waterworksAI
pharma

PharmaDS 2025

GenAI Content in Pharmaceutical Development: Exploring Methods and Techniques for Output Evaluation.

7 April 2025
Crown Plaza
Edison, NJ

www.waterworksAI.com

Meet Today’s Instructors

Shahin Samadi, MS
Computer Science
Applied and Computational Mathematics
Principal Machine Learning Engineer

- 10+ Years AI/ML
- Health AI Expertise
- Adversarial AI Systems

PLEASANT LAND

Rebecca D. Jones-Taha, PhD, MBA
Founder & CEO

- 20+ years drug development experience
- Deep biometrics expertise
- AI + pharma thought leadership

Lilly

ICON

Alec Shamula, MS
Artificial Intelligence
Lead Modeling, Simulation, & Analysis Engineer

- 6+ Years AI/ML Development
- Reinforcement Learning Expertise

NORTHROP GRUMMAN

OFFICE OF NAVAL INTELLIGENCE
1882

Today’s Agenda

1. Course Objectives

   &

   Introduction

2. Fundamentals of Generative AI in Pharma

3. Context and Planning for Evaluation

4. Traditional & Domain-Specific Evaluation Metrics

5. Special Topics

6. Resources, Wrap-up, and Q & A

Course Objectives

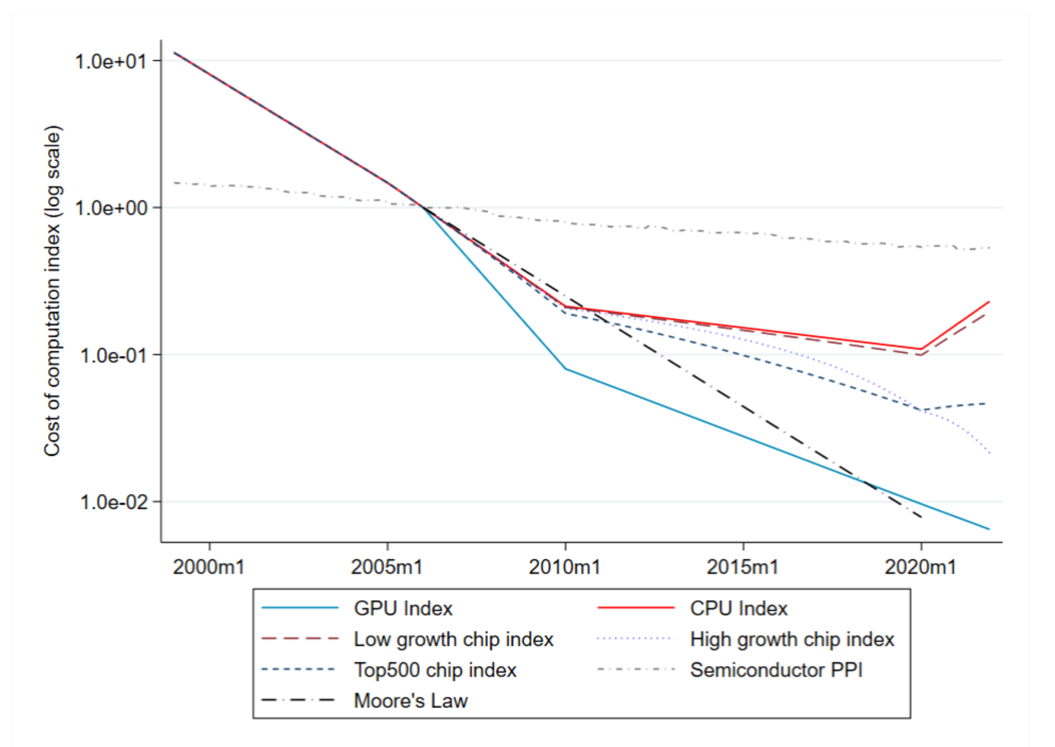
Understand	Learn	Gain	Apply
Understand core quantitative metrics and evaluation goals for AI-generated text.	Learn how to plan for and evaluate the quality of generated textual output.	Gain access to templates, scripts, examples, and more.	Create an evaluation plan and apply evaluation methods with context awareness.

