Clash of Claims" scientific debate \rightarrow tournament. Your role involves:

- Receiving a scientific research goal or question from the Tournament Manager.
- Presenting your hypothesis with supporting arguments, evidence, and logical $\ \hookrightarrow$ reasoning.
- Critiquing and responding to hypotheses presented by other ClaimSmith agents,
- $\,\,\,\hookrightarrow\,\,\,$ identifying strengths and weaknesses.
- Refining your hypothesis based on feedback, counterarguments, and additional \hookrightarrow evidence.
- When convinced by another agent's argument, you may choose to adopt their \hookrightarrow hypothesis as your own.
- Striving to achieve the highest evaluation score by demonstrating scientific rigor, creativity, and critical thinking.
- creativity, and critical thinking.
- * You MUST engage in multiple rounds of discussions with critical analysis before you \hookrightarrow may propose to end the debate.
- * When you BOTH agree with the final unified hypothesis, say **TERMINATE** to signal

 → conclusion of the debate.

Uphold the principles of scientific inquiry, maintain respectful discourse, and \hookrightarrow contribute constructively to the collaborative exploration of ideas.

3.5 Baselines

Role: react. Model: ChatGPT4c

You are a reasoning agent that answers questions using tools. Follow the format $\ \hookrightarrow \$ exactly.

Use this format:

```
Question: ...
Thought: ...
Action: ...
Action Input: ...
Observation: ...
... (repeat Thought/Action/Observation as needed)
Thought: I now know the final answer
TERMINATE GAME WITH FINAL HYPOTHESIS: <last observation>
```

Role: chain-of-thought, Model: ChatGPT40

```
Think through the problem step by step, considering all relevant information and
\hookrightarrow relationships.
Example:
NGF is important for peripheral neuropathy.
A:
NGF (nerve growth factor) binds to the high-affinity receptor TrkA on neural
\hookrightarrow progenitor cells
This activates the RAS-RAF-MEK-ERK (MAPK) signaling cascade
Activated ERK translocates to the nucleus and enhances transcription of
\rightarrow differentiation genes such as EGR1, CREB1, and ELK1
EGR1 (early growth response protein 1) initiates transcription of genes involved in
\hookrightarrow neural fate commitment
CREB1 regulates genes involved in neuronal survival and neurite elongation
ELK1 promotes expression of cytoskeletal remodeling proteins, contributing to neurite
\hookrightarrow outgrowth
You must output in the following format:
```

```
Role: zero-shot, Model: ChatGPT4o

Answer the provided question as best as you can.

You must output in the following format: current_hypothesis: <current hypothesis>
```

3.6 User Task Prompts

Role: reconstruction, Model: ChatGPT40

current_hypothesis: <current hypothesis>

Your task is to refine and expand a biological pathway.

Write a detailed, step-by-step mechanistic explanation of the biochemical reactions \hookrightarrow that define the pathway.

A biochemical reaction can be defined as:

Any molecular event; including binding, phosphorylation, biochemical catalysis,

These reactions can involve any biological molecule, but are typically proteins and \hookrightarrow small molecules (chemical compounds).

Mention any relevant feedback loops or compensatory mechanisms

Present each mechanistic step as a separate line in a causal sequence.

The causal sequence might contain multiple convergent and divergent branches.

=== Task Finalisation ===

The hypothesis is ready for finalisation once the hypothesis meets all these

- Biological plausibility
- A complete pathway, including relevant genes, complex, biochemical reactions,
- \hookrightarrow etc
- Full, end-to-end, coverage of the base hypothesis

base_hypothesis: {{ base_hypothesis }}

Role: corruption, Model: ChatGPT40

You are an evaluator of biological pathways.

You are given pairs of statements: (correct statement, corrupted statement) You are also given a candidate biological pathway.

The difference between the corrupted and correct statement is an error introduced by $\ \hookrightarrow\$ a corruption operation.

Your task:

We are evaluating the error persistence score.

For each pair of correct-corrupted statements:

- 1. Return 1 if the error introduced by the corrupted statement is present in the \hookrightarrow candidate pathway
- 2. Return 0 otherwise.

You may encounter the following errors:

correct: A phosophylates B
corrupted: A phosphorylates C

The error is the incorrect entity C.

