Retrieve to Explain: Evidence-driven Predictions with Language Models

Anonymous Authors[1]

Abstract

Language models hold incredible promise for enabling scientific discovery by synthesizing massive research corpora. Many complex scientific research questions have multiple plausible answers, each supported by evidence of varying strength. However, existing language models lack the capability to quantitatively and faithfully compare answer plausibility in terms of supporting evidence. To address this issue, we introduce Retrieve to Explain (R2E), a retrieval-based language model. R2E scores and ranks all possible answers to a research question based on evidence retrieved from a document corpus. The architecture represents each answer only in terms of its supporting evidence, with the answer itself masked. This allows us to extend feature attribution methods, such as Shapley values, to transparently attribute each answer's score back to its supporting evidence at inference time. The architecture also allows R2E to incorporate new evidence without retraining, including non-textual data modalities templated into natural language. We assess on the challenging task of drug target identification from scientific literature, a human-in-the-loop process where failures are extremely costly and explainability is paramount. When predicting whether drug targets will subsequently be confirmed as efficacious in clinical trials, R2E not only matches non-explainable literature-based models but also surpasses a genetics-based target identification approach used throughout the pharmaceutical industry.

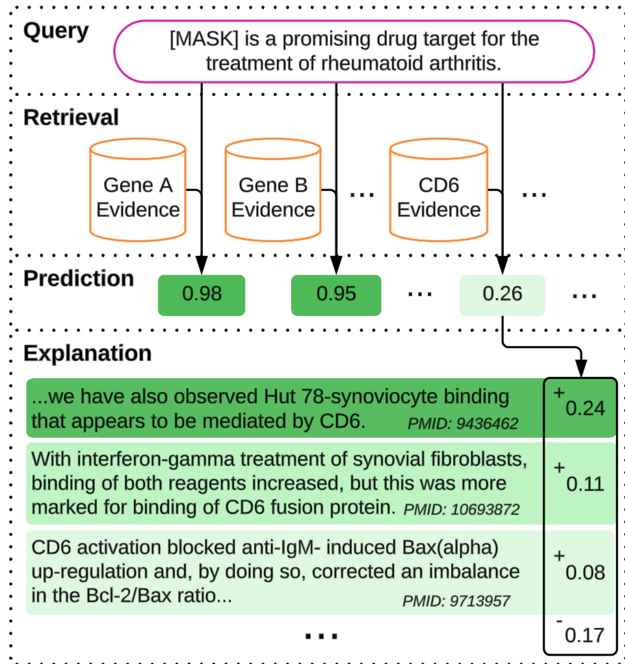

Figure 1. **R2E drug target identification example**. R2E makes predictions based on retrieved evidence and provides explanations in terms of the evidence. **Query**: User queries are phrased in cloze-style, where [MASK] can be filled from a set of potential answers (named entities). For target identification, answers are the set of protein-coding genes (potential drug targets), and the query specifies a disease. **Retrieval**: R2E retrieves the evidence most relevant to the query for each potential answer, where evidence here is taken from across the biomedical literature that mentions the specific answer. **Prediction**: The model scores each answer based on the supporting evidence. **Explanation**: Each answer score is directly and quantitatively attributed to its retrieved evidence using Shapley values. Here, the best evidence is indirect, based on the role of CD6 in mechanisms central to rheumatoid arthritis pathology.

1. Introduction

Language models can act as knowledge bases, supplying answers to factual user queries using only the learned parameters (Petroni et al., 2019; Brayne et al., 2022). They can also be provided with access to searchable knowledge bases to enable a retrieval-augmented approach to question answering (Chen et al., 2017; Lewis et al., 2020; Izacard & Grave, 2021).

Beyond answering factual queries, a searchable knowledge base could provide evidence for queries without known an-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review at ICML 2024 AI for Science workshop. Do not distribute.

swers, including scientific research questions (e.g. *What are some promising drug targets to treat rheumatoid arthritis?*). By proposing new hypotheses supported by both direct and indirect scientific evidence, AI models could facilitate the scientific discovery process (Paliwal et al., 2020; Aliper et al., 2023; Sourati & Evans, 2023).

For high-stakes settings where acting on model hypotheses is costly or risky, an explainable model can mitigate risk by allowing a human expert to inspect the evidence and reasoning behind each prediction before acting on it (i.e. a human-in-the-loop setup). Explainability can also help to identify model flaws or systemic biases, leading to improved performance and task alignment (Kulesza et al., 2015).

Here, we introduce Retrieve to Explain (R2E), an approach for language model prediction with faithful and quantitative explanations (Figure 1). Given a cloze-style user query, R2E first retrieves the most relevant evidence from an evidence corpus, partitioned according to each possible answer. We consider a set of answers comprised of named entities. The model then scores each answer based on its supporting evidence to generate a ranked list. The R2E architecture represents potential answers explicitly in terms of their supporting evidence. In particular, the feature space is the evidence itself, enabling explainability with feature attribution methods to infer the contribution of each piece of evidence to the prediction. Here, we use Shapley values (Shapley et al., 1953; Lundberg & Lee, 2017). In addition to explainability, we show that this evidence-oriented approach allows model predictions to be updated without retraining by modifying the corpus, such as introducing new evidence. Since R2E can generate a score for every answer in the answer set, it is particularly applicable in human-in-the-loop scenarios where many potential hypotheses are prioritized for user review.

With half of drugs failing to show efficacy when tested in human populations (Wong et al., 2019), we focus here on drug target identification, a critical and challenging early stage in drug discovery where specific genes or proteins (targets) are selected as the focus for developing treatments. For this task, we train R2E to score protein-coding genes by relevance to a user query based on a scientific literature corpus. We then augment the corpus with genetic associations by templating them into natural language, allowing the model to use both evidence sources. We show that Shapley values on individual pieces of evidence correlate with large language model (LLM) relevance assessments, which similarly correlate with human experts. Notably, when used to predict clinical trial outcomes, R2E significantly outperforms both genetics evidence, a widely recognised predictor in the pharmaceutical industry (Nelson et al., 2015; Trajanoska et al., 2023), and a few-shot chain-of-thought prompted GPT-4 baseline with retrieval-augmentation, a setup that in practice would also be prohibitively costly and sacrifices faithful explainability. R2E outperforms the genetics baseline even when supplied only with genetics evidence, suggesting that representing gene-trait associations in natural language improves generalization over a structured ontology. Finally, we show that R2E's explainability enables the use of LLMs to audit prediction reasoning, further improving performance.

Alongside the clinical trial outcomes, we evaluate the model on two additional target identification benchmarks and make all three new benchmarks publicly available (Appendix A).

We include an Impact Statement in Section 6. Our core contributions are as follows:

- We introduce R2E, a novel architecture for retrieval-based high-stakes question answering, which scores the plausibility of each answer directly in terms of its supporting evidence, and thereby enables faithful, quantitative explainability using evidence-level Shapley values.
- We evaluate on the challenging scientific discovery problem of drug target identification, showing that R2E is not only as predictive of clinical trial outcomes as non-explainable literature-based baselines, but also surpasses a genetics approach used throughout the pharmaceutical industry.
- We release three new benchmarks to address the lack of publicly-available datasets for target identification and to drive progress on this important problem.

2. Related work

2.1. Language Models with Retrieval

Many language models leverage retrieved text at inference time for question answering (Khandelwal et al., 2019; Karpukhin et al., 2020; Guu et al., 2020; Lewis et al., 2020; Lee et al., 2020; Izacard & Grave, 2021; Borgeaud et al., 2022; Izacard et al., 2022). R2E differs from these existing approaches by (1) scoring all possible answers in a given answer set and (2) faithfully and quantitatively attributing each answer's score to evidence passages using Shapley values. This approach follows from the application: R2E is designed for answering research questions that merit deep user engagement (e.g. identifying drug targets for a disease) as opposed to typical factual recall tasks (e.g. identifying a country's capital city). There is therefore a premium on generating multiple possible answers with explanations, to allow a human to investigate them.

R2E perhaps bears the most resemblance to kNN-LM (Khandelwal et al., 2019) which uses retrieval to improve next-token prediction. However, kNN-LM uses retrieval to augment a standard masked language model, while R2E is fully retrieval-based to enable evidence-driven explanations. The

Fusion-in-Decoder (FiD) approach (Izacard & Grave, 2021) also bears a resemblance to R2E; both merge each piece of evidence with the query independently before jointly processing. FiD is motivated by efficiency and performance. We are additionally motivated by explainability. Achieving faithfully explainable multi-label prediction with existing generative LLM architectures is largely infeasible, as discussed in depth in Appendix V.

2.2. Explainability & Data Attribution

R2E is inspired by SHAP (SHapley Additive exPlanations) (Lundberg & Lee, 2017), which explains model predictions by approximating feature-level Shapley values (Shapley et al., 1953). R2E extends feature attribution methods like SHAP to data, by using a retrieval-based architecture in which the feature space is comprised of evidence. R2E therefore also contrasts with explainability-focused training data attribution (TDA) methods (Hammoudeh & Lowd, 2024), such as representer point selection (Sui et al., 2021), which evaluates the impact of training examples on predictions. Instead, R2E uses the evidence in the corpus at inference time for both prediction and explanation. Among TDA methods, Data Shapley (Ghorbani & Zou, 2019) also assigns Shapley values to data. Data Shapley focuses on explaining model performance rather than inference-time predictions.

SimplEx (Crabbé et al., 2021) explains predictions by approximating a classified input in terms of a corpus of classified exemplars. SimplEx is general-purpose but indirect: the corpus illuminates black-box model predictions, but does not impact them. In contrast, the corpus drives model predictions in R2E.

2.3. Models for Hypothesis Generation

The use of models in generating or evaluating scientific hypotheses is an emerging area of research. Knowledge graphs (KGs) are a popular approach for novel hypothesis generation, because their structure enables multi-hop inference between unconnected nodes. Novel hypotheses have been generated by subject-area experts directly querying and inspecting a KG (Smith et al., 2021).

Sourati & Evans (2023) use KG patterns for material property prediction and drug re-purposing, additionally leveraging nodes for specific researchers to infer which discoveries are more or less likely to be discovered based on social dynamics. Paliwal et al. (2020) used tensor factorization on a biomedical KG to predict future research findings and clinical trial outcomes for therapeutic drug targets. Aliper et al. (2023) similarly employed a biomedical KG to predict clinical trial outcomes; they used a graph transformer network ensembled with a tabular model leveraging clinical trial design features. R2E differs from these approaches by enabling explainability of predictions from the evidence and operating directly on published research without needing to construct a KG.

In this vein, Tshitoyan et al. (2019) work with a materials science research corpus to identify new material properties. They use cosine similarity on unsupervised word embeddings, specifically word2vec (Mikolov et al., 2013). This approach resembles our parametric masked language model baseline, except that in our case embeddings are derived using a transformer. Tshitoyan et al. suggest that word2vec enables indirect inference similar to that in a KG; for instance, a material that has never been defined as thermoelectric may be mentioned alongside properties associated with thermoelectricity. We observe a similar phenomenon in R2E: for instance, a target never identified directly with a disease may still have been shown to regulate disease-relevant mechanisms (Figure 1) or to be genetically associated with relevant traits (Appendix U.7). R2E can use these indirect findings as support.

3. Methods

We consider the problem of scoring N potential answers $\mathcal{A} = \{a_i\}_{i=1}^{N}$ to a user query q such that they can be ranked from most to least relevant. To align with the training corpus (Section 3.1), we let q be a cloze-style statement (e.g. *[MASK] is a promising drug target for the treatment of osteoporosis.*), where each answer a_i represents a potential named entity for the mask location. Lewis et al. (2019) provides an approach to translate between cloze- and natural-style questions.

3.1. Masked Entity-Linked Corpus

Our approach uses a training corpus of textual passages, $\mathcal{D}$, each containing at least one named entity from the set of answer entities $\mathcal{A}$. Entity linking identifies and grounds entities in $\mathcal{A}$ in the corpus. For each passage, the span of every occurrence of a single entity is replaced by a [MASK] token. When the passage contains multiple unique entities in $\mathcal{A}$, we duplicate the passage with each masked in turn while the others appear as plain text. Each example is therefore a tuple (a, d) consisting of an answer entity identifier $a \in \mathcal{A}$ and a masked text passage $d \in \mathcal{D}$ in which that entity occurs.

In our specific use case of drug target identification, $\mathcal{A}$ consisted of 19,176 protein-coding gene entities, hereafter referred to collectively as *Genes*. Surface forms of the gene and corresponding protein (encoded by the gene), were grounded to the same entity under an assumption of 1:1 correspondence. $\mathcal{D}$ was a large-scale entity-linked corpus of 160 million sentences from scientific literature. For more details on the corpus and splits used, including temporal splits to avoid leakage, see Appendix B; for details on the

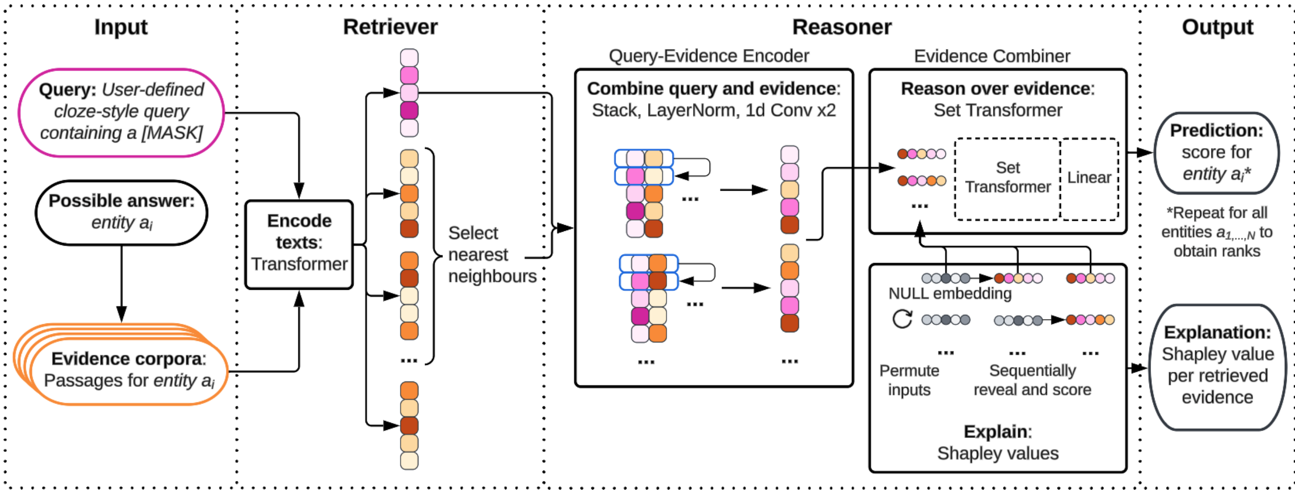

Figure 2. **R2E architecture schematic**. Illustration of R2E inference and explanation. **Input**: A user-defined cloze-style query, a possible answer (named entity) to evaluate, and a corpus of evidence passages corresponding to that answer entity with entity mentions replaced with [MASK]. **Retriever**: The query text is encoded with a transformer. All of the entity's evidence passages are encoded prior to inference, using the same encoder, and stored in a FAISS search index. The k evidence passages with highest cosine similarity to the query are retrieved. **Reasoner**: Each evidence embedding is stacked with the query embedding. The resulting query-evidence pairs are layer-normalised before each pair is combined at corresponding dimensions into a single embedding using convolutional layers. All combined pair embeddings are passed to a set transformer, followed by a linear layer and sigmoid to obtain the binary probability. Shapley values for each pair (corresponding to each piece of evidence) can be computed to quantitatively explain the prediction. **Output**: To rank a set of answer entities $a_{1\ldots N}$, binary probabilities are obtained independently for each. Shapley values attribute model predictions back to the evidence passages providing an explanation of the model's prediction.

entity linking method, see Appendix C. Note that $\mathcal{D}$ could support other use cases (e.g. biomarker identification, drug repurposing, biological mechanism selection) by adjusting $\mathcal{A}$.

3.2. Masked Language Model (MLM)

We first consider a parametric approach based on the pre-training method in Brayne et al. (2022). We use an encoder-only transformer (Vaswani et al., 2017), specifically a scaled-down version of PubMedBERT (Gu et al., 2021). For query passages $d^q \in \mathcal{D}^{\text{MLM}} \subset \mathcal{D}$ containing a masked answer $a_i \in \mathcal{A}$, we train to predict $p(a_i|d^q)$.