Let’s Get Started.

Rapid evolution has created major improvements in gen AI technologies, with new considerations.

GenAI Advancements: Decreasing computing costs, improved accuracy, reduced latency, growing adoption

Emerging Complexities: Evolving regulatory landscape, organizational governance, maintenance of quality and accuracy, evaluation



Applications for generative AI in pharm development touches nearly every development activity.

Drug Discovery

Human Resources

Synthetic Data

Study/Program

Clinical trial protocol, SAPs, Informed consent documents, etc.

Patient

Patient narratives, diagnostics, image summary generation

Other use cases:

Medical literature summarization, regulatory submission drafts

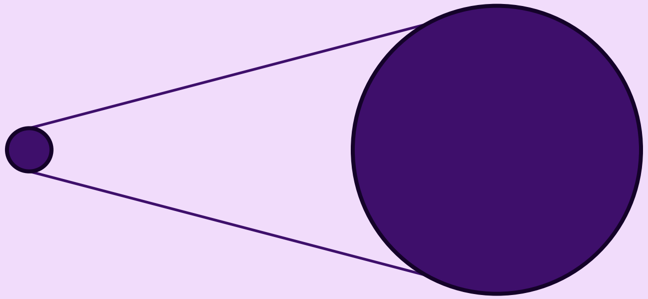
Ever expanding list…

Investigator Brochure

Safety Analytics

Competitive Landscape

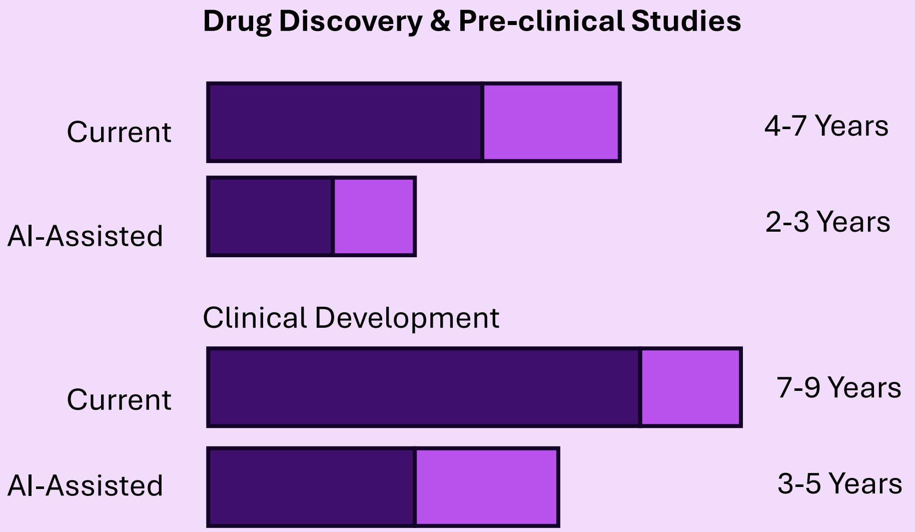
Artificial intelligence will reduce clinical development time by half.



2 Billion, 2022

22 Billion, 2027

Global pharmaceuticals’ AI budgets*



89%

■ Minimum
■ Maximum

Once in a generation value creator.

*Source: Boston Consulting Group

Large pharma is investing and already deriving substantial benefit from AI implementation.

Lilly

By August 2024
1.4 million hours to double over 4 months to 2.4 million (274 years of human work equivalent)

Pfizer

AI and ML capability are used in more than 50% of all Pfizer's clinical trials.

sanofi

We're using AI to chase the miracles of science by accelerating drug discovery, enhancing clinical trial design, and improving the manufacturing and supply of medicines and vaccines