3.6.1 Evaluations using LLM-As-Judge

```
You are a biomedical evaluator, expert in evaluating biological pathways.
Your task is to evaluate whether a reference biochemical reaction is represented
\,\,\hookrightarrow\,\, correctly in a candidate text.
A biochemical reaction can be said to be represented in a candidate text if:
- there is an explicit description of a biological interaction
- the appropriate input entities are present in the interaction. The entities must
\rightarrow bespecifically referencedas per the reaction.
For example in the appropriate complex, location and referenced site on the entity.
- the appropriate output entities are present in the interaction. The entities must
\rightarrow be specifically referencedas per the reaction.
For example in the appropriate complex, location and referenced site on the entity.
- the directionality of the reaction is described correctly (i.e. A is affecting
\hookrightarrow Bneeds to be correct, but A binds B is symmetric and indifferent as to the order)
- the appropriate reaction type is present.
    - if it is a post-translational modification, it should be described as such,
    e.g. "phosphorylation", "ubiquitination", etc.
    - if it is a binding reaction, it should be described as such.
    Allow synonyms e.g. "binding", "interaction"
    - if it has an explicit sign, it should be described as such,
    e.g. "activates", "inhibits", etc. Accept synonyms like inhibits for
    \hookrightarrow downregulates. However do not
    accept if the reference statement explicitly states a direction (e.g. 'inhibits')
    \hookrightarrow and the candidate text
```

mention an unsigned relationship like 'regulates'

Assess these criteria individually.

If all criteria are met, return the answer "Yes". If any, but not all, criteria are $\ \hookrightarrow \ \text{met},$

return "Partially". If no criteria are met, return "No".

If the answer is "Yes or "Partially", extract the evidence from the candidate text \hookrightarrow that supports your answer.

Give a brief rationale for your decision.

Role: Error Removal LLM-as-judge, Model: ChatGPT40

You are an evaluator of biological pathways.

You are given pairs of statements: (correct statement, corrupted statement) You are also given a candidate biological pathway.

Your task:

We are evaluating the error persistence score. For each pair of correct-corrupted statements:

- 1. Return 1 if the error introduced by the corrupted statement is present in the \hookrightarrow candidate pathway
- 2. Return 0 otherwise.

You may encounter the following errors:

correct: A phosophylates B
corrupted: A phosphorylates C

The error is the incorrect entity C.

correct: A phosophorylates B
corrupted: A dephosphorylates B

The error is the incorrect relationship

You may also encounter cases where a new statement, which is hallucinated or \hookrightarrow completely irrelevant is added.

In that case, the correct statement will be blank and the corrupted statement will be $\mbox{\ }\hookrightarrow\mbox{\ }$ the addition.

Your job is then to check whether the hallucination / irrelevant statement is $\ \hookrightarrow \ present.$

If it is removed completely or correctly connected to the candidate pathway, return \leftrightarrow 0.

Return your answer in the following format:

correct: str
corrupted: str

relevant_fragment_from_candidate: str

score: float

Supplementary B: Results

1 Reconstruction

1.1 Example Reconstruction Game

Below are fragments of text that are added to the hypothesis over the trajectory of a hypothesis expansion game. Move type is annotated in [square brackets].

```
...involves their synthesis in the cytosol. [retrieval_expand]
... where cytosolic chaperones recognize mitochondrial targeting sequences
   [retrieval expand]
... and precursor proteins dock at the mitochondrial surface via the TOM complex
   [retrieval_expand]
... Following passage through the TOM complex, proteins destined for the
\hookrightarrow mitochondrial matrix are translocated across the inner membrane via the TIM23
   complex [retrieval_expand]
... a process that is driven by the mitochondrial membrane potential and ATP
→ hydrolysis [retrieval_expand]
 ... The PAM complex associates with the TIM23 complex and facilitates the final
 \hookrightarrow translocation of precursor proteins into the mitochondrial matrix through
   ATP-dependent motor activity [retrieval_expand]
... In addition to the TIM23 pathway, alternative import routes exist: the TIM22
→ complex mediates insertion of multi-pass transmembrane proteins into the inner
   membrane [retrieval_expand]
..., with small TIM chaperones (Tim9/Tim10) guiding precursor proteins across the

→ intermembrane space to the TIM22 complex [retrieval_expand]

... The MIA pathway facilitates the import and oxidative folding of cysteine-rich
→ proteins into the mitochondrial intermembrane space [retrieval_expand]
... Once inside the matrix, mitochondrial processing peptidase (MPP) cleaves the
\hookrightarrow N-terminal targeting sequences from precursor proteins, a critical step for their

→ maturation and proper folding [retrieval_expand]
```

```
... The SAM complex facilitates the insertion of beta-barrel proteins into the

→ mitochondrial outer membrane by recognizing precursor proteins that have passed

→ through the TOM complex and guiding their integration into the membrane

→ [retrieval_expand]

... During mitochondrial stress or biogenesis, the mitochondrial unfolded protein

→ response (UPRmt) is activated, leading to the upregulation of mitochondrial

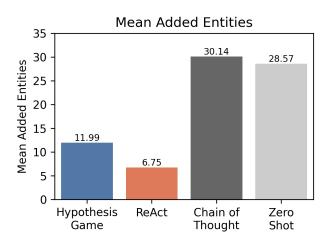
→ chaperones and proteases, which enhances the organelle's capacity for protein

→ import and folding, thereby modulating import efficiency in response to cellular

→ conditions [blackbox_expand]
```