The query embedding is the mean over output embeddings corresponding to [MASK] tokens. We take the dot product with a learned embedding for each possible answer $a_i \in \mathcal{A}$, then apply a bias and softmax to predict $p(a_i|d^q)\,\forall\, a_i \in \mathcal{A}$. We train with a cross-entropy loss. Pre-trained weights for domain-specific models are available, including PubMedBERT, but we train from scratch to avoid leakage from pre-training data in our temporally split evaluations.

This model is both a baseline (MLM) and the basis for the Retriever component of R2E (Section 3.3).

3.3. R2E Retriever

We now consider our semi-parametric approach, R2E, which leverages retrieval from an evidence corpus. R2E combines a Retriever module and a Reasoner module (Figure 2). See Appendix D for additional details of the R2E architecture, training and inference hyperparameters.

The MLM in Section 3.2 produces text embeddings that are trained to have a high inner-product with the paired answer embeddings in the answer set. We reasoned that two text embeddings would therefore have high similarity if they permit a similar distribution over answers, i.e. if they were semantically similar with respect to this task. This makes the MLM well-suited to identifying corpus passages that are relevant to the user query and so we used this MLM as the R2E Retriever.

We first used the MLM to embed each of the masked evidence passages in the evidence corpus $\mathcal{D}^e$, where $\mathcal{D}^e = \mathcal{D}^{\text{MLM}}$ for Reasoner training (Section 3.4; typically $D^e = \mathcal{D}$ at inference). We partitioned evidence embeddings according to the masked answer entity they contained, and created distinct FAISS search indices (Johnson et al., 2019) for each to enable efficient answer-specific retrieval.

At inference time, the user's cloze-style query q is encoded with the MLM. The Retriever selects the k evidence passages $[d^e_{1i}, ..., d^e_{ki}] \subset \mathcal{D}^e$ with the highest cosine similarity to q from each answer a_i's search index. For our experiments, $k = 64$. The query embedding and retrieved evidence embeddings for each answer are then the inputs to the Reasoner.

3.4. R2E Reasoner

Training Objective We train the Reasoner with a binary cross entropy loss to differentiate positive examples ($L = 1$) from negative examples ($L = 0$), i.e. to learn $p(L = 1|a_i, d^q)$ when taking an entity a_i and masked query d^q as input, where $d^q \in \mathcal{D}^q \subset \mathcal{D} \setminus \mathcal{D}^e$ ($\mathcal{D}^e$ excluded to avoid trivial inference by retrieving d^q from $\mathcal{D}^e$). Positive examples were constructed from pairs $(a_p, d^q) \forall d^q \in \mathcal{D}^q$, where a_p is the true masked answer in d^q. For each positive example, a corresponding negative example (a_n, d^q) was constructed by uniformly sampling $a_n \in \mathcal{A} \backslash \{a_p\}$. For each (a_i, d^q) pair, positive or negative, the Reasoner receives the top k evidence passages $[d^e_{1i}, ..., d^e_{ki}]$ fetched by the Retriever from the retrieval corpus of a_i. For negatives, due to the sampling of a_n, retrieved evidence corresponds to a different entity to the answer entity masked in the query. Under this negative sampling scheme, the objective $p(L = 1|a_i, d^q)$ is closely related to the MLM multiclass objective $p(a_i|d^q)$ at optimality (Appendix F); however, unlike multinomial regression, sampling negatives avoids needing to retrieve evidence for every possible alternative answer for each example during training.

Inference At inference time, we use $p(L = 1|a_i, q) \forall a_i \in \mathcal{A}$ to score and rank the full answer set for the cloze-style query q, using the evidence fetched for q. This requires $|\mathcal{A}|$ nearest neighbour searches and forward passes through the Reasoner; however, since retrieval and reasoning for each answer is independent of all other answers, the process can be fully parallelized subject to computational resources. See Appendix H for profiling of inference speeds.

Architecture The R2E Reasoner architecture is shown in Figure 2. First, the *query-evidence encoder* $f : \mathbb{R}^h \times \mathbb{R}^h \to \mathbb{R}^h$ combines each of the k evidence embeddings with the query independently. It stacks the evidence with the query to generate a tensor of size $[2, h]$; it then compresses the tensor into a vector of size $[1, h]$ using convolutional layers. The convolutional layers have a filter size of $[2, 1]$ across each embedding dimension h, encoding the relationship between the query and evidence in each dimension.

Next, the *evidence combiner* $g : (\mathbb{R}^h)^k \to [0, 1]$ generates $p(L = 1|a_i, q)$ from the k query-evidence embeddings. There is no inherent ordering among the k vectors, so we use a set transformer (Lee et al., 2019). Since the Reasoner sees the answer a_i only indirectly via the evidence embeddings, we can also view the score as the probability that the query and evidence embeddings are discussing the same entity.

3.5. R2E Explanations

R2E provides explanations in the form of Shapley values (Shapley et al., 1953; Lundberg & Lee, 2017) - the average expected marginal contribution of each piece of evidence to the overall model score for the query. Shapley values enable attribution of the model prediction back to pieces of retrieved evidence, such that they sum up to the overall score.

Multiple methods exist for rapidly approximating Shapley values on deep learning features (Lundberg & Lee, 2017). Defining each of the k inputs to the evidence combiner as a distinct feature gives a relatively small feature space. As a result, it is tractable to use a simple permutation sampling approach to approximate Shapley values. See Appendix E for the full algorithm and Appendix H for profiling.

During training, we replaced query-evidence features at random with a learned NULL embedding. In addition to acting as a regularizer (akin to dropout), introducing the NULL embedding during training ensured that the model could handle missing features robustly when estimating Shapley values. For each training example, the evidence dropout rate was sampled in Uniform$(0, 1)$.

3.6. Post-hoc Frequency Bias Correction

Many answer sets suffer from class imbalance. In drug target identification, some targets are significantly more well-studied than others. As a result, the learned model $p(a_i|q)$ can be strongly correlated with the prior $p(a_i)$.

While bias can be informative (e.g. reflecting the fact that some targets are involved in more diseases than others) it can also be misleading (e.g. reflecting publishing trends rather than underlying biology). To flexibly control for bias, we therefore use a corrected output based on the frequency of each answer in the training corpus and parameterized by $c \in [0, 1]$ (Appendix G). When $c = 0$ the scores and rankings are unaltered; when $c = 1$, the rankings reflect the pointwise mutual information (PMI) of the query and answer, inspired by the use of PMI in NLP co-occurrence statistics (Church & Hanks, 1990). In the results we report both uncorrected ($c = 0$; R2E-uncor) and partially corrected ($c = 0.5$; R2E-cor) rankings, with the latter value selected using a validation set (Appendix D). In Shapley value explanations, the bias correction can be represented as an additive feature.

4. Experiments and Results

We evaluate R2E performance on three datasets aligned with drug target identification, each involving prediction over *Genes*:

- **Held-out Biomedical Literature**: Predicting masked genes in biomedical literature sentences taken from abstracts published after the publication of the training data and retrieval corpus.
- **Gene Description Facts**: Predicting masked genes in sentences containing human-curated information about the gene, based on gene descriptions provided by UniProt (Consortium, 2022).
- **Clinical Trial Outcomes**: Retrospectively predicting success or failure in historical clinical trials based on evidence published before the trials, using the disease indication and drug target (gene).

Evaluation dataset statistics are summarised in Appendix I, and dataset access in Appendix A.

For *Gene Description Facts* and *Clinical Trial Outcomes*, we also construct *Evidence Annotations* datasets to evaluate the alignment of R2E explanations with expert reasoning. Specifically, we look at the strength of relationship between R2E Shapley values and GPT-4 (Achiam et al., 2023) binary annotations of whether each piece of explanatory evidence is relevant or irrelevant to the query. We validate GPT-4 annotations against human drug discovery expert annotations.

Given their greater orthogonality to the R2E training objective, we choose to focus on *Gene Description Facts* and *Clinical Trial Outcomes* in the main text, and include a complete set of results on *Held-out Biomedical Literature* in Appendix J (Table S4). In short, on *Held-out Biomedical Literature*, R2E outperformed all baselines. We found R2E is able to leverage retrieved literature that it was not trained on, further improving performance.

4.1. Metrics

For ranking *Genes* on *Held-out Biomedical Literature* and *Gene Description Facts*, we report mean reciprocal rank (MRR), mean rank (MR), hits@10 (h@10) and hits@200 (h@200). In the case of *Gene Description Facts*, we used macro ranking metrics to ensure each gene is given equal weight irrespective of frequency. For *Clinical Trial Outcomes* we primarily report AUROC, but include relative success results in Appendix U for consistency with Minikel et al. (2024). To compute p values for AUROC comparisons we use DeLong test. For relative success comparisons we use Z-test and report confidence intervals using Katz method (Katz et al., 1978). For *Evidence Annotations*, we report AUROC for the R2E Shapley scores of evidence sentences against GPT-4 annotations, and accuracy when validating GPT-4 against human expert annotations.

4.2. Baselines and Ablations

In addition to MLM (Section 3.2), we include two baselines throughout: FREQ and MCS. For FREQ, entities were scored according to their frequency in the training set of $\mathcal{D}$. For MCS (mean cosine similarity), each entity a_i was scored by computing $\frac{1}{64}\sum_{j=1}^{64}(d^e_{ji}\cdot q)/(\|d^e_{ji}\|\|q\|)$ for the query q.

For *Clinical Trial Outcomes*, we include a genetics baseline widely used throughout the pharmaceutical industry (setup described in detail in Appendix Q). Along with other relative success results in Appendix U, we also perform a comparison to a few-shot chain-of-thought prompted GPT-4 baseline with retrieval augmentation (setup detailed in Appendix V). For extensive ablations of R2E components, including the Retriever, Reasoner and literature bias correction, see Appendix L.

4.3. Predicting Genes in Gene Description Facts

Dataset Construction We first sought to validate that R2E could perform well on predicting protein-coding genes in human-curated facts describing gene function, before proceeding to the scientific discovery task of predicting clinical trial outcome. We extracted descriptions of protein functions for our *Genes* entities from UniProt (Universal Protein Resource) (Consortium, 2022). Each description is a human-written summary of function, and consists of one or more sentences. We used a combination of regular expressions and Claude 2.0[1] to extract a set of [MASK]-containing facts from each description. Further details of the source and preprocessing of the dataset, including the Claude prompt and an example gene description with extracted facts, are found in Appendix M. R2E was trained on, and retrieved from, all years of literature evidence for the *Gene Description Facts* evaluation.

We also constructed an *Evidence Annotations* dataset by having GPT-4 (prompt in Appendix N) annotate as query-relevant or irrelevant, all evidence for 50 randomly sampled *Gene Description Facts* query-entity pairs (positive examples), and the same 50 queries with randomly sampled alternative entities (negative examples), obtaining 6400 annotated query-evidence pairs. To validate GPT-4 annotations, a human drug discovery expert following the GPT-4 prompt annotated all evidence for a subset of 8 randomly sampled examples (4 positive, 4 negative; 512 annotations).

Results R2E substantially improved on all baselines, both with and without bias correction (Table 1). As expected, bias correction was helpful. R2E metrics here appear to

[1] https://www.anthropic.com/

Table 1. **Gene Description Facts**: R2E macro ranking metrics.

METRIC	BASELINES			R2E	
	FREQ	MCS	MLM	UNCOR	COR
MRR	<0.001	0.176	0.167	0.202	**0.260**
MR	8252	1776	2208	937	**599**
H@10	<0.001	0.309	0.296	0.349	**0.434**
H@200	0.013	0.622	0.590	0.701	**0.776**

show greater improvement over baselines than for the *Held-out Biomedical Literature* dataset in Table S4. This may reflect a tendency for gene description facts to describe more well-established knowledge than literature sentences; as a result, R2E may benefit from its access to such facts, when more directly stated in the retrieved evidence sentences. Additionally, there was a strong correlation between evidence Shapley values and GPT-4 relevance annotations (AUROC: 0.824). Combined with a 71.5% agreement rate between GPT-4 and human-expert annotations, the agreement between R2E and GPT-4 suggests that R2E has correctly learnt to prioritise evidence for its predictions. See Appendix O for examples.

4.4. Predicting Clinical Trial Outcomes

Dataset Construction We constructed a benchmark of gene-disease pairs (therapeutic hypotheses) from clinical trials as per Nelson et al. 2015, using the PharmaProjects database (Citeline) (1,449 success, 4,222 failure, Appendix P). This benchmark focused on *in vivo* efficacy of therapeutic hypotheses as demonstrated by transition of drugs from Phase II/III clinical trials to regulatory approval.

To avoid leakage due to reporting of clinical trial results in literature, we removed drugs investigated prior to 2005 (Appendix P) and used pre-2005 literature for R2E training and retrieval (Appendix B). We scored therapeutic hypotheses using a query template "*[MASK] is a promising drug target for the treatment of* {*DISEASE*}.", substituting {*DISEASE*} with the PharmaProjects disease (Appendix T).

As the ability of genetics methods such as locus-to-gene (Mountjoy et al., 2021) to predict successful clinical development (Nelson et al., 2015; Ochoa et al., 2022; Minikel et al., 2024) drives their wide use in target identification, we used the most recently published PharmaProjects-aligned dataset of genetics predictions (Minikel et al., 2024) (Appendix Q) as a competitive baseline. In order to validate our *Clinical Trial Outcomes* data, we corroborated the published result (Minikel et al., 2024) that the probability of clinical success of therapeutic hypotheses supported by genetics evidence is approximately double the probability without supporting genetics evidence (relative success: 1.98; 95% CI (1.76, 2.24); Appendix U.2).

Table 2. **Clinical Trial Outcomes**: AUROC for R2E with retrieval corpus of literature-alone, genetics-alone, or both combined.

MODEL	CORPUS	AUROC
GENETIC	GENETICS	0.545
FREQ	LITERATURE	0.561
MCS	LITERATURE	0.623
MLM	LITERATURE	0.630
R2E-UNCOR	GENETICS	0.579
R2E-UNCOR	LITERATURE	0.629
R2E-COR	LITERATURE	0.632
R2E-COR	BOTH	**0.633**
R2E-AUDIT	BOTH	**0.638**

We also constructed an *Evidence Annotations* dataset with GPT-4 (prompt in Appendix R) assessing the relevance of all 64 evidence passages for 100 *Clinical Trial Outcome* therapeutic hypotheses (50 success, 50 failure; randomly sampled), obtaining 6400 annotated query-evidence pairs. To validate GPT-4 annotations, a human drug discovery expert following the GPT-4 prompt annotated all evidence for 8 hypotheses for which they had most knowledge (4 success, 4 failure; 512 annotations).

Multimodality via Templating into Natural Language We assessed R2E's ability to reason from genetics by generating a sentence for every row in the genetics dataset used in the genetics baseline (77,645 total), with the simple template "*[MASK] is genetically associated with* {*MeSH name*}.". The MeSH name, as supplied in Minikel et al. 2024, was programmatically reformatted to better align with naming conventions in the biomedical literature (details in Appendix T). This genetics corpus was given to the R2E Retriever alone and in combination with the pre-2005 biomedical literature.

Results Table 2 shows primary results, while Appendix U includes several further results and detailed discussions, including on relative success (Appendices U.1-U.3; Figure S4). Overall, R2E variants incorporating biomedical literature matched or outperformed all baselines.

Notably, R2E significantly outperformed the widely-used genetics baseline (Genetic) when leveraging only the exact same underlying genetics data templated into sentences (R2E-uncor (genetic); $p < 0.001$). This could be explained by the language model's capacity to leverage "soft" semantic associations (e.g. recognizing correlations between diseases / traits beyond ontological similarity), as corroborated by the inspection of high-scoring genetics evidence (Appendix U.7; Figure S6). The addition of literature resulted in a significant further improvement ($p < 0.001$). The relative under-performance of models using genetics data alone compared to those using biomedical literature likely reflects the lack of genetic coverage of diseases, despite it being predictive when available. In contrast, the literature has broad coverage across diseases. Figure S5 (Appendix U.6) shows performance by disease area with greater variability for genetics. R2E also significantly outperformed the few-shot chain-of-thought prompted GPT-4 baseline with retrieval augmentation (Appendix U.4).