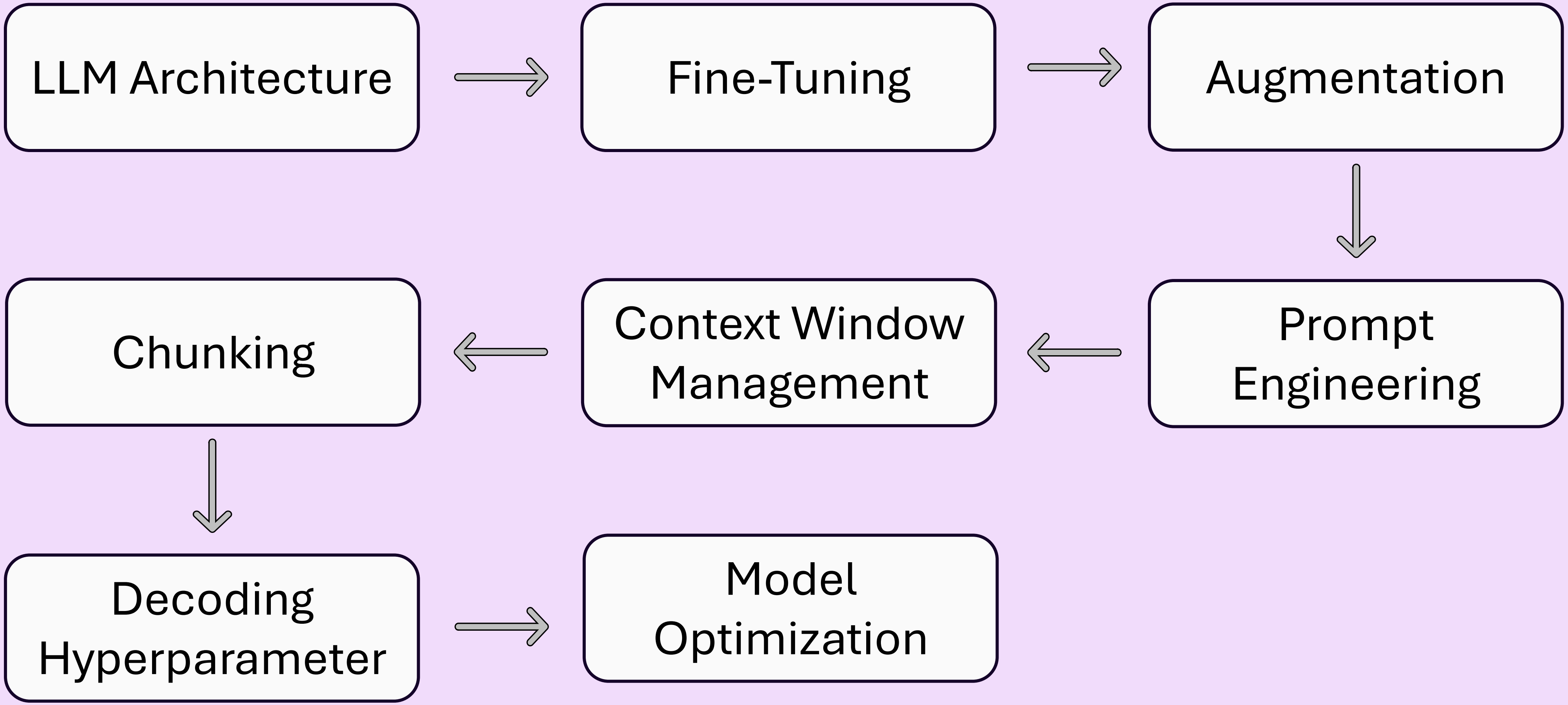
WE ARE TRANSITIONING TO AN AI-DRIVEN INDUSTRY.

THE WAY WE WORK IS CHANGING FOR GOOD.

Given the breadth of use of AI in pharma development, how can we have confidence in results?

Not so fundamental fundamentals.

Factors that Determine Your AI's Behavior



LLM Architecture

Fine-Tuning

Augmentation

Prompt Engineering

Context Window Management

Chunking

Decoding Hyperparameter

Model Optimization

LLM Architecture – How model design impacts performance, speed, and input types.

Dense vs Sparse

Dense

What is it?
Every token interacts directly with every other token, creating full attention connections throughout the entire model.