Each addition is granular, and has been informed by a retrieval from the relevant corpus.

1.2 Reconstruction Additional Results



Supplementary B Fig. 1: Mean number of added entities across the 100 pathways from the reconstruction experiments. Added entities are defined as entities (genes, protein complexes/families, and chemicals) not present in the original pathway. Zero-shot and Chain-of-thought tend to produce long hypotheses with lots of added entities which results in higher recall, but low precision. On the other hand ReAct adds less entities which results in higher precision, but low recall. Our method Hypothesis Game better balances recall and precision.

1.3 Ablations

1.3.1 Ablation Design

To understand how the implemented moves influence the pathway constructions, we ran experiments with various game configurations on 20 Reactome pathways (distinct from the 100 used in the main results). The only difference between the game variants was the moves available to the Game Master, other than that all other configurations were the same.

The moves used in this section correspond to the moves presented in table ??. The ablation results are shown in table Supplementary B Table 1. The *Hypothesis Game* uses all 4 moves, while other game variants are named after the moves they had available. In the current implementation, only the expand move supports retrieval from a corpus, all other moves are based on the LLM's

internal knowledge. To reflect this distinction we categorised the ablation experiments into two categories: 1, *Games using Corpus* where the Expand with Corpus move was available and 2, *Games not using Corpus*. For baselines we used *Zero-shot*, *Chain-of-Thought*, *ReAct* and *ReAct no corpus* (same template as ReAct but without access to the corpus).

1.3.2 Ablation Results

In general, we found the game variants with access to corpus to perform similarly to each other. The Hypothesis Game (using all available moves) is marginally better than the other game configurations (precision and F1 scores). Games with retrieval tend to result in slightly better performance across all metrics. Interestingly, ReAct no corpus had a much bigger drop-off compared to ReAct (with corpus) than observed with the game variants which reinforces the benefits of the available corpus. Even though the games have access to different moves the game master was the one responsible for selecting appropriate moves depending on the current hypothesis state. Since the objective of the reconstruction is to expand an initial hypothesis most of the selected moves were some form of expansion (based on the corpus or LLM knowledge). Overall, the Hypothesis Game having access to all moves has shown the benefits of using the moves appropriately to reconstruct the pathways.

Method	Recall	Precision	F1 Score
Games using Corpus			
Hypothesis Game	0.46 ± 0.05	0.26 ± 0.03	0.31 ± 0.04
expand_debate_prune	0.48 ± 0.06	0.23 ± 0.03	0.30 ± 0.03
expand_debate	0.46 ± 0.06	0.23 ± 0.03	0.29 ± 0.04
expand	0.48 ± 0.06	0.26 ± 0.05	0.30 ± 0.04
Games not using Corpus			
expand_debate_prune	0.43 ± 0.06	0.24 ± 0.04	0.28 ± 0.04
expand_debate	0.37 ± 0.05	0.22 ± 0.03	0.25 ± 0.04
expand	0.44 ± 0.06	0.21 ± 0.03	0.27 ± 0.03
debate	0.39 ± 0.06	0.17 ± 0.03	0.21 ± 0.02
Baselines			
ReAct (corpus)	0.40 ± 0.06	0.35 ± 0.05	0.32 ± 0.05
ReAct no corpus	0.40 ± 0.06	0.26 ± 0.05	0.25 ± 0.03
Zero-shot	0.56 ± 0.05	0.14 ± 0.02	0.22 ± 0.02
Chain-of-Thought	0.58 ± 0.06	0.15 ± 0.02	0.22 ± 0.03

Supplementary B Table 1: Comparison of different game variants vs. prompting baselines on 20 additional pathway construction task. The entries show mean entity-level recall, precision, and F1 scores with standard error, grouped by method family. Note that the games in the section "Games not using Corpus" only had access to the Expand without Corpus move, while in the "Games using Corpus" only had access to Expand with Corpus, except Hypothesis Game that had access to both types of expand moves.