There was only a marginal improvement from combining templated genetics evidence and the biomedical literature over literature alone. This could be explained by a combination of the $\sim$200:1 balance of literature to genetics-derived sentences in the evidence corpus, and the potential redundancy of the genetics evidence given information already represented in the literature. Additional approaches to combining data sources, with similar performance, are compared in Appendix U.5 (Table S8).

Evidence Shapley values correlated with binary GPT-4 relevance annotations (AUROC: 0.665) and GPT-4 with human-expert annotations (82.2% agreement rate). Together, the results suggest moderate agreement between the three sources on evidence relevance. See Appendix S for examples.

4.5. Auditing Explanation Evidence with GPT-4

We sought to assess the hypothesis that R2E explanations could enable human- or LLM-in-the-loop feedback to remove false positive evidence. Pooling R2E predictions on the *Clinical Trial Outcomes* dataset, we used GPT-4 to annotate the relevance of 20,000 query-evidence pairs with the highest Shapley values (computed on pre-sigmoid outputs). We then reran R2E-cor inference on the full dataset, replacing evidence labelled as irrelevant with the NULL embedding, referred to as R2E-audit. R2E-audit obtained a small but significant increase in AUROC (Table 2, $p = 0.004$). Said differently, by allowing evidence to be audited, R2E's explainability enabled further performance improvement. See Appendix R for the GPT-4 prompt and Appendix W for examples of auditing.

5. Limitations and Future Work

Retrieving evidence at inference time to make predictions has a cost: each answer score requires a vector search over the answer's evidence, followed by a model forward pass. In comparison, predicting with a multiclass model (MLM) requires a single forward pass without retrieval. However, retrieval and reasoning can be parallelized across answers for efficient scaling (Appendix H).

Retrieval-based inference has flexibility benefits beyond those explored here. By filtering retrieved evidence on document metadata, users could customize the ranking at inference time; with a scientific literature dataset, this could include filtering supporting evidence to specific timespans, publications, impact factors, paper sections, or keywords.

In Sections 4.3 and 4.4, we applied the model directly to downstream tasks; in the case of clinical trials, we simply adopted a one-size-fits-all query template. Instead, the system could be fine-tuned for the task of interest. Fine-tuning with human feedback is of particular interest here, since with R2E a user can focus on faulty *evidence use* (as opposed to a faulty prediction). Similarly, an LLM could be used to generate evidence-level labels for model fine-tuning in addition to inference-time auditing.

The evidence templating approach used for genetics is relatively general, and could be applied to other data modalities, such as transcriptomics evidence in drug discovery. However, care should be taken with respect to the distribution of the training data. For example, for scientific applications, evidence should be templated consistently with how it might be discussed in the literature corpus.

Performance gains might be made by scaling the Retriever and Reasoner, as well as extending to longer literature passages to increase context, for example paragraphs instead of sentences.

6. Conclusions

By retrieving evidence to make predictions, R2E enables faithful and quantitative explainability, a critical feature in complex, high-stakes decision-making. R2E matched or outperformed all target identification baselines across the three evaluation tasks. Combined with the proposed bias correction technique, this improves the ability to make informed predictions about novel and understudied, but promising targets. Finally, R2E outperformed a widely-used competing approach on the important and challenging task of predicting clinical trial efficacy outcomes, without task-specific fine-tuning. Performance was further improved by auditing R2E's prediction explanations using GPT-4, an approach made possible by the retrieval-based setup. We show here that retrieval can provide not only performance and flexibility advantages, but also significantly improved transparency into how the model reasons from evidence.

Impact Statement

This goal of this work is to advance the field of Machine Learning. There are many potential widely discussed societal consequences of developments in machine learning that we do not think warrant specifically highlighting.

In general, as detailed in Section 1, the explainability of R2E has the potential to positively impact the utility and adoption of models in high-stakes human-in-the-loop settings where explainability is often paramount, as exemplified by target identification. For target identification specifically, the improvements here could have significant positive consequences for the success of drug development programs and therefore the rate at which new more efficacious therapies become available to patients.

The application of R2E to predict and explain protein-coding genes in response to a user query is quite different to either the generality of large language models or the structural biology and chemistry foci of the AI-enabled biological tools most typically associated with any potential dual risk concern. As with other tools that facilitate biomedical research and understanding, the ability to identify and understand particular genes could be applied in a range of use cases. For this paper, we do not believe there to be material risks to highlight, especially noting: (1) We are not releasing proprietary training data, code, or model weights; (2) Explanations provided by R2E are either publicly-available extracts from the scientific literature or non-textual data templated in natural language, and can be interpreted by expert users in the context of their wider biomedical understanding, but do not significantly lower the barrier to entry for non-experts users; (3) R2E is predicting at the level of drug targets, with multiple complex downstream steps required to translate the identification of a target that may achieve a particular biological effect, into a capability to intervene on that target.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Aliper, A., Kudrin, R., Polykovskiy, D., Kamya, P., Tutubalina, E., Chen, S., Ren, F., and Zhavoronkov, A. Prediction of clinical trials outcomes based on target choice and clinical trial design with multi-modal artificial intelligence. *Clinical Pharmacology & Therapeutics*, 114(5): 972–980, 2023.

Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., Van Den Driessche, G. B., Lespiau, J.-B., Damoc, B., Clark, A., et al. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pp. 2206–2240. PMLR, 2022.

Brayne, A., Wiatrak, M., and Corneil, D. On masked language models for contextual link prediction. In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pp. 87–99. Association for Computational Linguistics, 2022.

Chen, D., Fisch, A., Weston, J., and Bordes, A. Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1870–1879, 2017.

Church, K. and Hanks, P. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29, 1990.

Citeline. Citeline's pharmaprojects database. https://api.pharmaintelligence.informa.com/v1/feed/drug/.

Consortium, T. U. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531, 2022.

Crabbé, J., Qian, Z., Imrie, F., and van der Schaar, M. Explaining latent representations with a corpus of examples. *Advances in Neural Information Processing Systems*, 34: 12154–12166, 2021.

Ghorbani, A. and Zou, J. Data shapley: Equitable valuation of data for machine learning. In *International conference on machine learning*, pp. 2242–2251. PMLR, 2019.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., and Poon, H. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23, 2021.

Guu, K., Lee, K., Tung, Z., Pasupat, P., and Chang, M. Retrieval augmented language model pre-training. In *International conference on machine learning*, pp. 3929–3938. PMLR, 2020.

Hammoudeh, Z. and Lowd, D. Training data influence analysis and estimation: A survey. *Machine Learning*, pp. 1–53, 2024.

Izacard, G. and Grave, E. Leveraging passage retrieval with generative models for open domain question answering. In *EACL 2021-16th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 874–880. Association for Computational Linguistics, 2021.

Izacard, G., Lewis, P., Lomeli, M., Hosseini, L., Petroni, F., Schick, T., Dwivedi-Yu, J., Joulin, A., Riedel, S., and Grave, E. Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*, 2022.

Johnson, J., Douze, M., and Jégou, H. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7 (3):535–547, 2019.

Jurafsky, D. and Martin, J. H. Speech and language processing (3rd (draft) ed.), 2019.

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, 2020.

Katz, D. et al. Obtaining confidence intervals for the risk ratio in cohort studies. *Biometrics*, 34(3):469–474, 1978. doi: 10.2307/2530610. URL https://doi.org/10.2307/2530610. Accessed 16 May 2024.

Khandelwal, U., Levy, O., Jurafsky, D., Zettlemoyer, L., and Lewis, M. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*, 2019.

Kulesza, T., Burnett, M., Wong, W.-K., and Stumpf, S. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*, pp. 126–137, 2015.

Lee, J., Lee, Y., Kim, J., Kosiorek, A., Choi, S., and Teh, Y. W. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 3744–3753, 2019.

Lee, J., Sung, M., Kang, J., and Chen, D. Learning dense representations of phrases at scale. *arXiv preprint arXiv:2012.12624*, 2020.

Lewis, P., Denoyer, L., and Riedel, S. Unsupervised question answering by cloze translation. *arXiv preprint arXiv:1906.04980*, 2019.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization, 2019.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Maglott, D., Ostell, J., Pruitt, K. D., and Tatusova, T. Entrez gene: gene-centered information at ncbi. *Nucleic acids research*, 33(suppl_1):D54–D58, 2005.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

Minikel, E. V., Painter, J. L., Dong, C. C., and Nelson, M. R. Refining the impact of genetic evidence on clinical success. *Nature*, pp. 1–6, 2024.

Mitchell, R., Cooper, J., Frank, E., and Holmes, G. Sampling permutations for shapley value estimation. *The Journal of Machine Learning Research*, 23(1):2082–2127, 2022.

Mountjoy, E., Schmidt, E. M., Carmona, M., et al. An open approach to systematically prioritize causal variants and genes at all published human gwas trait-associated loci. *Nature Genetics*, 53:1527–1533, 2021. doi: 10.1038/s41588-021-00945-5.

Nelson, M. R., Tipney, H., Painter, J. L., Shen, J., Nicoletti, P., Shen, Y., Floratos, A., Sham, P. C., Li, M. J., Wang, J., et al. The support of human genetic evidence for approved drug indications. *Nature genetics*, 47(8):856–860, 2015.

Ochoa, D., Karim, M., Ghoussaini, M., Hulcoop, D. G., McDonagh, E. M., and Dunham, I. Human genetics evidence supports two-thirds of the 2021 fda-approved drugs. *Nat Rev Drug Discov*, 21(8):551, 2022.

Paliwal, S., de Giorgio, A., Neil, D., Michel, J.-B., and Lacoste, A. M. Preclinical validation of therapeutic targets predicted by tensor factorization on heterogeneous graphs. *Scientific reports*, 10(1):18250, 2020.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library, 2019.

Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., and Miller, A. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2463–2473, 2019.

Shapley, L. S. et al. A value for n-person games. *Contributions to the Theory of Games II*, 1953.

Smith, D. P., Oechsle, O., Rawling, M. J., Savory, E., Lacoste, A., and Richardson, P. J. Expert-augmented computational drug repurposing identified baricitinib as a treatment for covid-19. *Frontiers in Pharmacology*, 12: 709856, 2021.

Sourati, J. and Evans, J. A. Accelerating science with human-aware artificial intelligence. *Nature Human Behaviour*, 7(10):1682–1696, 2023.

Sui, Y., Wu, G., and Sanner, S. Representer point selection via local jacobian expansion for post-hoc classifier explanation of deep neural networks and ensemble models. *Advances in neural information processing systems*, 34: 23347–23358, 2021.

Trajanoska, K., Bhérer, C., Taliun, D., Zhou, S., Richards, J. B., and Mooser, V. From target discovery to clinical drug development with human genetics. *Nature*, 620 (7975):737–745, 2023.

Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., Persson, K. A., Ceder, G., and Jain, A. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, 571(7763):95–98, 2019.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Wong, C. H., Siah, K. W., and Lo, A. W. Estimation of clinical trial success rates and related parameters. *Biostatistics*, 20(2):273–286, 2019.

Zaharia, M., Xin, R. S., Wendell, P., Das, T., Armbrust, M., Dave, A., Meng, X., Rosen, J., Venkataraman, S., Franklin, M. J., Ghodsi, A., Gonzalez, J., Shenker, S., and Stoica, I. Apache spark: A unified engine for big data processing. *Commun. ACM*, 59(11):56–65, 2016.

A. Accessing Evaluation Datasets

We make the three performance evaluation datasets used in this paper publicly available as part of the Supplementary Material, licensed under CC BY-NC-SA 4.0. Specific licensing information for the datasets is as follows:

- *Clinical Trials Outcomes* is licensed under CC BY-NC-SA 4.0. We have permission from Citeline PharmaProjects to publicly release the subset of their data that is used here.
- *Gene Description Facts* is licensed under CC BY-NC-SA 4.0. It is adapted from "Universal Protein Resource (UniProt)" by Uniprot Consortium, used under CC BY 4.0.
- *Held-out Biomedical Literature* validation and test dataset sentences are courtesy of the National Library of Medicine.

B. Masked Entity-Linked Corpus, Dataset Splits & Sizes

The large-scale corpus of scientific documents consisted of open access PubMed abstracts and PMC full texts as well as paid access Springer, Wiley and Elsevier full texts. We performed entity linking using a proprietary method (Appendix C), however any entity linking approach may be used (e.g. dictionary matching). Individual sentences were used as passages.

We filtered to sentences in the corpus that contained both: i) one or more protein-coding genes (entity set referred to as *Genes*), and ii) one or more non-gene grounded biomedical entities (e.g. diseases, biological pathways etc.), to select for an informative corpus. This process yielded 160 million sentences.

We created three distinct corpus splits $\mathcal{S}_1$, $\mathcal{S}_2$, and $\mathcal{S}_3$ (Figure S3). These splits were generated at the level of entire documents to reduce the occurrence of highly similar sentences between splits.

For *Held-out Biomedical Literature* (Appendix J) and *Clinical Trial Outcomes* (Section 4.4) experiments, where evaluation queries were associated with metadata for year of publication and earliest clinical development date respectively, a temporal year split setup was used to ensure models trained on and retrieved from sentences prior to the start year of the evaluation data. Specifically, for these year split experiments, $\mathcal{S}_1$ and $\mathcal{S}_2$ were random samples from *before* the split year with 1.5 million sentences allocated to $\mathcal{S}_2$ and the remainder to $\mathcal{S}_1$. $\mathcal{S}_3$ contained all sentences from documents *after* the split year. A split year of 2005 was used for *Clinical Trial Outcomes* ($|\mathcal{S}_1| = 16.2$ million sentences), and a split year of 2020 for *Held-out Biomedical Literature* ($|\mathcal{S}_1| = 112$ million sentences).

For *Gene Description Facts* experiments (Section 4.3), where evaluation queries did not correspond to a particular year, no year split was used. Specifically, $\mathcal{S}_1$, $\mathcal{S}_2$, and $\mathcal{S}_3$ were all random samples of the corpus, with 1.5 million sentences allocated to each of $\mathcal{S}_2$ and $\mathcal{S}_3$, and the remainder to $\mathcal{S}_1$ (157 million sentences).

Training, validation and testing datasets were then constructed for both R2E Retriever / MLM and R2E Reasoner, by using the appropriate $\mathcal{S}_1$, $\mathcal{S}_2$, and $\mathcal{S}_3$ splits.

For the R2E Retriever / MLM, training and validation datasets were composed as follows:

- $\mathcal{D}^{\text{MLM}}_{train} = \mathcal{S}_1$
- $\mathcal{D}^{\text{MLM}}_{val} = \mathcal{S}_2$

For the R2E Reasoner, for each of train, validation and test, both retrieval and query corpora were needed, to ensure query sentences were not also included in the retrieval corpus. We use $\mathcal{D}^e$ to refer to a retrieval corpus of evidence sentences and $\mathcal{D}^q$ to refer to the query corpus of sentences. The datasets were composed as follows:

- $\mathcal{D}^e_{train} = \mathcal{S}_1$
- $\mathcal{D}^q_{train} = \mathcal{S}_2$
- $\mathcal{D}^e_{val} = \mathcal{S}_1 \cup \mathcal{S}_2$
- $\mathcal{D}^q_{val} \subset \mathcal{S}_3$
- $\mathcal{D}^e_{test} = \mathcal{S}_1 \cup \mathcal{S}_2$

- $\mathcal{D}^q_{test} \subset \mathcal{S}_3 : \mathcal{D}^q_{test} \cap \mathcal{D}^q_{val} = \emptyset$, i.e. a held-out subset of $\mathcal{S}_3$, without overlap with $\mathcal{D}^q_{val}$

The above splitting procedure is illustrated in Figure S3 for the case of the 2020 year split setup used for *Held-out Biomedical Literature* experiments. For this *Held-out Biomedical Literature* setup, the disjoint subsets sampled from $\mathcal{S}_3$ and used to create overall validation ($\mathcal{D}^q_{val}$) and test ($\mathcal{D}^q_{test}$) queries, are those used to report ranking metric evaluations over all genes in *Genes*; namely the:

- *Held-out Biomedical Literature* validation dataset: 1 sentence per gene, sampled from publicly-available abstract section sentences from 2020 onwards. Used for hyperparameter selection and ablation experiments described in Appendices D & L respectively.
- *Held-out Biomedical Literature* test dataset: 1 sentence per gene per year for 2020 onwards, sampled from publicly-available abstract section sentences. Used for evaluations described in Section 4 and Appendix J, including evaluation of the MLM and other baselines.