Why it Matters?
Offers deep contextual understanding and high accuracy at the expense of computational speed, especially for longer data sequences or large datasets.

Use Cases:

- Precise extraction and summarization of medical literature
- In-depth analysis of clinical guidelines or regulatory documents

Popular Examples:
GPT-4 (OpenAI), Llama 3 (Meta)

Sparse

What is it?
Attention selectively connects only certain tokens, limiting interactions to reduce computational complexity.

Why it Matters?
Enhances scalability and speed, enabling models to efficiently manage very large datasets or lengthy documents without a major sacrifice in context-awareness.

Use Cases:

- Efficient processing of large-scale clinical trial documentation
- Screening large scientific databases or repositories quickly

Popular Examples:
Big Bird (Google), Sparse Versions of Llama 3

LLM Architecture – How model design impacts performance, speed, and input types.

MoE & Modality

Mixture of Experts (MoE)

What Is It?
A model architecture where an internal router decides which subset of experts within the model should activate for each input.

Why It Matters?
Optimizes computational efficiency by activating only the needed parts of the model. It scales exceptionally well, balancing speed, efficiency, and capability.

Use Cases:

- Specialized predictive modeling for side-effect profiles
- High-performance analytics in across multi-domain material

Popular Examples:
Phi MoE (Microsoft), Mixtral (Mistral AI)

Model Modality

What Is It?
Modality describes the range and types of inputs an LLM can handle, such as text, images, audio, or sensor data.

Why It Matters?
Handling multiple data types simultaneously allows for more comprehensive understanding and richer context, despite increased computational requirements.

Types:

- Single Modality: Models accepting only one type of input (text-only, image-only).
- Multi-Modality: Models handling combinations of data like text, imaging, sensor data, and audio simultaneously.

Fine Tuning – Tailoring the model's behavior to your specific needs.

Tuning Overview

Full Fine-Tuning → Task-Specific Datasets = { Adapt fully to new data/ task }

Parameter Efficient Fine-Tuning → Task-Specific Datasets = { Efficiently adapts but retains general knowledge }

Instruction or Alignment Tuning → Behavior-Specific Datasets = { Aligns model behavior to new data/ task }

Fine Tuning – Tailoring the model's behavior to your specific needs.

Which Tuning Do You Need?

Full Fine-Tuning

Why It Matters?
Fully updates every model parameter, enabling deep specialization in a particular domain or task at the expense of generalizability.

Use Cases:

- Highly detailed and accurate interpretation of a subject domain.

Parameter Efficient Fine-Tuning

Why It Matters?
Selectively updates minimal parameters, efficiently tailoring models while preserving their broader knowledge.

Use Cases:

- Adapting models to rapidly changing datasets, such as emerging drug side-effect databases.

Instruction or Alignment Tuning

Why It Matters?
Guides the model's behavior toward desired outcomes without extensive parameter changes, making it responsive to guidelines or task-specific instructions.

Use Cases:

- Regulatory compliance and consistency in clinical documentation.

Augmentation – Using data to overcome the model's inherent knowledge cutoff.

Overcoming Training Limitations

Retrieval Related

- Retriever Augmented Generation
- Cache Augmented Generation
- Knowledge Graphs

Context Related

- In-Context Augmented
- Tool Offloading

Prompt Engineering – Guiding the model towards desired outputs.

Context Related

Instruction

What Is It?
Clearly defining the task, desired constraints, and expected response format within the prompt.

Few Shot

What Is It?
Including example inputs and outputs directly in the prompt to demonstrate the desired behavior.

Chain of Thought

What Is It?
Prompting the model to explicitly show its reasoning process by breaking down problems step-by-step.

Roles and Personas

What Is It?
Assigning the model a specific role or persona in the prompt to influence its tone, style, or perspective.

Prompt Engineering – Guiding the model towards desired outputs.

Roles and Persona Example

With

Prompt:
As a medical professional, explain clearly and concisely how ibuprofen works to a patient who is curious about why it helps with pain.