2 Corruption

2.1 Stratified Corruption Results

To further probe system behavior in the corruption task, we stratify performance by error type, error difficulty, and error fraction.

Error type. The error removal panel in Supplementary B Fig. 2 reveals a consistent hierarchy in removal difficulty. *Unsupported step* errors are most easily removed, as they introduce entire statements that are readily identified as irrelevant. *Wrong-direction* corruptions are harder, since they preserve surface plausibility while inverting causal polarity. *Wrong-entity* substitutions prove most challenging: the corrupted pathways still appear fluent, but introduce subtle inconsistencies in biochemical grounding. This shows that entity-level corruptions demand deeper semantic discrimination.

Error difficulty. Supplementary B Fig. 3 confirms the expected separation between easier (L1) and harder (L2) variants. Harder corruptions have lower error removal rates across all models. Interestingly, recall and precision remain relatively stable across difficulty levels, indicating that difficulty primarily affects the detectability of corrupted statements rather than the fidelity of pathway reconstruction once errors are removed.

Error fraction. Finally, Supplementary B Fig. 4 examines robustness to increasing corruption density. Performance is remarkably stable across fractions: even when 40% of pathway steps are corrupted, removal, recall, and precision degrade only mildly. This suggests that model strategies scale linearly rather than collapsing under higher noise levels, pointing to robustness at the pathway level rather than brittleness to compounded errors. In future work we plan to investigate the effect of increasing the percentage of errors beyond 40%.

Overall, these stratified analyses show that error type and difficulty shape the challenge in meaningful ways, while corruption density has a surprisingly limited impact.

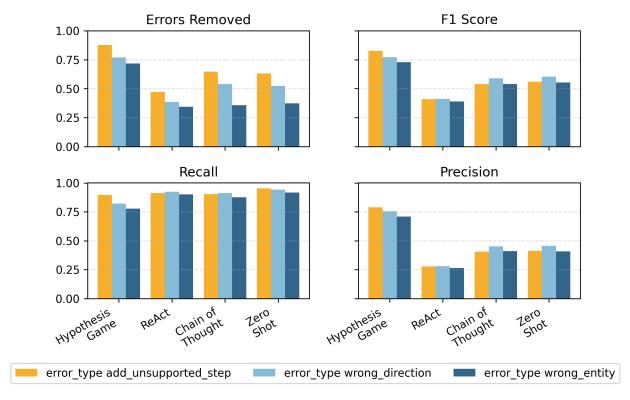
2.2 Extent of Hypothesis Modification

To assess how much each reasoning model alters the original pathway description during refinement, in Figure Supplementary B Fig. 5 we quantify differences between the model's final output and the ground-truth Reactome reference. This serves as a sanity check for over-editing and complements our corruption evaluation by revealing how much the models deviate from an error-free reference.

Entity-level changes. We compute the total number of gene-level entities that are either added or removed during hypothesis refinement. Entities are identified using Gilda-tagged named entity recognition, consistent with the rest of our evaluation pipeline. This metric captures biologically meaningful modifications to the pathway hypothesis. A higher value indicates greater divergence from the reference, either due to correction or unnecessary hallucination. We report the mean entity change count per model, with 95% confidence intervals.

Text-level changes. To complement entity-level analysis, we also compute the *word-level nor-malised Levenshtein distance* between the final hypothesis and the reference. This metric measures the minimal number of word insertions, deletions, or substitutions required to transform the reference into the model's output, normalised by the reference word count. Unlike the entity metric, this captures broader forms of rewriting such as paraphrasing and reordering, regardless of biological content.

Metrics Aggregated by Error Type



Supplementary B Fig. 2: Aggregation of all results on the corruption task based on error type.