Note the key difference between this 2020 year split setup for *Held-out Biomedical Literature*, and the setups for the other two evaluation datasets were:

- Different year splits (as described above)
- The queries used in evaluation were derived from those specific evaluation datasets, not a held-out split of the literature corpus (i.e. $\mathcal{D}^q_{eval} \neq \mathcal{D}^q_{test}$)

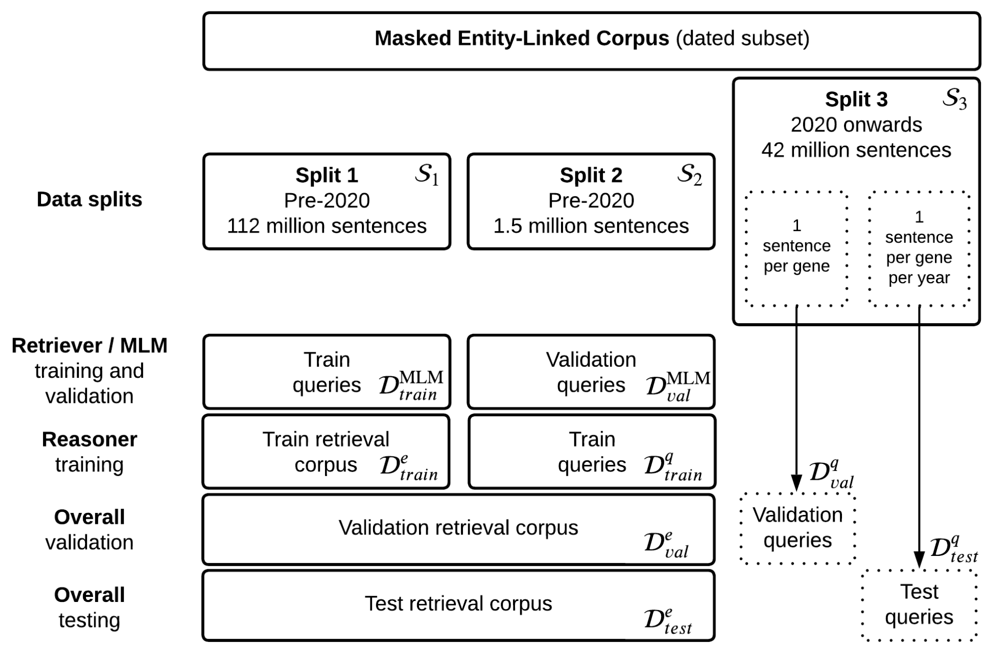

Figure S3. ***Masked entity-linked corpus*** **for** ***Held-out Biomedical Literature*** **experiments**. Here we illustrate how the *masked entity-linked corpus* was partitioned to enable Reasoner/MLM and Retriever training, validation, and testing. Specifically the example of a 2020 year split setup is shown, as was used for *Held-out Biomedical Literature* experiments.

C. Entity Linking

We used a proprietary entity linking methodology based on dictionaries of entities and synonyms, as well as an abbreviation detection algorithm and a model that resolves type ambiguities based on the context of each mention. The dictionaries were created from several sources.

1. External ontologies.
2. Human annotations of synonyms discovered by machine learning methods.
3. Automatic synonym generation to cover e.g. variation in punctuation, Greek letters and plurals of terms.

For the protein-coding gene target entities, referred to as *Genes* and used throughout the paper, we ground both gene and protein forms to the same entity, under the assumption of a 1:1 relationship between a gene and the protein it encodes.

D. R2E Hyperparameters

The R2E model was implemented using PyTorch deep learning library (Paszke et al., 2019).

All sentences were tokenized, and then truncated and padded to a length of 128, using the same vocabulary as PubMedBERT (Gu et al., 2021). Pre-processing of training examples for both Retriever and Reasoner training was done using Apache Spark (Zaharia et al., 2016). The Retriever and Reasoner were trained sequentially, each for 10 epochs on a single Tesla V100 GPU, with a total training time of approximately 1 week.

The final R2E Retriever architecture, as well as the MLM baseline, consisted of a scaled down version of PubMedBERT (Gu et al., 2021) trained from scratch on the task described in 3.2, with 4 hidden layers, 4 attention heads, an intermediate size of 512, a hidden size of 256, and total size of 10 million parameters. Final Retriever/MLM training used a batch size of 512, a categorical cross-entropy loss, and an AdamW optimizer (Loshchilov & Hutter, 2019) with a learning rate of 0.0001 and no weight decay.

The R2E architecture is summarised in Figure 2. The final query-evidence encoder component of the R2E Reasoner architecture consisted of a layer normalisation across all concatenated query/evidence pairs, then two conv1d layers each with kernel size of 1 (first layer: 2 input channels, 8 output channels; second layer: 8 input channels, 1 output channel) across each query/evidence pair individually. The final evidence combiner component of the R2E Reasoner architecture consisted of a set transformer (Lee et al., 2019) over all query-evidence embeddings returning a single embedding, followed by a linear layer and sigmoid to output a binary probability. The set transformer had 4 heads, 2 induced set attention blocks with 32 inducing points for the encoder, and a pooling by multihead attention followed by two set attention blocks in the decoder. The Reasoner had a total size of 2 million parameters. After freezing the Retriever weights, the final Reasoner training used a batch size of 2048, binary cross-entropy loss, and AdamW optimizer with a learning rate of 0.0001 and weight decay of 0.001. For both training and inference, 64 evidence sentences were retrieved for a given query. A post-hoc frequency bias correction factor of 0.5 was used for the R2E-cor variant (Section 3.6 and Appendix G for details of post-hoc correction).

The post-hoc frequency bias correction factor selection and architectural comparison ablations (Appendix L) were based on MRR for a 2020 year split model, on a *Held-out Biomedical Literature* validation set containing one biomedical literature cloze-style query sentence per gene in *Genes* from publicly-available abstract sections (Appendix B). The resulting 15477 validation set queries were therefore sentences published from 2020 onwards, and retrieval corpus sentences published prior to 2020. The learning rate was chosen to reduce training time while maintaining training stability, and the batch size selected to optimise GPU utilisation. We did not evaluate variations of model scale and leave this to future work.

E. Approximating Evidence Shapley Values

We used a simple Monte Carlo method to approximate Shapley values, combined with antithetical sampling for variance reduction (Mitchell et al., 2022). The Shapley value was approximated as

$$\phi_i \approx \frac{1}{2M} \sum_{j=1}^{M} \left([g(S_j \cup \{i\}) - g(S_j)] + [g(\bar{S}_j \cup \{i\}) - g(\bar{S}_j)] \right) \tag{1}$$

where ϕ_i is the approximate Shapley value of feature i (an encoded query/evidence pair), M is the chosen number of sampled permutations, S_j is the set of features preceding i in the j-th permutation sample, $g(S_j)$ is the Reasoner output when only the features S_j are unmasked, $g(S_j \cup \{i\})$ is the Reasoner output when feature i is unmasked in addition to S_j, and $\bar{S}_j$ corresponds to the set of features preceding i in the reverse of the j-th permutation sample (equivalently, the set of features following i in the j-th permutation sample). The sum of the Shapley values over features plus the score when all features are NULL equates to the final score. Depending on the purpose, we use either the post-sigmoid output or the logit score for g. We use $M = 100$ whenever Shapley values are computed as part of this paper. See Appendix H for profiling of Shapley computation.

Algorithm 1 Generate permutation-approximated Shapley attributions for a single query.

Input: Number of permutations M, query-evidence embeddings $E = \{d_1^{qe}, \ldots, d_k^{qe}\}$, missing evidence embedding $NULL$, model forward function $g(\cdot)$
Output: Shapley value of query-evidence embeddings: $\phi_1, \ldots, \phi_k$
Initialize $\phi_i = 0$ for $i = 1, \ldots, k$
$2M$ antithetical sample of permutations p^j for $j \in 1, \ldots, 2M$ of the feature indices $1, \ldots, k$
where $p^{M+i} = \text{ReverseOrder}(p^i)$
$\tilde{E}_0 = \{\text{NULL}, \ldots, \text{NULL}\}$, with $|\tilde{E}_0| = k$
$s_0 \leftarrow g(\tilde{E}_0)$
for all $j \in \{1, \ldots, 2M\}$ **do**
  for all $i \in \{1, \ldots, k\}$ **do**
    $\tilde{E}_i^j \leftarrow \{d_{p^j[1]}^{qe}, \ldots, d_{p^j[i]}^{qe}, \text{NULL}, \ldots, \text{NULL}\}$, with $|\tilde{E}_i^j| = k$
    $s_i \leftarrow g(\tilde{E}_i^j)$
    $\phi_{p^j[i]} \leftarrow \frac{j-1}{j}\phi_{p^j[i]} + \frac{1}{j}(s_i - s_{i-1})$ {Cumulative average of marginals for feature $p^j[i]$ across permutations}
  end for
end for

F. Relationship between Multinomial and Binary Objectives

R2E is trained to predict the probability that a given query-entity pair is "true", i.e. that it came from a real occurrence in the literature and was not randomly generated. Given the labels $L \in \{0, 1\}$, the query (masked sentence) variable Q, the named entity answer variable A, the Reasoner parameters θ and the fixed Retriever parameters ψ, the model is trained to predict

$$\frac{1}{1 + \exp(-z(a_i, q_i))} \approx P(L = 1 | Q = q_i, A = a_i; \theta, \psi) \tag{2}$$

where $z(a_i, q_i)$ is the logit output of the network in response to a specific example i, i.e.

$$\begin{aligned} z(a_i, q_i) \approx &\log(P(L = 1 | Q = q_i, A = a_i; \theta, \psi)) \\ &- \log(P(L = 0 | Q = q_i, A = a_i; \theta, \psi)). \end{aligned} \tag{3}$$

Here, when $L = 0$, the example i corresponds to a negative example where Q and A have been chosen independently. Consider the case where the specific parameters θ and ψ have been learned such that the equality in Eq. 3 holds exactly; we are interested in the output in this case. We therefore assume the optimal output $z^*(a_i, q_i)$ and exclude the parameters.

The equation can be re-written using Bayes' Theorem,

$$\begin{aligned} z^*(a_i, q_i) = &\log(P(Q = q_i, A = a_i | L = 1)) + \log(p(L = 1)) - \log(P(Q = q_i, A = a_i)) \\ &- \log(P(Q = q_i, A = a_i | L = 0)) - \log(p(L = 0)) + \log(P(Q = q_i, A = a_i)). \end{aligned} \tag{4}$$

In our training setup, positive and negative examples are sampled equally often, i.e.

$$\log(p(L = 1)) = \log(p(L = 0)). \tag{5}$$

As a result, Eq. 4 simplifies to

$$z^*(a_i, q_i) = \log(P(Q = q_i, A = a_i | L = 1)) - \log(P(Q = q_i, A = a_i | L = 0)) \tag{6}$$

$$\tag{7}$$

Using the product rule

$$\begin{aligned} z^*(a_i, q_i) = &\log(P(A = a_i | Q = q_i, L = 1)) + \log(P(Q = q_i | L = 1)) \\ &- \log(P(A = a_i | Q = q_i, L = 0)) - \log(P(Q = q_i | L = 0)). \end{aligned}$$

The distribution over queries is also equal for positive and negative labels, as each query sentence is chosen for each condition once per epoch, simplifying to

$$z^*(a_i, q_i) = \log(P(A = a_i|Q = q_i, L = 1)) - \log(P(A = a_i|Q = q_i, L = 0)). \tag{8}$$

The distribution over named entity answers is independent of the query when conditioned on $L = 0$, because negative samples are chosen by randomly pairing queries and entities. So the second term here corresponds to our negative sampling distribution. Therefore, the output at optimality corresponds to

$$\begin{aligned} z^*(a_i, q_i) &= \log(P(A = a_i|Q = q_i, L = 1)) - \log(P(A = a_i|L = 0)) \\ &= \log(P(A = a_i|Q = q_i, L = 1)) + \log(|\mathcal{A}|) \end{aligned} \tag{9}$$

since the probability of choosing a given answer a_i as a negative sample during training is $\frac{1}{|\mathcal{A}|}$. Comparing to the optimal logit output of the MLM model, we see a close relationship:

$$z^{*,\mathrm{MLM}}(a_i, q_i) = \log(P(A = a_i|Q = q_i, L = 1)) + \log(Z) \tag{10}$$

where Z is the partition function (the MLM includes $L = 1$ implicitly as all examples are positive). The optimal logit outputs for the models therefore scale up to their respective normalization factors.

G. Post-hoc Frequency Bias Correction as Trading off Log Probability and Mutual Information

From Equation 9 in Appendix F, we find that the optimal model output logit scales with $\log(P(A|Q, L = 1))$, i.e. the probability of the answer given the query assuming a real example ($L = 1$). This score will be highly correlated with the prior distribution over the answer set, particularly for an imbalanced dataset (like the mentions of *Genes* in the scientific literature corpus used in the paper).

One approach to counteract the literature bias, if desired, is to instead consider the pointwise mutual information between a given answer and a given query:

$$\mathrm{PMI}(A = a; Q = q) = \log \frac{P(A = a|Q = q)}{P(A = a)}. \tag{11}$$

PMI is widely used in the NLP community to measure associations between keywords in a corpus, based on their marginal occurrence counts and joint co-occurrence counts (Jurafsky & Martin, 2019). Similarly, we find that it offers a straightforward means of correcting for class imbalance after training the model.

For a model that predicts a multiclass output (like the MLM), we can directly adapt the output. Specifically, after normalizing the outputs to remove $\log(Z)$, where Z is the partition function,

$$\begin{aligned} z_c^{\mathrm{MLM}}(a_i, q_i) &= z^{\mathrm{MLM}}(a_i, q_i) - c \cdot \log P(A = a_i|L = 1) \\ &\approx \log P(A = a_i|Q = q_i, L = 1) - c \cdot \log P(A = a_i|L = 1) \end{aligned} \tag{12}$$

where $P(A = a_i|L = 1)$ is estimated by the proportion of passages in the corpus where a_i is the correct answer. When $c = 0.0$, the two approaches are equivalent; while when $c = 1.0$, the output approximates the PMI score in Equation 11. Stronger corrections penalize common answers, and the score is only positive if the model's estimated answer probability for the given query is higher than the frequency-based prior.

In R2E, we instead note that the optimal logit score in Equation 9 already reflects PMI *if* the negative sampling probability $P(A = a_i|L_i = 0)$ was chosen to reflect the prior distribution over answers in the dataset, $P(A = a_i|L = 1)$. We therefore consider a negative distribution $P_c(A = a_i|L = 0)$ that trades off between a uniform distribution $\frac{1}{|\mathcal{M}|}$ and one based on the answer prior in the training corpus:

$$P_c(A = a_i|L = 0) = \frac{C(a_i)^c}{\sum_{i=1}^{|\mathcal{A}|} C(a_i)^c} \tag{13}$$

where $C(a_i)$ is the count of occurrences of answer a_i as a masked entity in the training corpus. When $c = 1$, this corresponds to the background distribution of a_i in the training corpus $P(A = a_i|L = 1)$; when $c = 0$, it corresponds to the uniform distribution $\frac{1}{|\mathcal{A}|}$.

One possible approach to bias correction is to set a fixed c during training and use the resulting negative sampling distribution in Equation 13. However, this approach grants less flexibility in terms of the desired bias correction at inference time. We therefore continue to use the fixed uniform distribution $\frac{1}{|\mathcal{A}|}$ and instead introduce a correction factor

$$f_c = \log \frac{1}{|\mathcal{A}|} - \log P_c(A = a_i | L = 0). \tag{14}$$

Applying this correction to the logit output of R2E after training (Equation 9) yields

$$z(a_i, q_i) + f_c \approx \log(P(A = a_i | Q = q_i, L = 1)) - \log P_c(A = a_i | L = 0) \tag{15}$$

which reflects a log probability estimate when $c = 0$ and a pointwise mutual information estimate when $c = 1$. We found that the best performance in terms of MRR on the *Held-out Biomedical Literature* validation dataset (Appendix B), was achieved with a partial correction of $c = 0.5$. We refer to this as R2E-cor, and refer to the case with $c = 0.0$ as R2E-uncor.

