

**Initial Search**  
Query combining terms from LLMs and Genetics/Medicine

Records identified: 57,558, from which:  
PubMed: 34,923  
arXiv: 5,972  
bioRxiv: 11,213  
medRxiv: 5,450

Identification

Screening

Eligibility

Included

**Automated duplicates merging**

Records after duplicates removed: 51,613

**Primary Semantic Analysis**

Perform frequency analysis, identify key-phrases, filter relevant articles

**Key phrases:**

**Group 1:** LLM-related terms (LLM, GPT, BERT, transformer, etc.)

**Group 2:** Clinical terms (clinical, syndrome, complex trait, etc.)

**Group 3:** Genetics-related terms (inherited, GWAS, genome, etc.)

Records after initial filtering: 576

Screening

**Manual duplicates merging**

Records after duplicates removed: 550

**Initial Filtering**

**Classify to:**

- **irrelevant** (not related to transformers and/or diagnostics, genetics),
- **partially relevant** (it does not use LLM/transfomers OR area is not genetics and diagnostics)
- **relevant** (research/tool on the application of the transformers and diagnostics and genetics)

Result: 298 (184 partially relevant articles, 114 fully relevant articles)

**Manual Addition**

Manually add several relevant articles from other sources (27):

- **partially relevant:** 19
- **fully relevant:** 8

**Excluded**

Irrelevant articles: 251

**Final Categorization**

Final set (325):

- partially relevant: 203
- fully relevant: 122

**Usage in the analysis**

195 articles were used for analysis (73 **partially relevant** and 122 **relevant**):

- **Introduction:** 14
- **Knowledge Navigation:** 34
- **Clinical Data Analysis:** 43
- **Genetic Data Analysis:** 32
  - **Communication:** 28
- **Related Research Areas:** 44
- **Discussion (techniques, data, and bias):** 28

The categories and subcategories may overlap.