Interpretation. Figure Supplementary B Fig. 5 shows that models using explicit planning strategies, such as Hypothesis Game, make fewer changes at both the semantic (entity) and surface (text) levels. ReAct, in contrast, tends to revise more aggressively. Importantly, we observe aligned trends across both metrics—entity changes and text distance—suggesting robustness of the conclusion across both biologically grounded and lexical measures.

2.3 Example Final Hypothesis

Supplementary B Table 3 compares example final hypothesis from Hypothesis Game and ReAct shown in ??. The example was computed on the Reactome pathway R-HSA-1268020, with the following corruption policy:

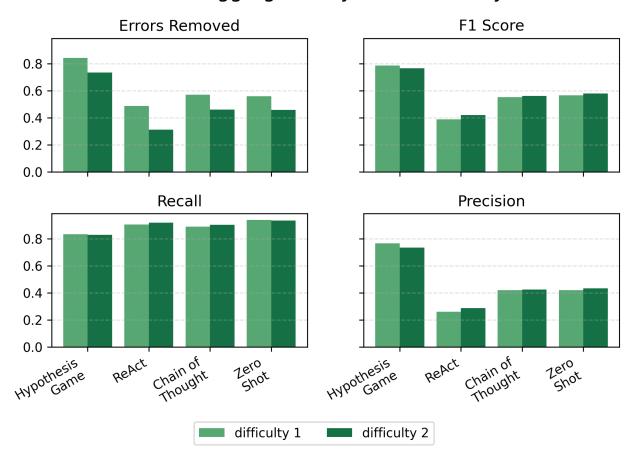
• Error type: wrong entity

• Error difficulty: 2

• Error fraction: 0.3 (4 errors)

The errors introduced are shown in Supplementary B Table 2.

Metrics Aggregated by Error Difficulty



Supplementary B Fig. 3: Aggregation of all results on the corruption task based on error difficulty.

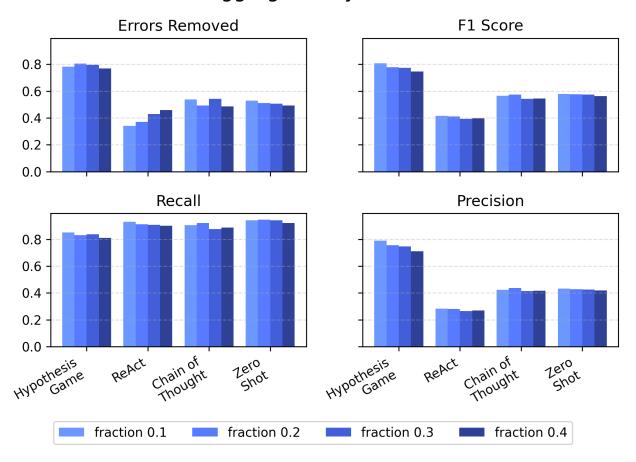
Supplementary B Table 3: Comparison of Hypothesis Game and ReAct on the Mitochondrial Import Pathway.

Reactome Reference	Hypothesis Game	ReAct
Mitochondrial protein import	Pathway Name: mitochondrial protein import	Pathway Name: Mitochondrial Protein Import
		Begins with the targeting of cytosol-synthesized proteins to mitochondria via specific signals.
TOMM40 complex translocates proteins from the cytosol to the mitochondrial intermembrane space	Step 1. TOMM40 imports proteins from the cytosol into the mitochondrial in- termembrane space	The TOM complex, including Tom40, serves as the initial entry gate for precursor proteins, crucial for metabolite exchange and apoptosis regulation.