The bias correction can be straightforwardly identified as an additional additive feature during Shapley value estimation to communicate its impact to the user. For under-represented answers, it can be seen as compensating for "missing" evidence, e.g. due to the lack of research on a particular target.

H. R2E Inference Speed

We profiled R2E for both prediction and explanation. We used CPUs only, though GPUs could be used to achieve additional speed-ups by reducing the time taken for the forward pass.

H.1. Prediction

For prediction on CPUs, the MLM baseline took ∼140ms over one query on one core, obtaining scores for all 19,176 genes via a single forward pass. By comparison, the non-negligible components of R2E inference time are:

1. The batched forward pass over 19,176 query-evidence pairs (one for each gene), through the Reasoner - ∼7.4s on one core, and scales linearly with cores
2. Vector searches over the 19,176 FAISS indices corresponding to each gene, for the Retriever - ∼27s on one core, 1.5s on 40 cores or <0.15s if one core per index

Since the evidence is split into separate retrieval indices for each of the potential answers, the top evidence from each can be found in parallel. Therefore, search can generally scale more efficiently than for a traditional single FAISS index. To optimise inference, the forward pass should be run in batches while the search results for each potential answer are returned from each corresponding FAISS index. As a result, the total time is then largely defined by the maximum time for the above two stages of batched forward pass and vector search, given the relevant parallelisation.

These results assume exact brute force vector search (IndexFlatIP search indices from FAISS (Johnson et al., 2019)) with a complexity of $O(nd)$, where n is the number of vectors in the given search index and d is the dimensionality of each vector. While vector search was not a bottleneck in our setup, if inference speed were a concern as the retrieval corpus scales, there are many out-of-the-box options for more efficient approximate nearest neighbour search indices, including within FAISS. The R2E profiling results above also assume access to a machine with ∼300GB memory for the FAISS indices; fast inference is achieved on widely available resources.

H.2. Explanation

For inference time explanations, we compute Shapley values using the permutation-based method detailed in Appendix E, using $M = 100$ permutations (200 with antithetical sampling). With 64 evidence sentences retrieved for a given query, this results in 12,800 evidence set variations required to compute all 64 Shapley values. Therefore, <10 forward passes are required, with a reasonable batch size. Given the small size of the Reasoner module (2 million parameters), generating an explanation takes ∼5 seconds using a single CPU only.

We also note that more efficient methods exist for approximating Shapley values (Lundberg & Lee, 2017), particularly for deep networks. However, since Shapley value efficiency is neither our primary focus nor prohibitive, we used a permutation-based approach (Appendix E).

I. Evaluation Dataset Statistics

The total sizes of all test/evaluation datasets are shown in Table S3.

Table S3. **Evaluation dataset statistics**

DATASET	SUBSET	COUNT
HELD-OUT BIOMEDICAL LITERATURE	2020	14429
	2021	14859
	2022	15074
GENE DESCRIPTION FACTS		60839
GDF EVIDENCE ANNOTATIONS (HUMAN EXPERT)	QUERY-GENE PAIRS	8
	POSITIVES:NEGATIVES	4:4
	EVIDENCE	512
GDF EVIDENCE ANNOTATIONS (GPT-4)	QUERY-GENE PAIRS	100
	POSITIVES:NEGATIVES	50:50
	EVIDENCE	6400
CLINICAL TRIAL OUTCOMES (2005 ONWARDS)	SUCCESSES	1449
	FAILS	4222
CTO EVIDENCE ANNOTATIONS (HUMAN EXPERT)	QUERY-TARGET PAIRS	8
	SUCCESSES:FAILS	4:4
	EVIDENCE	512
CTO EVIDENCE ANNOTATIONS (GPT-4)	QUERY-TARGET PAIRS	100
	SUCCESSES:FAILS	50:50
	EVIDENCE	6400

J. Predicting Genes in Held-out Biomedical Literature

Dataset Construction For all experiments in this section, we trained the MLM (R2E Retriever) and R2E Reasoner only on biomedical literature data published prior to 2020. Except where specified, R2E also only retrieved data published prior to 2020 (Figure S3). We then constructed a *Held-out Biomedical Literature* evaluation dataset from publicly-available paper abstracts. We generated a balanced dataset to obtain results reflecting performance across all 19,176 genes, not biased to the most well-studied (discussed further in Appendix K). We sampled one sentence per unique gene in *Genes* for each of the years 2020, 2021, and 2022; further details in Appendix B.

Results R2E improved on the baselines over all year subsets, both with and without bias correction (Table S4). Bias-corrected R2E improved on uncorrected performance, consistent with the use of a balanced evaluation dataset. For completeness, we show results on an imbalanced dataset (without stratification by gene in *Genes*) in Appendix K.

To test R2E's ability to leverage retrieved literature that it was not trained on, we enabled retrieval up to the year preceding the query sentence publication (rather than strictly prior to the 2020 training data cutoff). This improved performance (R2E-cor-updated, Table S4).

K. Comparison of Models on a Non-Stratified Held-out Biomedical Literature Dataset

Gene mention counts are extremely imbalanced in the literature. In the training data, of the 19,176 protein-coding genes, the most-well studied has approximately 2 million mentions, while the least studied 10,000 genes all have less than 1,000 mentions. For our *Held-out Biomedical Literature* dataset we used stratified sampling (stratification by gene in *Genes*) to obtain a class balanced test dataset, with equal counts of each gene to avoid dominance of well-studied genes. By preventing reliance of models on the gene frequency distribution prior, a class-balanced setup is especially challenging. Strong performance across the genome is desirable because understudied genes are of particular interest in drug discovery, when seeking new ways to treat a disease.

While our focus is therefore on balanced performance across the genome (results in Appendix J), for completeness, we also evaluated R2E on a dataset of 20,000 randomly-sampled publicly-available abstract sentences published from 2020 onwards, obtaining an imbalanced dataset *without* stratification by gene in *Genes*. As expected, the frequency-based baseline performs significantly better here relative to the stratified dataset in Table S4, reflecting that ability to rely on the frequency

Table S4. **Held-out Biomedical Literature**: Ranking metrics on a dataset consisting of one sentence per gene in *Genes* for each year of 2020, 2021, and 2022. MLM and R2E trained on data published prior to 2020. MCS, R2E-uncor and R2E-cor also retrieved data published prior to 2020. R2E-cor-updated retrieved up to the year before the publication year of the query sentence.

Metric	Query year	Baselines			R2E		
		Freq	MCS	MLM	Uncor	Cor	Cor-Updated
MRR	2020	<0.001	0.182	0.181	0.198	**0.233**	-
	2021	<0.001	0.172	0.169	0.187	**0.215**	**0.223**
	2022	<0.001	0.167	0.164	0.178	**0.205**	**0.219**
MR	2020	7661	3280	3465	2803	**2489**	-
	2021	7834	3568	3789	3032	**2695**	**2544**
	2022	7931	3770	4016	3287	**2902**	**2623**
H@10	2020	<0.001	0.268	0.269	0.291	**0.333**	-
	2021	<0.001	0.251	0.252	0.274	**0.313**	**0.324**
	2022	<0.001	0.243	0.243	0.260	**0.295**	**0.312**
H@200	2020	0.014	0.443	0.438	0.484	**0.521**	-
	2021	0.014	0.422	0.416	0.456	**0.497**	**0.509**
	2022	0.013	0.404	0.398	0.435	**0.473**	**0.496**

Table S5. **Non-stratified Held-out Biomedical literature**: R2E ranking metrics on a random subsplit (not stratified by gene in *Genes*) of query sentences published from 2020 onwards (20,000 queries), for an R2E model trained and retrieving from data prior to 2020.

Metric	Baselines			R2E	
	Freq	MCS	MLM	Uncor	Cor
MRR	0.026	**0.405**	0.399	0.403	0.350
MR	2321	**1114**	1305	1140	1456
H@10	0.070	0.520	0.519	**0.523**	0.500
H@200	0.304	0.691	0.686	**0.699**	0.686

distribution prior. Ranking metrics show similar performance for R2E, MCS and MLM (Table S5). In comparison, on the more challenging stratified setup R2E markedly outperforms baselines (Table S4). Comparing R2E and MLM, R2E's superior balanced performance across the genome could be explained by it's access to a knowledge base even for the most rare genes, avoiding the need to memorise knowledge of genes rarely seen at training time in the model parameters. R2E obtains superior performance on less studied genes without sacrificing performance on well-studied genes.

L. Architecture Ablation Experiments

We performed ablations of all core R2E architectural components, including the Reasoner, Retriever and frequency bias correction. A *Held-out Biomedical Literature* validation set was used for ablations experiments, consisting of one sentence per gene in *Genes* sampled from publicly-available abstract sentences published from 2020 onwards (as described in Appendices B & D), for an R2E model trained and retrieving from data prior to 2020. The results are summarised in Table S6. All ablations resulted in a drop in performance across all ranking metrics, demonstrating the benefit of R2E components.

L.1. Reasoner

The MCS baseline (Section 4.2) acts as an ablation of the R2E Reasoner, since it relies solely on query-evidence cosine similarities of the Retriever to obtain a score.

We also selectively ablated the convolutional query-evidence encoder component of the R2E Reasoner (Section 3.4) by substituting that component for a parameter-free Hadamard product between the query embedding and each evidence embedding. The Hadamard product was chosen in order to incorporate an inductive bias towards the cosine similarity.

L.2. Retriever

We ablated our task specific Retriever (Sections 3.2 & 3.3), by replacing it with an off-the-shelf biomedical transformer. Specifically we used a PubMedBERT model initialised with its published weights (Gu et al., 2021) as the Retriever. We created sentence embeddings by taking the mean over outputs corresponding to [MASK] tokens. This Retriever had a larger hidden size with 768 dimensional query and evidence embeddings. The R2E Reasoner was therefore linearly scaled to match this hidden size.

We also evaluated the MLM baseline (Section 3.2), which acts as an ablation of R2E in its entirety, taking a fully parametric approach to prediction.

L.3. Post-hoc frequency bias correction

We report results with and without bias correction.

Table S6. **Architecture ablations**: Ablated versions of R2E-uncor on a validation dataset consisting of one sentence per gene in *Genes* sampled from sentences published from 2020 onwards, while training on and retrieving from data prior to 2020. Hadamard: substituting the convolution layers of the Reasoner with a Hadamard product. PubMedBERT: substituting the Retriever for the PubMedBERT model.

METRIC	R2E		REASONER ABLATIONS		RETRIEVAL ABLATIONS	
	COR	UNCOR	MCS	HADAMARD	PUBMEDBERT	MLM
MRR	**0.211**	0.181	0.163	0.166	0.134	0.163
MR	**2873**	3210	3726	3260	3606	3945
H@10	**0.302**	0.262	0.241	0.253	0.207	0.242
H@200	**0.482**	0.443	0.409	0.441	0.389	0.404

M. Further Details on Creation of Gene Description Facts Dataset

We downloaded UniProt FTP server data version 2023_01 and extracted descriptions of protein functions for our set of protein-coding gene entities (*Genes*) from UniProt (Universal Protein Resource), used under CC BY (4.0), (Consortium, 2022) (by pulling "text" from UniProt entities with type "function" in the "comment" field). Each entry is a human-written description of function, and consists of one or more sentences.

After dropping all descriptions containing fewer than four words, we converted each description into a set of single-sentence facts as follows:

1. Descriptions were split into individual sentences and PubMed IDs removed, using regular expression operations.
2. Each sentence was converted into a fact containing a "[MASK]" referring to the gene and "[MASK]" in place of all gene mentions, using one-shot prompted Claude 2.0 language model from Anthropic (prompt template below). Sentences which Claude determined did not contain a suitable fact, were dropped.
3. "[MASK]"-containing facts were extracted from the Claude completion, and facts without any "[MASK]" mention were dropped.

For example, the description for the protein corresponding to gene *ELF2* is:

> "Isoform 1 transcriptionally activates the LYN and BLK promoters and acts synergistically with RUNX1 to transactivate the BLK promoter. Isoform 2 may function in repression of RUNX1-mediated transactivation."

From this description, the following facts were extracted for the evaluation dataset:

- [MASK] isoform 1 transcriptionally activates the LYN and BLK promoters and acts synergistically with RUNX1 to transactivate the BLK promoter.
- [MASK] isoform 2 may function in repression of RUNX1-mediated transactivation.

The following one-shot prompt template was used to convert sentences from pulled UniProt gene descriptions into [MASK]-containing facts. The gene GENE_NAME and UNIPROT_DESCRIPTION_SENTENCES were substituted into the template for each sentence-gene pair in the dataset, prior to querying Claude 2.0 via Anthropic's API.

```
{HUMAN_PROMPT}
# THE TASK:
You are an expert biologist. You will be given a set of sentences from a
DESCRIPTION of a GENE from UniProt.

Your instructions are to go one-by-one through each sentence in the
DESCRIPTION, and:
1. If the sentence states a fact about the specified GENE convert the
sentence into a FACT according to the FACT formatting shown in the <example>
below. 2. If, and only if, the sentence does not state any information
about the GENE, you may skip the sentence and indicate this with
"sentence[nb] SKIPPED" as shown in the <example> below.

# FORMATTING:
Here's an example input and output contained in the <example> XML tags,
to illustrate the format in which FACTs should be stated, including how to
indicate that a sentence has been skipped.

<example>

Input:
GENE: PGP
DESCRIPTION sentences:
<sentence1>Glycerol-3-phosphate phosphatase hydrolyzing glycerol-3-phosphate
into glycerol.</sentence1>
<sentence2>Thereby, regulates the cellular levels of glycerol-3-phosphate a
metabolic intermediate of glucose, lipid and energy metabolism.<\sentence2>
<sentence3>Was also shown to have a 2-phosphoglycolate phosphatase activity
and a tyrosine-protein phosphatase activity.</sentence3>
<sentence4>However, their physiological relevance is unclear
(PubMed:26755581).</sentence4>
<sentence5>In vitro, has also a phosphatase activity toward ADP, ATP, GDP
and GTP (By similarity).</sentence5>
<sentence6>Further work is needed to understand this.</sentence6>
<sentence7>(Microbial infection) Involved in replication of Rubella virus.
</sentence7>

Output:
Here are complete set of [MASK]-containing FACTs for each sentence about PGP:
<sentence1_fact>[MASK] is a glycerol-3-phosphate phosphatase that hydrolyzes
glycerol-3-phosphate into glycerol.</sentence1_fact>
<sentence2_fact>[MASK] regulates cellular levels of glycerol-3-phosphate, a
metabolic intermediate of glucose, lipid and energy metabolism.
</sentence2_fact>
<sentence3_fact>[MASK] has 2-phosphoglycolate phosphatase activity and
tyrosine-protein phosphatase activity.</sentence3_fact>
<sentence4_fact>sentence4 SKIPPED</sentence4_fact>
<sentence5_fact>In vitro, [MASK] has phosphatase activity toward ADP, ATP,
GDP and GTP.</sentence5_fact>
<sentence6_fact>sentence6 SKIPPED</sentence6_fact>
```

```
<sentence7_fact>[MASK] is involved in replication of Rubella virus.
</sentence7_fact>

</example>

# FACT REQUIREMENTS
You must note the following requirements, when constructing each FACT:
1. Each and every FACT must include one or more [MASK] tokens representing
the GENE.
2. All references to or synonyms of the GENE anywhere in a FACT, must also
be replaced by [MASK].
3. Only include information explicitly stated in the DESCRIPTION sentence
when extracting a FACT - do not elaborate with any additional information
from elsewhere.
4. You must go through every sentence.
5. You can only skip a sentence if it contains no information about the
GENE, and you must indicate this by stating the sentence was SKIPPED in
the corresponding sentence FACT XML tags.

# THE FINAL GENE AND DESCRIPTION SENTENCES
Now, paying attention to all the above instructions and example, please go
one-by-one through each sentence in the following DESCRIPTION and extract
each FACT for the stated GENE:

Input:
GENE: {GENE_NAME}
DESCRIPTION sentences:
{UNIPROT_DESCRIPTION_SENTENCES}

{AI_PROMPT}
Output:
Here are complete set of [MASK]-containing FACT(s) for each sentence about
{GENE_NAME}:
<sentence1_fact>
```

N. Further Details on Creation of Explanation Annotations for Gene Description Facts Dataset

We constructed *Evidence Annotations* for the *Gene Description Facts* dataset by having GPT-4 annotate query relevance for all evidence across 50 randomly sampled query-entity pairs (positive examples) and the same 50 queries but with a randomly sampled alternative entity (negative examples), resulting in 6400 query-evidence pairs (100 queries each retrieving 64 pieces of evidence) with a binary annotation.