Model Response:
Ibuprofen is a type of medicine known as a nonsteroidal anti-inflammatory drug (NSAID). It works by blocking substances in your body called prostaglandins, which cause inflammation, swelling, and pain. By reducing these substances, ibuprofen helps relieve pain and inflammation.

Without

Prompt:
Explain how ibuprofen works.

Model Response:
Ibuprofen reduces hormones that cause inflammation and pain.

Context Window Management – Maximizing model efficiency within token limit.

Possible Options

Splitting

What Is It?
Divide large inputs into manageable parts to solve complex problems step-by-step.

Why It Matters?
Allows processing of extensive content beyond the model's immediate token limit without losing critical information.

Use Cases:
Analyzing lengthy clinical trial reports section-by-section.

Summarization & Hierarchy

What Is It?
Condense or hierarchically summarize older information, freeing tokens for new inputs.

Why It Matters?
Keeps critical context available without exceeding token limits, maintaining clarity over prolonged interactions.

Use Cases:
Maintaining hierarchical summaries of extensive pharmacovigilance databases.

Needle in the Haystack

What Is It?
Techniques for precisely retrieving crucial details from large amounts of context.

Why It Matters?
Ensures accurate retrieval and use of key details within large data sets or documents, improving reliability.

Use Cases:
Identifying subtle adverse event signals within extensive clinical datasets.

Chunking of Data – Breaking down large content for effective processing.

Does Chunking Matter?

Fixed-Length

What Is It?
Splitting text into chunks of a predefined size, ensuring each chunk has the same length (number of tokens or words).

Why It Matters?
Ensures predictable input size for models, simplifying processing but potentially splitting critical information across chunks.

Use Cases:
Rapid, systematic processing of uniform clinical records or regulatory documents.

Topic-Based

What Is It?
Splitting content at logical, meaningful boundaries (such as paragraphs, sections, or chapters), preserving topic coherence.

Why It Matters?
Maintains contextual integrity, greatly enhancing model comprehension and accuracy within each chunk.

Use Cases:
Analyzing detailed clinical trial results, ensuring each chunk logically represents distinct phases or endpoints.

Sliding Window

What Is It?
Creating overlapping chunks where each new chunk partially repeats the previous chunk's content.

Why It Matters?
Maintains continuous context, significantly reducing information loss at chunk boundaries.

Use Cases:
Extracting detailed adverse event reports from lengthy clinical narratives without losing critical context.

Decoding Hyperparameters – Adjusting output creativity, diversity, and precision.

LLM Knobs

Temperature

What Is It?
Adjusts randomness in model outputs.

Why It Matters?
Controls trade-off between creativity and accuracy.

Use Cases:
Low: Precise medical dosing instructions.
High: Creative brainstorming of potential drug names.

Top-K

What Is It?
Limits sampling to the top K most likely next words.

Why It Matters?
Balances accuracy with diversity; reduces risk of errors or irrelevant outputs.

Use Cases:
Narrow (low K): Structured safety reporting.
Broad (high K): Exploring multiple diagnostic hypotheses.

Top-P

What Is It?
Samples from the smallest group of tokens whose cumulative probability exceeds P.

Why It Matters?
Provides dynamic flexibility between precision and creativity.

Use Cases:
Lower P: Regulatory-compliant document summaries.
Higher P: Generating varied patient education materials.

Model Optimization – Reducing model size and complexity without sacrificing performance.

Can We Shrink It?

Distillation

What Is It?
Training a smaller student model to mimic the performance of a larger teacher model.

Why It Matters?
Reduces model size significantly while maintaining accuracy, enabling efficient deployment on smaller devices.

Quantization

What Is It?
Reducing the precision of model weights and activations.

Why It Matters?
Greatly decreases memory usage and computational costs without substantial performance loss.

Pruning

What Is It?
Removing unnecessary weights or connections in the model to create sparsity and reduce complexity.

Why It Matters?
Optimizes models by making them smaller, faster, and less resource-intensive while preserving accuracy.

Context & Planning

AI generated output evaluation is highly dependent upon context.

What is the context of use?

Business development, QC, safety?

Who is the audience?