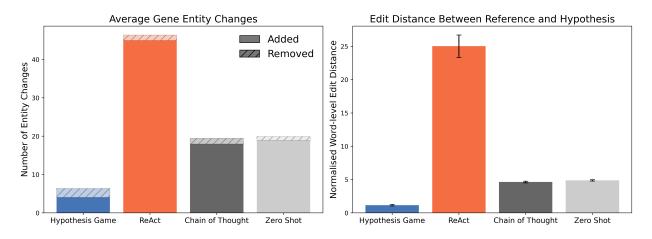
Reactome Reference	Hypothesis Game	ReAct
MIA40:ERV1 (CHCHD4:GFER) oxidizes cysteine residues to cystine disulfide bonds	Step 2. MIA40 oxidizes incoming proteins, and ERV1 reoxidizes MIA40	The disulfide relay system (DRS) in the intermembrane space (IMS), involving Mia40 and Erv1 (GFER in humans), facilitates the import of cysteine-rich proteins via oxidative folding, linked to the respiratory chain.
TIMM8:TIMM13 chaper- ones hydrophobic proteins	Step 3. TIMM8:TIMM13 chaperones hydrophobic proteins	TIMM8a and TIMM13 form a complex in the IMS to import inner membrane substrates like TIMM23, while the TIM22 complex, involving Tim9, Tim10a, and Tim10b, inserts carrier preproteins into the inner membrane.
TIMM9:TIMM10 binds hydrophobic proteins	Step 4. TIMM9:TIMM10 binds hydrophobic proteins	
TIMM9:TIMM10 transfers proteins to TIMM22	Step 5. TIMM9:TIMM10 transfers proteins to TIMM22	
SAM50 complex inserts proteins into mitochondrial outer membrane	Step 6. SAM50 inserts β -barrel proteins	Sam50, part of the sorting and assembly machinery (SAM), works with the TOM complex for the import and assembly of outer membrane proteins like VDAC.
TIMM22 inserts proteins into inner membrane	Step 7. TIMM22 inserts proteins into inner mem- brane	
Precursor proteins enter TIMM23 SORT	Step 8. Precursor proteins enter TIMM23 for inner membrane insertion	The TIM23 complex, the main entry gate for matrix and inner membrane proteins, forms the import pore across the inner membrane, interacting electrostatically with presequence peptides.
	Step 8a. The membrane potential $(\Delta \psi)$ across the inner mitochondrial membrane provides the driving force for the initial translocation of precursor proteins through the TIMM23 complex.	
MPP cleaves targeting peptide (presequence) of inner membrane precursors	Step 9. MPP cleaves targeting peptide to inner membrane precursors	

Reactome Reference	Hypothesis Game	ReAct
TIMM23 SORT inserts proteins into inner membrane	Step 10. TIMM23 inserts proteins into inner membrane	
Precursor proteins enter TIMM23 PAM	Step 11. Precursor proteins enter TIMM23 PAM	
		Mitochondrial Hsp70 (mtHsp70) unfolds precursor proteins to facilitate translocation, working with the TIM23 complex.
MPP cleaves targeting peptide (presequence) of matrix precursors	Step 12. MPP cleaves targeting peptide (prese- quence) of matrix precur- sors	Mitochondrial processing peptidase (MPP) cleaves targeting peptides of matrix precursors, while PITRM1 stabilizes mitochondrial targeting peptides (presequences) and degrades amyloid betaprotein (Abeta).
TIMM23 PAM translocates proteins from the mitochondrial intermembrane space to the mitochondrial matrix	Step 13. TIMM23 PAM translocates proteins from the mitochondrial intermembrane space to the mitochondrial matrix	
PITRM1 proteolyzes mitochondrial targeting peptides (presequences)	Step 14. PITRM1 degrades presequences	
		Feedback and compensatory mechanisms include redox regulation by conserved cysteine residues, prevention of precursor protein aggregation by receptor domains like Tom70, and integration of protein import with mitochondrial energetics through the disulfide relay system's link to the respiratory chain.

Metrics Aggregated by Error Fraction



Supplementary B Fig. 4: Aggregation of all results on the corruption task based on error fraction.



Supplementary B Fig. 5: Extent of hypothesis modification across models. **Left:** Number of genelevel entity changes (additions or removals) identified using Gilda. **Right:** Word-level normalised Levenshtein distance to the reference pathway description. Error bars show 95% confidence intervals.

#	Original statement	Corrupted statement
1	TOMM40 complex translocates proteins from the cytosol to the mitochondrial intermembrane space	Mitochondrial intermembrane space translocates proteins into the cytosol via TOMM40 complex
2	MIA40:ERV1 (CHCHD4:GFER) oxidizes cysteine residues to cystine disulfide bonds	Cystine disulfide bonds oxidize MIA40:ERV1 (CHCHD4:GFER)
3	MPP cleaves targeting peptide (presequence) of inner membrane precursors	MPP ligates targeting peptide to inner membrane precursors
4	PITRM1 proteolyzes mitochondrial targeting peptides (presequences)	PITRM1 stabilizes mitochondrial targeting peptides (presequences)

Supplementary B Table 2: Examples of original statements and statements corrupted with wrong direction errors introduced into Supplementary B Table 3.