We used GPT-4 to obtain relevant/irrelevant annotations for this task by using the following prompt, substituting in GENE_DESCRIPTION_FACT and EVIDENCE_SENTENCE:

```
You are a scientific expert working on target identification in drug
discovery.

Your task is to use your expertise to evaluate whether a piece of evidence
(referred to as EVIDENCE) about a masked target from an academic paper (in
the form of a sentence), provides relevant support to a specified biological
fact about that masked target (referred to as FACT). Please explain your
reasoning first before giving your answer. Provide your final answer by
stating either <answer>RELEVANT</answer> or <answer>IRRELEVANT</answer>
to indicate whether the evidence sentence provides relevant support for
```

```
the FACT or not. Please also indicate your confidence in your answer by
writing one of <confidence>HIGH</confidence> or <confidence>LOW</confidence>.

Here are some examples demonstrating proper formatting and reasoning in a
response:
<example>
H: Your FACT of interest is:
<fact>
[MASK] also inhibits Wnt signaling by binding to the CTNNB1 protein,
preventing interaction of CTNNB1 with TCF7L2/TCF4.
</fact>

Here is the EVIDENCE sentence to assess:
<evidence>
In the HTB-72 melanoma cell line, [MASK] inhibited melanoma cell growth
and additionally, [MASK] also induced apoptosis.
</evidence>

A:The evidence sentence makes no indication that [MASK] is involved in Wnt
signaling or anything else mentioned in the given biological fact. Therefore,
it is <answer>IRRELEVANT</answer> with <confidence>HIGH</confidence>.
</example>

<example>
H: Your FACT of interest is:
<fact>
[MASK] is a probable serine protease that plays a role in hearing.
</fact>

Here is the EVIDENCE sentence to assess:
<evidence>
type hearing loss, as occurred in subject SB114-206, caught our attention
(Figure 1a) because alterations in [MASK], a serine protease, are also known
to cause progressive ski slope{type hearing loss.
</evidence>

A:The evidence sentence states that [MASK] is a serine protease that can affect
hearing so it directly supports our biological fact being true of [MASK].
Therefore, it is <answer>RELEVANT</answer> with <confidence>HIGH</confidence>.
</example>

Now here is your real task.
Your FACT of interest is:
<fact>
{GENE_DESCRIPTION_FACT}
</fact>

Here is the EVIDENCE sentence to assess:
<evidence>
{EVIDENCE_SENTENCE}
</evidence>
```

O. Examples of Evidence Annotations for Gene Description Facts Dataset

Here we show an example of each of true positive, true negative, false positive and false negative R2E evidence Shapley values versus GPT-4 relevant/irrelevant *Evidence Annotations*, for *Gene Description Facts*. These examples also illustrate the level of biological complexity in the *Gene Description Facts* task.

1. **True positive**: R2E assigns a large, positive Shapley value to a piece of evidence that GPT-4 labels as relevant. R2E and GPT-4 recognise this as a relevant piece of evidence.

```
Query: [MASK] regulates actin dynamics in motor neurons by inhibiting
the GTP-binding activity of ARF6, leading to ARF6 inactivation.

Target: C9orf72

Evidence: The observation that the GTP-bound forms of Rac1 and Arf6 are
increased when [MASK] is repressed in primary mouse motor neurons or
in patient iPSC-derived motor neurons strongly suggests that [MASK] is
essential for the GTP exchange activity of Arf6 in particular.
(PMID: 27723745)

R2E-generated Shapley value: 0.0391

GPT4 completion: A: The evidence sentence directly supports the biological
fact. It states that when [MASK] is repressed, the GTP-bound forms of ARF6
are increased, suggesting that [MASK] is essential for the GTP exchange
activity of ARF6. This aligns with the fact that [MASK] regulates actin
dynamics in motor neurons by inhibiting the GTP-binding activity of ARF6,
leading to ARF6 inactivation. Therefore, it is <answer>RELEVANT</answer>
with <confidence>HIGH</confidence>.
```

2. **True negative**: R2E assigns a negative Shapley value to a piece of evidence that GPT-4 labels as irrelevant. R2E and GPT-4 recognise this as an irrelevant piece of evidence.

```
Query: [MASK] binds the 3'-UTR of TRPV1 mRNA and represses TRPV1
translation, which is required to maintain normal thermoception.

Target: DCTN6

Evidence: Gene co-expression analyses for SLC2A4 were conducted with
West Breast [ 28 ], and the results showed that SLC2A4 is co-expressed with
FADD, BLOC151, RHOB, [MASK], CELF2, SNTB2, NPPB, TIE1, FGFR1, IDH1, ECH1 ,
etc. (Fig. 5 D). (PMID: 34525987)

R2E-generated Shapley value: -0.0074

GPT4 completion: A: The evidence sentence only mentions that [MASK] is
co-expressed with SLC2A4 and other genes, but it does not provide any
information about [MASK] binding to the 3'-UTR of TRPV1 mRNA or
repressing TRPV1 translation. Therefore, it is <answer>IRRELEVANT</answer>
with <confidence>HIGH</confidence>.
```

3. **False positive**: R2E assigns a large, positive Shapley value to a piece of evidence that GPT-4 labels as irrelevant. R2E seems to consider the mention of magnesium concentration in relation to [MASK] sufficient to support the fact that ”[MASK] can use Mg (2+)”.

```
Query: [MASK] can use either Ca(2+) or Mg(2+) equally.

Target: AGTR2

Evidence: Magnesium concentrations were 2 mol/L, except for [MASK]
(2.5 mol/L). (PMID: 11841622)

R2E-generated Shapley value: 0.0981

GPT4 completion: A: The evidence sentence only provides information
about the magnesium concentration in relation to [MASK], but it does not
provide any information about [MASK]'s ability to use either Ca(2+) or
Mg(2+) equally. Therefore, it is <answer>IRRELEVANT</answer> with
<confidence>HIGH</confidence>.
```

4. **False negative**: R2E assigns a negative Shapley value to a piece of evidence that GPT-4 labels as relevant. R2E fails to recognize that this is a relevant piece of evidence.

```
Query: [MASK] is required for mitochondrial tubular network
and cristae organization.

Target: GHITM

Evidence: Decreased [MASK] expression and loss of cristae organisation
occur at an early stage of DIO and represent an IRF5-dependent mechanism
that may contribute to loss of microenvironmental homeostasis and
development of insulin resistance (Fig. 8 ). (PMID: 36042203)

R2E-generated Shapley value: -0.0090

GPT4 completion: A: The evidence sentence states that decreased [MASK]
expression leads to loss of cristae organization, which directly
supports the biological fact that [MASK] is required for cristae
organization. Therefore, it is <answer>RELEVANT</answer> with
<confidence>HIGH</confidence>.
```

P. Further Details on Creation of Clinical Trial Outcomes Dataset

Following the protocol described in Nelson et al. 2015 and Minikel et al. 2024, we filtered drugs from the commercial PharmaProjects (Citeline) drug development database. We filtered out drugs that were combination treatments and diagnostic drugs. In addition to the Nelson et al. 2015 filtering protocol, we applied a further temporal filtering of drugs to avoid temporal leakage (Section 4.4). Each drug was attributed an 'earliest evidence year', the earliest year that could be extracted from a mix of free-text and structured data fields in each PharmaProjects drug record. All dates were extracted from either: a "key events" field, which has well structured but heterogeneously populated dates; or free text fields giving details about preclinical, Phase I, Phase II and Phase III development or a general description of a drug's development trajectory. From the free text fields, all 4 digit date-like strings which did not occur in contexts with common failure modes were extracted using the regex `(?<=[^0-9a-zA-Z\=\%])([0-9]{4})(?=[\,\\\s\;)])(?![\s*m+g+l+])`. In brief, 4 digits, in brackets, followed by a comma, whitespace or backslash, and not subsequently followed by characters indicating quantitative measurements (namely 'm', 'g' and 'l'). Anomalous dates introduced by the regex were removed by dropping any dates that were more than 50 years from the median of the dates for a drug record. Across all of these date fields the earliest date was attributed to the drug and all indications it was tested against and used to include or exclude drugs from the analysis. The earliest development date for a drug is therefore conservative with regards the first time a drug was tested at Phase II / III for a disease. We excluded all drugs whose earliest development year was before 2005.

From the remaining drugs, we extracted therapeutic hypotheses, as described by a combination of a drug's protein targets and the diseases the drug had been tested against. We discretized therapeutic hypotheses using the PharmaProjects assigned MeSH (https://www.ncbi.nlm.nih.gov/mesh/) and Entrez (Maglott et al., 2005) ontology identifiers for the genes and diseases respectively. Nelson et al. 2015 and Minikel et al. 2024 investigate the transition between all trial phases. We use only a subset that focuses on the *in vivo* efficacy of therapeutic hypotheses. As such, we kept only the therapeutic hypotheses related to drugs tested at Phase II or III, or pre-Registration, Registration or Launched with regulatory approval. We kept only the therapeutic hypotheses where there were no drugs in active development and therefore whose clinical efficacy could be determined.

Therapeutic hypotheses that had made it to Phase II or III and have no drugs in active clinical development were assumed to have failed to demonstrate *in vivo* clinical efficacy while drugs that had made it to pre-Registration and above were said to have 'succeeded'. These are the positive and negative labels in the *Clinical Trial Outcomes* dataset.

In constructing the *Clinical Trial Outcomes* dataset we made the assumption that ceased development is indicative of a therapeutic hypothesis failing to show efficacy in a human population. We highlight that there is likely to be noise in these negative labels: drug programmes can be prosecuted or abandoned for a range of commercial reasons rather than biological ones, drug programmes may fail because sponsors failed to identify an appropriate patient population, or drug programmes may fail for pharmacological reasons peculiar to the candidate molecule.

Q. Genetics Baseline for the Clinical Trial Outcomes Dataset

Data for the genetics baseline was downloaded from the supplementary data of Minikel et al. 2024 (https://github.com/ericminikel/genetic_support/tree/sio/data) and reproduced using the methodology described in Minikel et al. 2024, briefly summarised below.

In the supplementary data, table *assoc.tsv* contains the full set of genetic associations that were templated into natural language in Section 4.4. These already-curated genetic associations were filtered further as per Minikel et al. 2024, removing all rows with a "source" of 'OTG' and an "l2g_share" < 0.5.

There exists ontological mismatch between sources of genetic evidence and diseases referenced in the PharmaProjects data. As such, the *Clinical Trial Outcomes* dataset is joined to the genetic association data by matching exactly on gene identity, and on a measure of MeSH-MeSH similarity for diseases / traits.

The table *sim.tsv.gz* contains a full list of pairwise MeSH - MeSH similarities used in this joining of datasets. The similarity measure is a composite information criterion measure of similarity on the MeSH ontology tree; see Minikel et al. 2024 for details.

The continuous score for the genetics baseline for each therapeutic hypotheses in the *Clinical Trial Outcomes* dataset is the maximum similarity to a genetics association across all the genetic association data, where 1 implies an exact disease-disease match and 0 means the there is no path between the entities in the MeSH ontology, or there is no genetic association data available for the gene anywhere in the genetic association data.

R. Further Details on Creation of Evidence Annotations for Clinical Trial Outcomes Dataset

We constructed *Evidence Annotations* for the *Clinical Trial Outcomes* dataset by having GPT-4 annotate (as relevant or irrelevant) all evidence for 50 *Clinical Trial Outcome* therapeutic hypotheses associated with trial success, as well as 50 with trial failures, both randomly sampled, resulting in 6400 query-evidence pairs (100 queries each retrieving 64 pieces of evidence) with a binary annotation.

Separately and using a similar approach, we created the dataset of evidence annotations used for auditing explanations as described in Section 4.5. In this case, we computed R2E Shapley values (computed on pre-sigmoid outputs) for all retrieved evidence over all *Clinical Trial Outcomes* dataset examples, ordered the evidence by Shapley value, and selected the 20,000 evidence sentences with highest Shapley values. We then ran relevant/irrelevant annotations on this subset using GPT-4.

We used the combined pre-2005 literature and templated genetics corpus for both tasks. Relevant/irrelevant annotations were obtained through the use of GPT-4, using the following prompt, substituting in DISEASE_OF_INTEREST and EVIDENCE_SENTENCE:

```
You are a scientific expert working on target identification in drug
```

```
discovery.

Your task is to use your expertise to evaluate a piece of evidence
(referred to as EVIDENCE) for a potential drug target for a specified
disease (referred to as DISEASE). Specifically you must indicate whether
the EVIDENCE about a masked target (in the form of a sentence from an
academic paper), provides relevant evidence that the drug target might be
promising for developing a treatment for the DISEASE. If the EVIDENCE
sentence does not make any link to the biology of the specified DISEASE,
then it is not relevant. Please explain your reasoning first before giving
your answer. Provide your final answer by stating either
<answer>RELEVANT</answer> or <answer>IRRELEVANT</answer>. Please also
indicate your confidence in your answer by writing one of
<confidence>HIGH</confidence> or <confidence>LOW</confidence>.

Here are some examples demonstrating proper formatting and reasoning in
a response:
<example>
H: Your DISEASE of interest is Sarcopenia.

Here is the EVIDENCE sentence, containing a masked target, to assess:
<evidence>
Many studies also described exercise-induced increases in transcriptional
and translational levels of FGFR1, [MASK], and/or KLB [29,33,35,36].
</evidence>

A:The evidence sentence makes no indication that [MASK] plays a role in
Sarcopenia, therefore it is <answer>IRRELEVANT</answer> with
<confidence>HIGH</confidence>.
</example>

<example>
H: Your DISEASE of interest is Amyotrophic Lateral Sclerosis.

Here is the EVIDENCE sentence, containing a masked target, to assess:
<evidence>
Therefore, further study is needed to clarify where [MASK] functions
during lysosome trafficking and neurite outgrowth.
</evidence>

A:The evidence sentence implies that [MASK] may play a role in biological
mechanisms directly related to ALS, but the phrasing is ambiguous, therefore
it is <answer>RELEVANT</answer> with <confidence>LOW</confidence>.
</example>

<example>
H: Your DISEASE of interest is Lung Adenocarcinoma.

Here is the EVIDENCE sentence, containing a masked target, to assess:
<evidence>
Conversely, [MASK] mRNA and protein expression is higher in a variety of
tumor tissues, including lung cancer [ 7 , 20 ].
</evidence>
```

```
A:The evidence sentence states that [MASK] expression is upregulated in tumor
tissues including our specific cancer type of lung cancer, therefore it is
<answer>RELEVANT</answer> with <confidence>HIGH</confidence>.
</example>


Now here is your real task.
Your DISEASE of interest is {DISEASE_OF_INTEREST}.

Here is the EVIDENCE sentence, containing a masked target, to assess:
<evidence>
{EVIDENCE_SENTENCE}
</evidence>
```

S. Examples of Evidence Annotations for Clinical Trial Outcomes Dataset

Here we show an example of each of true positive, true negative, false positive, and false negative R2E evidence Shapley values versus GPT-4 relevant/irrelevant *Evidence Annotations*, for *Clinical Trial Outcomes* data.