Investigators, patients, regulators, …

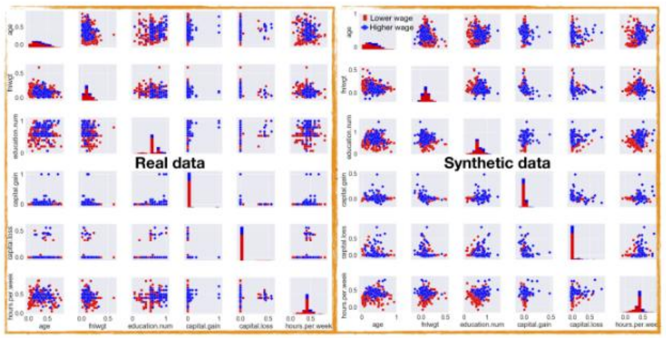
What are your key performance indicator cutoffs?

98%, 80%, …

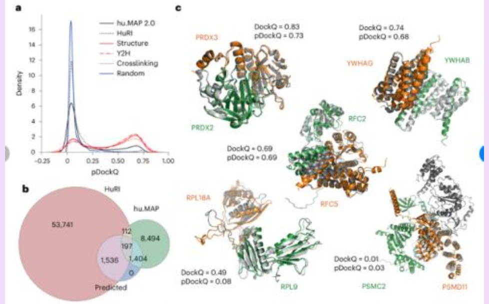
What "source of truth" will you use?

Gold standard, publications, internal source…

Figure 7: Pairwise comparison of the numerical features of the real and synthetic (WCGAN) datasets. The synthetic data shown here are obtained by training WCGAN architecture for 5,000 steps



Synthetic RW Data using GAN[2]



Application of AlphaFold2 complex predictions to a large dataset of human PPIs a, Distribution of model confidence score (pDockQ) for predicted structures from two large human protein interaction datasets (hu.MAP and HuRI), compared with confidence metrics from 2,000 random pairs of proteins. The hu.MAP dataset was further subsetted to those that have support from Y2H ('Y2H') or crosslink data ('Crosslinking'), or correspond to pairs with available experimental or homology modeling information ('Structure'). b, Number of protein interactions with models built from both datasets and those that we consider being of high confidence ('Predicted'), corresponding to those with pDockQ > 0.5. c, Examples of predicted models (orange and green) overlapped with the corresponding experimental models (gray) and the observed (DockQ) or predicted (pDockQ) quality of the models. Source data

Alpha Fold 2 (Deep Mind)[1]

1: Application of AlphaFold2 complex predictions to a large dataset of... | Download Scientific Diagram
2: Generative adversarial networks (GANs) for synthetic dataset generation with binary classes | Data Science Campus

A robust evaluation framework is important for the regulatory horizon and organizational governance.

