**Supplementary Text 2: LLM-Assisted Annotation Pipeline**

EnrichGT implements a sophisticated large language model (LLM)-assisted annotation pipeline designed to generate biologically coherent interpretations of enrichment analysis results while systematically mitigating hallucination risks through multi-layered validation strategies. We position this system as a hypothesis generation tool rather than an authoritative source of biological knowledge, requiring all LLM-generated interpretations to undergo expert review and literature validation—consistent with standard practices in computational biology.

**1.Model Specifications and Implementation**

All primary analyses presented in this manuscript utilized Deepseek V3 (0324 checkpoint), with GPT-4.1 serving as an auxiliary validation model throughout development and testing phases. The temperature parameter was consistently set to 0.6 across all experiments to balance creative interpretation with reproducible consistency. This setting was selected after preliminary optimization to maximize diverse biological insights while maintaining factual accuracy and avoiding overly deterministic responses.

Our multi-model comparison functionality was systematically developed and validated using three distinct LLM architectures: Deepseek V3, OpenAI GPT-4.1, and Google Gemini 2.5 Pro. While the methodological framework is intentionally model-agnostic to ensure broad applicability, we recommend using recent large-parameter foundation models rather than task-specialized variants, as the former demonstrate superior performance in complex biological reasoning tasks. The software package provides users with full control over temperature settings (values below 1.0) and model selection to optimize results for specific use cases.

**2.Structured Pre-Processing Through Pathway Clustering**

Central to our hallucination mitigation strategy is the mandatory transformation of raw enrichment results into structured, biologically coherent clusters before any LLM processing. We construct document-term matrices (DTMs) where each enriched pathway serves as a "document" and constituent genes function as "terms." For Over-Representation Analysis (ORA) results, we use hit genes (overlapping genes from the input list); for Gene Set Enrichment Analysis (GSEA), we utilize core enrichment genes that contribute most significantly to the enrichment score.

Pairwise similarities between pathways are quantified through cosine distance calculations on their vector representations, followed by hierarchical clustering with user-selectable linkage methods (complete, average, or single linkage). When analyzing results from multiple pathway databases (GO, KEGG, Reactome), we concatenate their respective DTMs to enable joint clustering, allowing identification of functional similarities spanning different annotation systems.

This pre-clustering step is fundamental to hallucination mitigation for four critical reasons: (1) it anchors all subsequent LLM analysis in curated pathway databases with established biological validity, (2) it dramatically reduces input complexity and noise that often trigger hallucinations, (3) it provides structured context that guides models toward biologically plausible interpretations, and (4) it enables LLMs to focus on established functional relationships rather than speculating about unknown gene interactions. This approach fundamentally differs from naive applications that present LLMs with overwhelming gene lists or verbose enrichment tables, which consistently produce poor results due to context limitations and lack of biological metadata.

**3.Prompt Engineering and Reproducibility Framework**

We developed standardized, rigorously tested prompt templates for three sequential annotation stages to ensure complete reproducibility across different users, datasets, and model implementations. These templates underwent iterative refinement to optimize biological accuracy and consistency:

***Stage 1 - Pathway Theme Summarization:***

Here is a list of reorganized biological themes derived from enrichment

analysis. As a professional biologist, please concisely summarize the

overarching theme of these pathways, their biological implications, and

possible interpretations. Provide evidence where necessary. Please avoid

misleading and over-claiming replies. Use an academic writing style and

limit the response to ~250 words. Output plain text only (no markdown).

The included pathways are: [PATHWAY\_LIST]

***Stage 2 - Gene-Level Functional Annotation:***

Here is a set of genes that may share similar biological functions.

Previous analysis suggests their potential functions may include

[BIOFUNS]. As a professional biologist, please help summarize their

potential biological roles, highlight which ones may be most noteworthy,

and provide your insights. Please avoid misleading and over-claiming

replies. Write in a single paragraph (~250 words) without bullet points.

Output plain text only (no markdown). The genes include: [GENE\_LIST]

***Stage 3 - Theme Title Generation:***

You are a biologist. I will provide you with a piece of biological text.

[PRERES] Please help me generate a title for the biological topic. Only

return the title itself—choose the single best (and only one) title. Do

not return any other content.

Each template incorporates explicit instructions for the model to: (1) adopt a "professional biologist" persona, (2) implement strict word limits, (3) avoid over-claiming or speculation, (4) maintain academic writing standards, and (5) produce plain text output without formatting. These design choices ensure consistent, professional-grade outputs across different model implementations.

**4.Multi-Layered Hallucination Mitigation Strategies**

Our approach implements several interconnected validation strategies that significantly reduce the probability of factually incorrect outputs:

**Sequential Context Building:** The multi-stage prompting strategy prevents information overload that frequently causes LLMs to generate speculative or contradictory statements. By first establishing high-level pathway themes (Stage 1) before analyzing specific genes (Stage 2), we create a progressive context-building framework where each stage validates and constrains the next. This hierarchical approach ensures that gene-level interpretations remain grounded in pathway-level biological context established from curated databases, significantly reducing the likelihood of biologically implausible predictions.

**Multi-Model Cross-Validation:** Recognizing that individual models possess distinct training biases and knowledge gaps, we implemented a systematic cross-validation framework (via R function “egt\_llm\_multi\_summary”, for example, “egt egt\_llm\_multi\_summary(result, list(`GPT4.1` = chat, `DeepSeekV3` = chat2))”) enabling users to compare outputs across multiple LLM architectures simultaneously. This approach allows researchers to: (1) identify biological insights with strong inter-model consensus, indicating higher confidence in their validity, (2) detect model-specific interpretations or potential hallucinations requiring additional verification, (3) leverage complementary knowledge strengths across different model training datasets, and (4) maintain scientific rigor by treating LLM outputs as hypothesis-generating tools rather than definitive biological conclusions.

Constraint-Based Output Control: The explicit instruction to "avoid misleading and over-claiming replies" in each prompt template, combined with strict word limits, prevents verbose speculation that often accompanies LLM hallucinations. The requirement for academic writing style and plain text output further constrains model responses toward factual, evidence-based statements. Additionally, the progressive validation through our three-stage approach means that significant factual errors in early stages are likely to be detected through inconsistencies in subsequent analyses.

**5.Quality Control and User Warnings**

The EnrichGT software package generates additional warning messages when LLM functions are activated, emphasizing that all interpretations produced by LLMs must undergo expert review and validation against literature. This practice aligns with standard procedures commonly followed in computational biology and reinforces the positioning of this system as a sophisticated hypothesis generation tool rather than an authoritative knowledge source.

This comprehensive validation framework ensures that LLM outputs are systematically grounded in established biological knowledge from curated databases and cross-validated across multiple models, providing a scientifically rigorous foundation for biological hypothesis generation while minimizing hallucination risks inherent to large language model applications.