1. **True positive**: R2E assigns a large, positive Shapley value to a piece of evidence that GPT-4 labels as relevant. R2E and GPT-4 recognise this as a relevant piece of evidence.

```
Query: [MASK] is a promising drug target for the treatment of
myelodysplastic syndromes.

Target: BIRC2

Evidence: [MASK] protein expression was found to be present in a majority
of leukemic blasts from untreated patients, but absent in normal
differentiating myeloid progenitor cells. (PMID: 14666661)

R2E-generated Shapley value: 0.4057

GPT4 completion: A: The evidence sentence indicates that [MASK] protein is
expressed in leukemic blasts, which are abnormal cells associated with
leukemia, a type of blood cancer. Myelodysplastic Syndromes (MDS) are a
group of disorders caused by poorly formed blood cells or ones that don't
work properly, which can often progress to leukemia. The fact that [MASK]
is not present in normal myeloid progenitor cells, which are cells that
give rise to all types of blood cells, suggests that it could be involved
in the abnormal development of these cells in MDS. Therefore, it is
<answer>RELEVANT</answer> with <confidence>HIGH</confidence>.
```

2. **True negative**: R2E assigns a negative Shapley value to a piece of evidence that GPT-4 labels as irrelevant. R2E and GPT-4 recognise this as an irrelevant piece of evidence.

```
Query: [MASK] is a promising drug target for the treatment of crohn
disease.

Target: LANCL2

Evidence: Because of the structural similarity to LanC, we postulate that
[MASK] may play a role as a component of a peptide-modifying complex.
(PMID: 11762191)
```

```
R2E-generated Shapley value: -0.0160

GPT4 completion: A: The evidence sentence suggests that [MASK] may be
involved in peptide modification, but it does not provide any direct link
or relevance to Crohn Disease. Therefore, it is <answer>IRRELEVANT</answer>
with <confidence>HIGH</confidence>.
```

3. **False positive**: R2E assigns a large, positive Shapley value to a piece of evidence that GPT-4 labels as irrelevant. R2E seems to misunderstand an abbreviation that GPT-4 is able to correctly parse.

```
Query: [MASK] is a promising drug target for the treatment of chronic
myelomonocytic leukemia.

Target: MAP3K1

Evidence: This observation indicates that [MASK] may be a useful
therapeutic target in the treatment of CML. (PMID: 14586403)

R2E-generated Shapley value: 0.2329

GPT4 completion: A:The evidence sentence suggests that [MASK] could
be a potential therapeutic target for Chronic Myeloid Leukemia (CML).
However, the disease of interest is Chronic Myelomonocytic Leukemia,
not CML. Despite both being types of leukemia, they are distinct diseases
with different pathologies. Therefore, the evidence is
<answer>IRRELEVANT</answer> with <confidence>HIGH</confidence>.
```

4. **False negative**: R2E assigns a negative Shapley value to a piece of evidence that GPT-4 labels as relevant. R2E fails to recognize that this is a relevant piece of evidence.

```
Query: [MASK] is a promising drug target for the treatment of heart
failure.

Target: RXFP1

Evidence: However, a recent study has identified the relaxin receptor as a
G-protein-coupled orphan receptor, [MASK], that can couple to adenylate
cyclase to increase cAMP ( Hsu et al ., 2002 ). (PMID: 12381685)

R2E-generated Shapley value: -0.0167

GPT4 completion: A: The evidence sentence suggests that [MASK] is a
G-protein-coupled orphan receptor that can couple to adenylate cyclase
to increase cAMP. This is relevant to heart failure as G-protein-coupled
receptors and cAMP signaling pathways are known to play crucial roles in
cardiac function and heart failure pathology. Therefore, it is
<answer>RELEVANT</answer> with <confidence>HIGH</confidence>.
```

T. Details on Cleaning MeSH Terms for Templating

The canonical MeSH name for each MeSH identifier is used in templates for generating R2E queries in the *Clinical Trial Outcomes* evaluation and generating genetics-derived sentences for augmenting the literature evidence.

The MeSH names underwent light reformatting to align them to how they might be expressed in natural language. Simply, we lowercase the MeSH name, split on commas, and reverse order the resulting list. So for example the MeSH name *Leukemia, Myelomonocytic, Chronic* becomes *chronic myelomonocytic leukemia*.

U. Further Results on Predicting Clinical Trial Outcomes and Genetic Evidence

U.1. Relative Success

For a given prediction threshold, we compute relative success of model predictions as:

$$RS = \frac{(\text{True Positive}/\text{Predicted Positive})}{(\text{False Negative}/\text{Predicted Negative})} \tag{16}$$

Where relevant, we use Katz method (Katz et al., 1978) for confidence intervals and Z-test for comparisons.

U.2. Results for Diseases with Genetic Insight

Previous analyses of genetic methods for target identification have restricted to evaluating only on diseases with at least one piece of genetics data and for which therefore genetics could be expected to be informative (those with 'genetic insight') (Minikel et al., 2024). In Minikel et al. (2024), diseases were deemed to have genetic insight if there was at least one genetic association between a gene and disease with a MeSH-MeSH similarity of > 0.7. This subsetting of therapeutic hypotheses was used to obtain the widely published relative success of $\sim$2 in predicting clinical trial outcome success from genetic data.

We validated our *Clinical Trial Outcomes* dataset by corroborating this result by similarly restricting post-2005 therapeutic hypotheses to diseases with genetic insight, and using a MeSH-MeSH similarity threshold of $>$0.8 as the threshold for positive predictions as per Minikel et al. 2024. At this threshold, the genetics baseline makes 500 positive predictions across the 4,056 therapeutic hypotheses, with a Relative Success of 1.98, 95% CI (1.76, 2.24). In comparison, R2E-cor predicting on literature obtained a relative success of 2.17 (95 % CI (2.44, 1.93)) making the same number of positive predictions.

For completeness, we also show AUROC results after restricting to diseases with genetic insight in Table S7, with trends in AUROC similar to the results without restriction shown in the main text - rationale for the latter below (Appendix U.3).

Table S7. **Clinical Trial Outcomes on therapeutic hypotheses with genetic insight**: AUROC for R2E retrieving from literature-alone, genetics-alone, or both; in comparison to baselines, when subsetting therapeutic hypotheses just to those where the disease has at least one genetic association in the genetics baseline.

MODEL	CORPUS	AUROC
GENETIC	GENETICS	0.588
FREQ	LITERATURE	0.552
MCS	LITERATURE	0.634
MLM	LITERATURE	0.638
R2E-UNCOR	GENETICS	0.618
R2E-UNCOR	LITERATURE	0.636
R2E-COR	LITERATURE	0.643
R2E-COR	BOTH	**0.647**
R2E-AUDIT	BOTH	**0.651**

U.3. Results for All Diseases

When comparing to predictions using literature evidence, restricting to diseases with genetic insight as described above, would undervalue literature as an evidence source; literature can be expected to be informative about a wider range of diseases. Therefore, for AUROC results in the main text (2) we instead show performance against all diseases in the *Clinical Trial Outcomes* data, without restriction to those with genetic insight.

In Figure S4, we show the relative success for a given number of positive predictions for each model, by varying thresholds for each model. The relative success of the genetics baseline is below that of all R2E models using literature evidence, across all model thresholds (S4), as well as largely below the R2E model using genetics-evidence only. As expected, compared to when restricting to diseases with genetic insight (Appendix U.2), the genetics baseline (using the same >0.8 threshold) has a lower relative success (1.72, 95% CI (1.54, 1.93)) when predicting for all diseases.

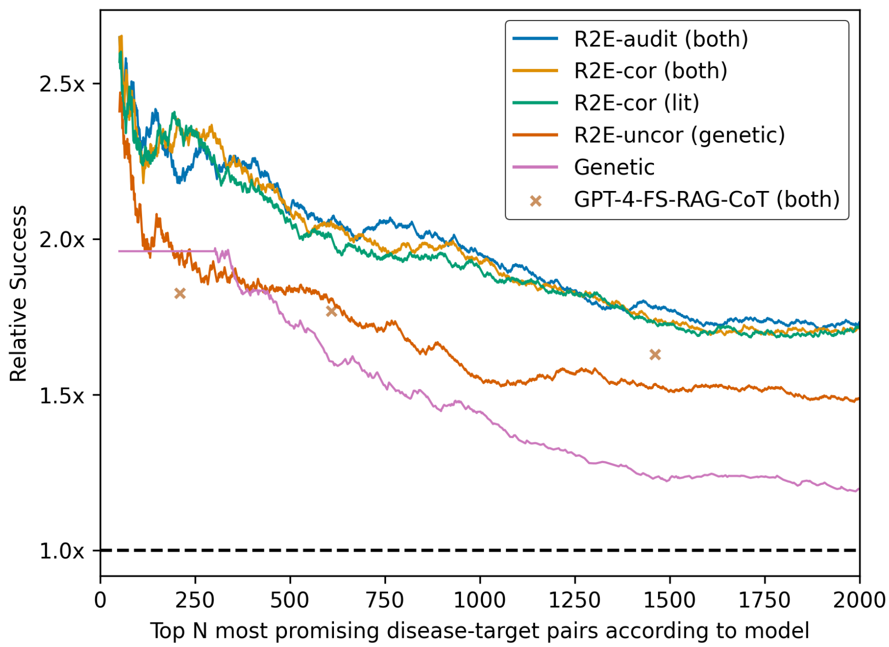

Figure S4. **R2E Relative Success on Clinical Trial Outcomes**. Relative success for a given number of positive predictions (x-axis) for each model. The different numbers of positive predictions was achieved by varying the threshold for a positive prediction for each model.

U.4. GPT-4-FS-RAG-CoT Baseline

The few-shot, chain-of-thought prompted GPT-4 baseline with retrieval augmentation (see Appendix V for details of setup) had lower relative success than all R2E models using literature evidence, at all thresholds (Figure S4). When matching thresholds to obtain 609 positive predictions, R2E-cor (both) (relative success: 2.05; 95% CI (1.86, 2.26)) significantly outperformed the GPT-4 baseline (relative success: 1.77; 95% CI (1.59, 1.97)) using the same evidence (Z test, $p = 0.043$).

U.5. Results with Forced R2E Retrieval of Genetics

Table S8 shows that in the multi-modal context (with a corpus of sentences from the biomedical literature and from the genetics data), forcing retrieval of genetics evidence does not change the AUROC.

Table S8. **Methods of multimodality for Clinical Trial Outcomes**: AUROC for R2E-uncor and R2E-cor with three different methods of multi-modalility: (1) Retrieve from a single corpus containing both genetics and literature sentences (single index); (2) Retrieve up to four sentences from the genetics corpus - where possible - and retrieve the remaining sentences from the literature corpus (separate index); and (3) R2E scores evidence from the genetics and the literature corpora separately and the final score is the mean of the two (post-hoc aggregation).

Model	Corpus	Method	AUROC
R2E-uncor	Both	Single index	0.631
R2E-cor	Both	Single index	0.633
R2E-uncor	Both	Separate index	0.631
R2E-cor	Both	Separate index	0.633
R2E	Both	Post-hoc aggregation	0.633

U.6. Performance by Disease Area

Figure S5 shows that there is substantial variation in performance across disease areas and modality. The variability is especially pronounced for the genetics baseline and R2E using only genetics-evidence, consistent with the reduced disease coverage of genetics compared to the literature. The magnitude of difference in performance between R2E retrieving from genetics alone and R2E retrieving from literature, varies by disease area. This may indicate disease areas for which alternative predictive modalities to genetics might be being represented in the literature.

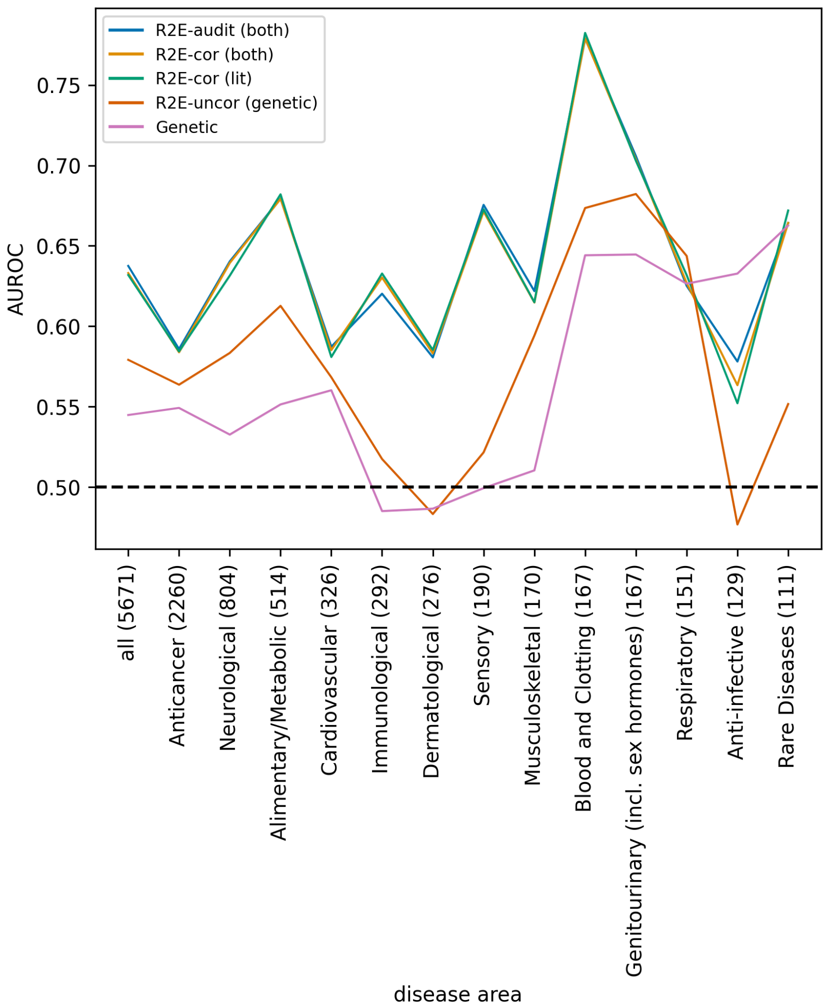

Figure S5. **R2E performance across disease areas**. AUROC in each PharmaProjects annotated disease area with more than 100 therapeutic hypotheses. Predictions by R2E retrieving from literature-alone (R2E-cor (lit)), genetics-alone (R2E-uncor (genetic)), both genetics and literature (R2E-cor (both)), or genetics and literature with LLM auditing (R2E-audit (both)); in comparison to the genetics baseline (Genetic). The number of therapeutic hypotheses for each disease area are given in brackets.

U.7. R2E Benefits from Soft Semantic Matching

In evidence auditing experiments detailed in Section 4.5, where high Shapley value evidence sentences were annotated by GPT-4 as relevant or irrelevant to the given query, 527/809 of the annotated genetics sentences were annotated as relevant, evidencing that R2E can appropriately leverage genetic evidence. Note that 268 of these 527 genetic evidence sentences was related by R2E to a disease that was neither a substring of, nor contained, the *Clinical Trial Outcomes* disease.

Figure S6 shows the distribution of MeSH-MeSH ontological similarity, between the clinical trial disease and the genetics evidence disease / trait, as calculated by Minikel et al. 2024, for these 527 relevant-annotated genetic query-evidence pairs with high Shapley scores (Section 4.5). Note that when calculating relative success in Minikel et al. 2024, the threshold MeSH-MeSH similarity for positively linking between therapeutic hypotheses and genetic association data was 0.8. By contrast, we observed that R2E can also perform “soft” semantic matching between the query and the genetics evidence. For example, R2E picked up on the following trait-trait pairs with a MeSH similarity < 0.2: (erythrocyte count and anemia), (eosinophilia and asthma), (astrocytoma and brain neoplasms). This highlights the shortcomings of a universal threshold based on ontological similarity metrics (genetics baseline) versus semantic matching and reasoning through natural language (R2E).

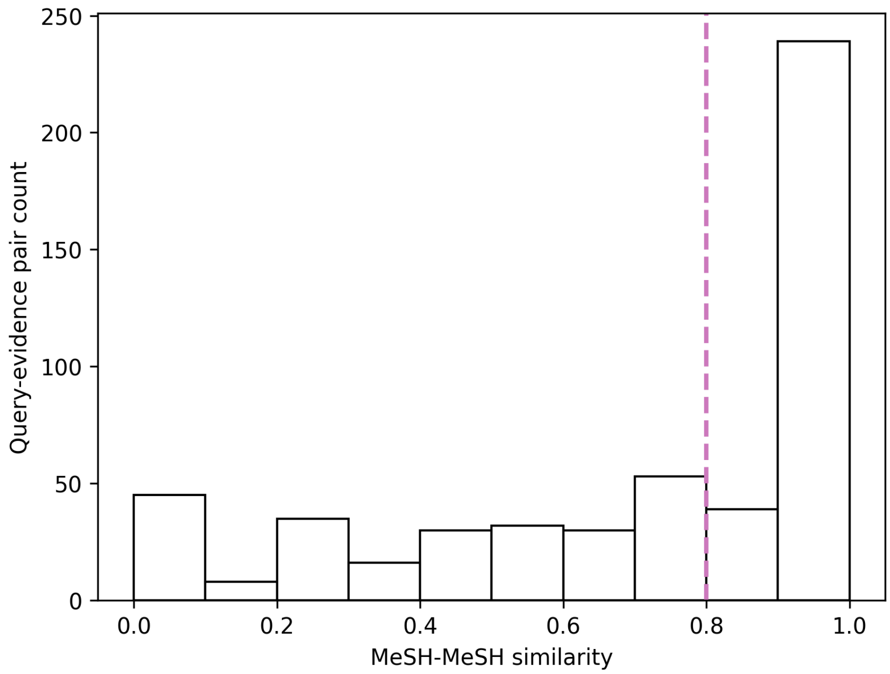

Figure S6. **MeSH-MeSH ontological similarity of query-genetics evidence pairs with high Shapley scores**. Histogram showing the distribution of similarities between MeSH terms for diseases / traits in genetics evidence annotated as “Relevant” by GPT-4, and diseases in the clinical trial therapeutic hypothesis queries, when the evidence had a high Shapley score (Section 4.5). The dashed line indicates the threshold used in the Minikel et al. 2024 genetics baseline to assign genetic evidence to a therapeutic hypothesis.

