Supplementary Material 2

GeneResolver: A Traceable Hybrid Multi-Agent Pipeline for Etiology-Aware Gene Prioritization

1 Etiology-classification dataset construction

The etiology-classification dataset was built in three stages over an author-assembled disease seed list. Because its construction involved author-defined rules and LLM-assisted screening, vignette generation, and mechanism verification, it is treated as a silver-standard dataset. Exact prompts are reproduced below.

1.1 Stage 0: author-assembled disease seeds

A fixed disease list was assembled by the authors through online searches for diseases described in public literature and disease resources as having single-gene, chromosomal, somatic, multifactorial, or non-genetic mechanisms.

1.2 Stage 1: Full-text case retrieval

For each seed disease, PubMed was queried through NCBI E-utilities using:

```
"{disease_name}"[Title] AND
"Case Reports"[Publication Type]
AND ("Adult"[MeSH Terms] OR
"Middle Aged"[MeSH Terms] OR "Aged"[MeSH Terms])
NOT ("Infant"[MeSH Terms] OR "Child"[MeSH Terms]
OR "Adolescent"[MeSH Terms])
```

Candidate PMIDs were resolved to PMC identifiers. Full text was fetched from the NCBI BioC endpoint, with Europe PMC full-text XML as a fallback. The final build used full-text-only mode, so abstract-only candidates were rejected before LLM screening.

One LLM call per candidate applied the following screening prompt:

```
Check this case report against two requirements for
target_disease. source_content_level tells you whether
you're reading the full article ("full_text") or only the
abstract ("abstract_only") - treat abstract_only
judgments as lower confidence, since age and primary-vs-
incidental status are often clearer in the full case
presentation.

1. is_adult: the patient is an adult (18+) at the time of
   the reported presentation. If age isn't explicitly
```

```
   stated, infer conservatively from clear contextual cues
   (occupation, "elderly", specific decade of life, etc.),
   if genuinely unclear, set is_adult to false rather
   than guessing.
2. is_primary_diagnosis: target_disease is THE main
   condition this case report is about, not a secondary/
   incidental finding, past medical history item, or
   comorbidity in a report primarily about something else.
   A report titled/focused on a different condition that
   merely mentions the patient also has target_disease
   does NOT qualify.
3. is_incidental_findings: target_disease is not merely an
   incidental finding in the report.

Return a JSON object with: is_adult; age_reported (string,
e.g. "34 years" or "elderly" or "not stated");
is_primary_diagnosis; confidence (low, moderate, or high);
patient_key (a short identifying label for this patient,
e.g. "58F with..."); and exclusion_reason (empty string if
both checks pass).
```

1.3 Stage 2: fact extraction and naratives generation

One LLM call per accepted case extracted patient facts and rewrote the report as a concise, first-presentation narrative. It excluded the final diagnosis, treatment, and confirmatory results:

```
From this case report about a patient with
target_disease, do two things in one response:

1. Extract patient facts using only what's reported:
   patient_key (short identifying label, e.g. "37M
   with..."), presentation_facts, family_history,
   exposure_history, prediagnostic_findings.
2. Write a concise first-presentation clinical case from
   those facts only. First-presentation is a first doctor
   visit and runned tests, the treatment, final testing and
   differentials are not established yet. Final diagnosis
   and tests result for the final diagnosis should not be
   included in the vignette.

Return a JSON object with: patient_key;
presentation_facts; family_history; exposure_history;
prediagnostic_findings; and vignette (a single plain-text
string, not a nested object).
```

1.4 Stage 3: patient-specific mechanism verification

The disease-level target class remained the default label, but a second call checked the article's explanation for the target disease in the individual patient. The prompt explicitly prevented a downstream complication or incidental trigger from replacing the mechanism of the seed disease:

```
Read the case report and find final explanation for
why THIS patient has target_disease specifically, not a
separate complication, or an incidental trigger, that
merely occurred in an already-diagnosed patient.
expected_mechanism is target_disease's established cause;
keep it unless the article's own final explanation for
target_disease itself specifically differs. Pick ONE label:

single_gene (one gene's own variant or small indel,
including a deletion confined to that one gene, e.g. one
exon of it); chromosomal (an aneuploidy, or a deletion/
duplication/rearrangement spanning multiple genes or a
chromosome segment, even if inherited); somatic
(postzygotic, acquired in only some of the patient's
cells/tissues, not inherited); multifactorial (polygenic
and/or an established combination of genetic predisposition
with an environmental/infectious/autoimmune trigger);
non_genetic (infectious, toxic/drug, nutritional,
mechanical, autoimmune, or other acquired cause, not rooted
in an inherited or postzygotic DNA change).

Return JSON: label; somatic_subgroup ("single_gene" or
"multifactorial" - only if label is "somatic", else "");
reason (explain the specific mechanism/finding that
justifies this label, one or two sentences).
```

This patient-specific verification corrected 41 of 220 labels (18.6%) relative to the disease-level default.

The complete 220-case distribution was 69 single-gene, 28 chromosomal, 14 somatic, 57 non-genetic, and 52 multifactorial.