V. A Few-Shot, Chain-of-Thought, RAG Baseline using GPT-4 on Clinical Trial Outcomes

V.1. Existing LLMs and Retrieval Augmented Generation

We first detail why in general, to the best of our knowledge, generative LLMs such as GPT-4, do not solve the problem we seek to address with R2E, of explainable multi-label prediction from textual data, i.e.:

- Score and rank each answer in the answer set
- Faithfully attribute the score to evidence passages

With access to token probabilities, one option could be to constrain LLM completions to synonyms of entities and compute completion probabilities to rank entities. For explainability one might then consider retrieval augmented generation (RAG). However, question-answering-style generative explanations are often not faithful and are not quantitative - they do not assess the direct, quantitative impact of a piece of evidence on the model score, and they risk hallucination. Additionally, running LLMs with separate gene-specific prompts in a RAG setup, for each of 19,176 genes for every query, would be prohibitively expensive.

For models restricted to API access only, such as GPT-4, it is not possible to use completion probabilities to rank genes. Prompting to directly generate a ranked list of targets returns well-known targets and omits explainability. Using a RAG approach for each gene independently, one could have the model specify a score to partially rank targets, or at least identify a subset of targets the LLM deems promising. However, once again this does not ensure faithful explainability and would be similarly prohibitively expensive for ranking 19,176 genes for every query.

V.2. Setup for Comparison to GPT-4 Baseline

Despite the points raised above on prohibitive expense (Appendix V.1), in the case of *Clinical Trial Outcomes*, since the evaluation only requires prediction on individual disease-target pairs, rather than full rankings of all genes for each query, a direct comparison to the latter RAG-based approach using GPT-4 is possible. Despite the described lack of faithful explainability and the practical inappropriateness of this approach to the problem addressed by R2E, for academic interest only we include a comparison. We also highlight that since it is not possible to use a GPT-4 model only trained on pre-2005 data, it may be advantaged in comparison to R2E when evaluating on our dataset of *Clinical Trial Outcomes* from 2005 onwards.

Specifically, we use a chain-of-thought, few-shot prompted GPT-4 in a RAG style setup. For each disease-target pair we provide the same evidence set of up to 64 sentences as seen by R2E, and have GPT-4 predict whether the evidence supports the masked target as promising or not for developing a treatment for the given disease, as well as a level of confidence in the prediction out of very low, low, medium, high, very high. We summarise the findings in Appendix U.4, and show the results in terms of relative success in Figure S4 with the following 3 different cutoffs used to determine positive predictions, corresponding to the three data points shown in the figure:

- At least very low confidence in the target being promising
- At least high confidence in the target being promising
- At least very high confidence in the target being promising

V.3. Prompting for GPT-4 Baseline

The following few-shot, chain-of-thought, retrieval-augmented prompt was used for the GPT-4 baseline, where we substitute DISEASE OF INTEREST and EVIDENCE SENTENCES for the particular evaluation query. The PMIDs included inline in this prompt are not passed to GPT-4, but are included in order to properly reference these works in this manuscript.

```
You are a scientific expert working on target identification in drug
discovery.

Your task is to use your expertise to evaluate whether a potential drug
target could potentially be promising for a given disease (referred to as
DISEASE). You must make your evaluation based on a provided set of evidence
about the drug target (referred to as EVIDENCE), identifying if any of the
EVIDENCE could directly or indirectly suggest the target could be promising.

Please explain your reasoning first before giving your answer.

Provide your final answer by stating either <answer>PROMISING</answer> or
<answer>NOT PROMISING</answer>.

Please also indicate your confidence in your answer by writing one of:
- <confidence>VERY HIGH</confidence>
- <confidence>HIGH</confidence>
- <confidence>MEDIUM</confidence>
- <confidence>LOW</confidence>
- <confidence>VERY LOW</confidence>.

Note that the name of the target will be hidden in the EVIDENCE set. Mentions
of the target have been replaced with '[MASK]'. This is because you should
```

make your prediction based on the evidence itself, not based on the particular target.

Here are some illustrative examples of the task demonstrating proper formatting and reasoning in a response.

<example>

TASK: Your DISEASE of interest is lung adenocarcinoma.

Here is the set of EVIDENCE about the target:
<evidence>
1. Thus, [MASK] is also a novel prognostic biomarker and therapeutic target for NSCLC. [PMID: 36215859]
2. In the present study, we demonstrated that [MASK] was significantly upregulated in tumor tissues and associated with poor clinical prognosis of NSCLC. [PMID: 32855383]
3. As expected, EMT-related gene sets were significantly enriched in the [MASK]-high expression phenotype, suggesting that [MASK] may contribute to TGF-β-induced EMT of NSCLC cells (Supplementary Fig. S6A). [PMID: 32855383]
4. [MASK] is highly expressed in non small cell lung cancer tissues and is associated with poor prognosis. [PMID: 32167655]
5. Collectively, this study supports that [MASK] is a key regulator in IL-6/JAK2/STAT3 axis and mediates EGFR inhibitor resistance in lung adenocarcinoma. [PMID: 36990047]
6. The cell growth was suppressed after [MASK] was knocked out in established PC9 sg-[MASK] cells, which confirmed that [MASK] is essential for cell survival of NSCLC (Figure S1I). [PMID: 31607564]
</evidence>

Please evaluate whether any of the provided EVIDENCE suggests that the target could be promising for lung adenocarcinoma.

RESPONSE: [MASK] is increased in NSCLC (a subtype of lung adenocarcinoma) tissues and is associated with poor survival (EVIDENCE: 2, 4). [MASK] is linked to relevant NSCLC mechanisms including EMT and EGFR resistance (EVIDENCE: 3, 6). There is also supporting assay data to suggest inhibiting [MASK] would prevent NSCLC cell growth (EVIDENCE: 1).

Conclusion: <answer>PROMISING</answer>.
Confidence: <confidence>VERY HIGH</confidence>.
</example>

<example>

TASK: Your DISEASE of interest is multiple sclerosis.

Here is the set of EVIDENCE about the target:
<evidence>
1. Given, that [MASK] does not have functionality in the MHC-II antigen presenting pathway, it is possible that [MASK] promotes MS pathogenesis via inflammasome activation. [PMID: 30817945]
2. [MASK] is an emerging pharmacological target for cancer immunotherapy and the control of inflammatory autoimmunity, including rheumatic conditions such

```
as AS (36, 37). [PMID: 33617882]
3. A Functional Variant in [MASK] Predisposes to Multiple Sclerosis
[PMID: 22253828]
4. In the light of foregoing discussion [MASK] can be envisaged as a relevant
target for prevention and treatment of autoimmune diseases. [PMID: 36740089]
5. In this study, we present, to our knowledge, the first mechanistic studies
performed to uncover why polymorphisms in [MASK] are associated with increased
susceptibility to MS. [PMID: 34810226]
</evidence>
Please evaluate whether any of the provided EVIDENCE suggests that the target
could be promising for multiple sclerosis.

RESPONSE: There is a possible mechanistic link from [MASK] to multiple
sclerosis pathogenesis via inflammasome activation (EVIDENCE: 1). [MASK]
is genetically linked to multiple sclerosis (EVIDENCE: 3), which is
potentially via a mechanistic function (EVIDENCE: 5). [MASK] has been
described as a therapeutic target for similar autoimmune diseases
(EVIDENCE: 2, 4).

Conclusion: <answer>PROMISING</answer>.
Confidence: <confidence>HIGH</confidence>.
</example>

<example>

TASK: Your DISEASE of interest is idiopathic pulmonary fibrosis.

Here is the set of EVIDENCE about the target:
<evidence>
1. The antimicrobial peptide YD attenuates inflammation via miR-155 targeting
[MASK] during liver fibrosis. [PMID: 33532183]
2. Although [MASK]-/- mice reacted similarly to WT mice when allowed to
recover from an acute DSS-induced injury ( Figure 1) and exhibited signs of
improved repair ( Figure 2), they had an increased inflammatory response
compared to WT animals ( Figures 5A and 5B ). [PMID: 20226691]
3. Consistent with their response to acute DSS treatment and their enhanced
tissue repair phenotype, [MASK]-/- mice were more resistant to chronic colitis
compared to WT animals, gaining weight by the end of the experiment as compared
to WT mice that lost 5% of their initial body weight ( Figure 5C).
[PMID: 20226691]
4. [MASK]-/- mice showed a comparable phenotype to WT mice in the acute model
of DSS colitis, but expressed an increased mortality when DSS exposure was
prolonged to 15 days. [PMID: 20346770]
5. Altogether, these data suggested that [MASK]-/- mice have an increased
ability to recruit macrophages, which leads to increased production of
inflammatory and tissue repair factors. [PMID: 20226691]
6. Yan et al. [ 301 ] recently reported that the anti-fibrotic properties of
AMP YD were mediated through the miR-155/[MASK]/NF-kB pathway.
[PMID: 34496967]
7. [MASK] is an inhibitor of caspase 1, and Dupaul-Chicoine et al . showed
that [MASK] -/- mice are resistant to acute and chronic (but not sustained)
DSS-induced colitis [PMID: 20425920]
</evidence>
```

```
Please evaluate whether any of the provided EVIDENCE suggests that the target
could be promising for idiopathic pulmonary fibrosis.

RESPONSE: The evidence largely points to [MASK] having a role in inflammation
rather than specifically fibrosis (EVIDENCE: 2, 3, 4, 5, 7). None of the
evidence specifies that [MASK] is expressed in the lung and none of the
evidence provides direct support for the role of [MASK] in IPF. However, there
is indirect linking of [MASK] to fibrosis via an indirect mechanism in a
different disease context to idiopathic pulmonary fibrosis (EVIDENCE: 1, 6).

Conclusion: <answer>NOT PROMISING</answer>.
Confidence: <confidence>LOW</confidence>.
</example>

Now here is your real task.

Your DISEASE of interest is {DISEASE OF INTEREST}.

Here is the set of EVIDENCE about the target:
<evidence>
{EVIDENCE SENTENCES}
</evidence>

Please evaluate whether any of the provided EVIDENCE suggests that the target
could be promising for {DISEASE OF INTEREST}.
```

W. Examples of Auditing of Evidence for Clinical Trial Outcomes Dataset

Here we show examples of query-evidence pairs that GPT-4 annotated as irrelevant and to which R2E assigned a large positive Shapley value, as identified during the auditing experiments detailed in Section 4.5. With each example we also report the overall R2E prediction score for the associated query-target pair before and after the auditing process. The change in score from before to after the auditing process accounts for the removal of the specified “false positive” (according to GPT-4) piece of evidence, as well as any other query-evidence “false positive” pairs removed by the audit. We also include the final clinical trial status of the associated query-target pair in each example.

1. **Negative evidence example**: GPT-4 correctly identifies the negation in the evidence sentence, which suggests that there is not good evidence for the involvement of [MASK] in Tourette’s syndrome. R2E fails in this case and assigns a large positive Shapley value to this irrelevant piece of evidence.

```
Query: [MASK] is a promising drug target for the treatment of tourette
syndrome.

Target: HTR1A

Clinical Trial Outcome: Failure

R2E score pre-audit: 0.9831

R2E score post-audit: 0.9395

Evidence: In conclusion, our study does not provide evidence that the
[MASK] gene plays either a major or a minor role in the genetic
predisposition to schizophrenia, bipolar affective disorder, or Tourette’s
syndrome. (PMID: 8546152)
```

```
R2E-generated Shapley value: 0.0371

GPT4 completion: A:The evidence sentence explicitly states that the [MASK]
gene does not play a major or minor role in the genetic predisposition to
Tourette's syndrome, therefore it is <answer>IRRELEVANT</answer> with
<confidence>HIGH</confidence>.
```

2. **Failed entity linking example**: R2E incorrectly understands ”ICH” to refer to the disease of interest ”ichytosis”, which is a skin disorder, and therefore assigns a large Shapley value to this irrelevant piece of evidence. GPT-4 correctly infers that ”ICH” refers here to intracerebral hemorrhage based on the surrounding context in the evidence sentence, and is therefore irrelevant to ichytosis.

```
Query: [MASK] is a promising drug target for the treatment of ichthyosis.

Target: TNF

Clinical Trial Outcome: Success

R2E score pre-audit: 0.8920

R2E score post-audit: 0.7040

Evidence: [MASK] likely plays a central and pathogenic role during ICH,
because antisense oligonucleotides that target [MASK] reduced [MASK]
protein levels and neuronal cell death within the hematoma as well as
neurobehavioural deficits following ICH. (PMID: 11409424)

R2E-generated Shapley value: 0.1088

GPT4 completion: A: The evidence sentence suggests that [MASK] plays a
central and pathogenic role during ICH, and that reducing [MASK] protein
levels can alleviate symptoms. However, the ICH referred to in the
sentence is likely intracerebral hemorrhage, not ichthyosis. Therefore,
the evidence is <answer>IRRELEVANT</answer> with
<confidence>HIGH</confidence>.
```

3. **Incorrect understanding example**: R2E assigns a large positive Shapley value to a piece of evidence that seems to have no relevance (perhaps reasoning incorrectly based on the shared ”retin” token in the disease name and the compound name). GPT-4 correctly identifies the evidence as irrelevant.

```
Query: [MASK] is a promising drug target for the treatment of retinitis
pigmentosa.

Target: NGFR

Clinical Trial Outcome: Failure

R2E score pre-audit: 0.6419

R2E score post-audit: 0.3103

Evidence: [MASK] was solubilized in methylene chloride and fluorinated
using diethylaminosulfurtrifluoride (DAST) as the fluorinating agent
```

```
(Fig. 1) asdescribed for the synthesis of all- trans-retinoyl-fluoride
[25]. (PMID: 11762191)

R2E-generated Shapley value: 0.3807

GPT4 completion: A: The evidence sentence describes a chemical process
involving [MASK], but does not link this process to the biology or
pathology of Retinitis Pigmentosa. Therefore, it is
<answer>IRRELEVANT</answer> with <confidence>HIGH</confidence>.
```

4. **Indirect evidence example**: GPT-4 is arguably too restrictive in its definition of relevance here. B-cell chronic lymphocytic leukemia (B-CLL) and mantle-cell lymphoma share enough significant disease biology that this is still a potentially relevant piece of evidence. This bias against indirect evidence was an issue we saw in several auditing examples we examined, and could be mitigated by adjusting the prompting strategy to GPT-4 in future work.

```
Query: [MASK] is a promising drug target for the treatment of mantle-cell
lymphoma.

Target: TNFRSF10B

Clinical Trial Outcome: Failure

R2E score pre-audit: 0.5400

R2E score post-audit: 0.0381

Evidence: Cells from 10 B-CLL patients previously studied by flow
cytometry and from an additional three B-CLL patients (L-244, L-243 and
L-249) were also positive for [MASK] protein by Western blot analysis
( Figure 3 ). (PMID: 11753607)

R2E-generated Shapley value: 0.5914

GPT4 completion: A: The evidence sentence indicates that [MASK] protein is
present in B-CLL (B-cell Chronic Lymphocytic Leukemia) patients. However,
B-CLL is a different type of lymphoma than Mantle-Cell Lymphoma. While they
are both types of B-cell lymphomas, they have different characteristics and
may not share the same drug targets. Therefore, the evidence does not
directly link [MASK] to Mantle-Cell Lymphoma. It is
<answer>IRRELEVANT</answer> with <confidence>HIGH</confidence>.
```