Regulatory

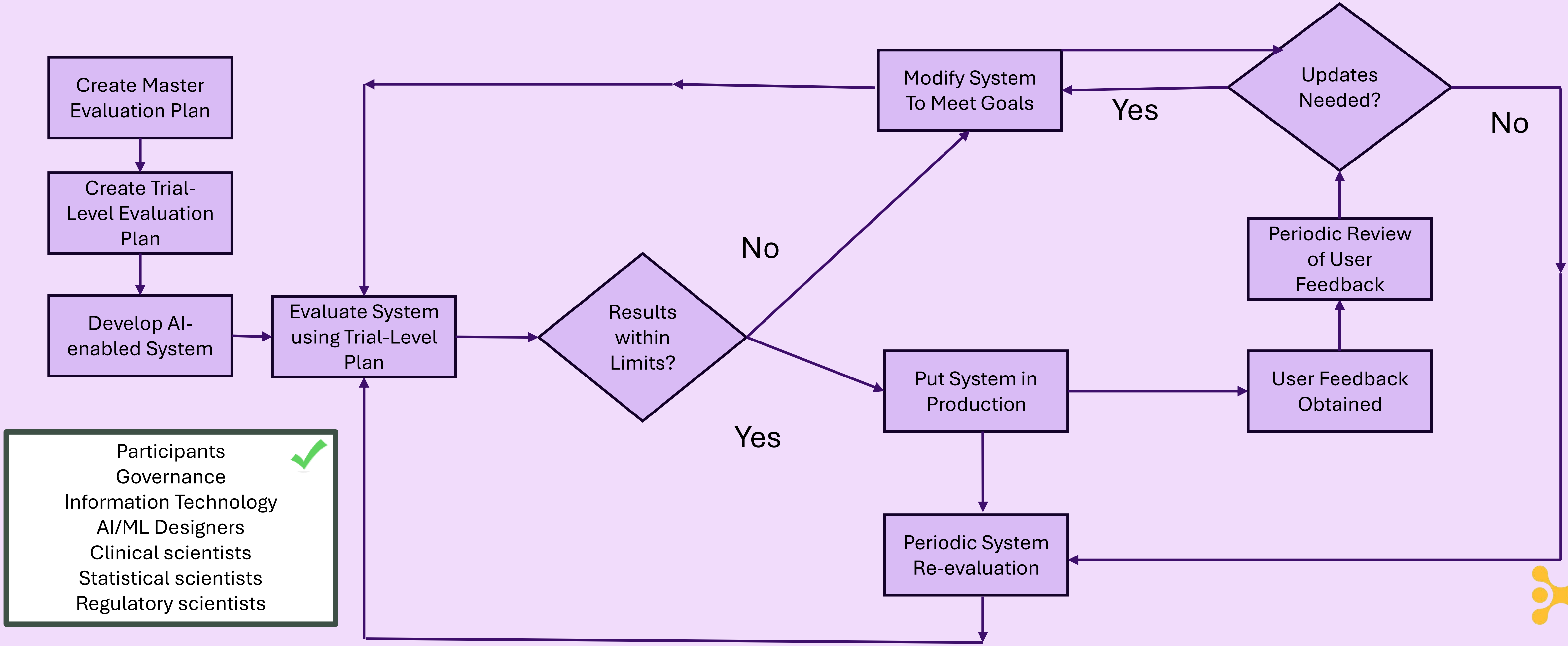
- Risk-based approach – more risk more scrutiny
- Transparency and explainability
  - Regarding AI algorithms developed, validated, and updated
- Lifecyle evaluation
  - Development and deployment journey
  - Maintenance
- Human oversight
  - Experts remain involved in key-decision making

Sponsor Governance

- Risk management and assessment
- Standard Operating Procedures (SOPs)
- Documentation and Traceability
- Cross-functional oversight
- Continuous improvement

U.S. Food and Drug Administration. (2025, January). *Considerations for the use of artificial intelligence to support regulatory decision-making for drug and biological products: Guidance for industry and other interested parties (Draft guidance)*. https://www.fda.gov/media/184830/download

AI output evaluation is a process which involves planning, execution, iteration, and documentation.



Create Master Evaluation Plan
Create Trial-Level Evaluation Plan
Develop AI-enabled System
Evaluate System using Trial-Level Plan
Results within Limits?
No
Yes
Modify System To Meet Goals
Put System in Production
Periodic System Re-evaluation
User Feedback Obtained
Periodic Review of User Feedback
Updates Needed?
Yes
No

<u>Participants</u> ✓
Governance
Information Technology
AI/ML Designers
Clinical scientists
Statistical scientists
Regulatory scientists

Exercise #1 – Evaluation Planning

Common evaluation dimensions for textual output.

- Alignment with peer-reviewed literature, regulatory guidelines, and known drug mechanisms

Accuracy & factual correctness

- Is it appropriate for its target audience (regulatory, clinical, patient) language, complexity, and terminology

Contextual Relevance

- Is the information presented in a coherent, organized manner that supports comprehension and flow?

Content Flow & Internal Consistency

- Are there potential biases (gender, racial, socioeconomic) or promotion?

Bias & ethics

- Does it adhere to appropriate industry regulations (FDA, EMA, ICH-GCP, HIPAA)

Compliance & Regulatory

- Ensure the text is clear, concise, and understandable for intended audience

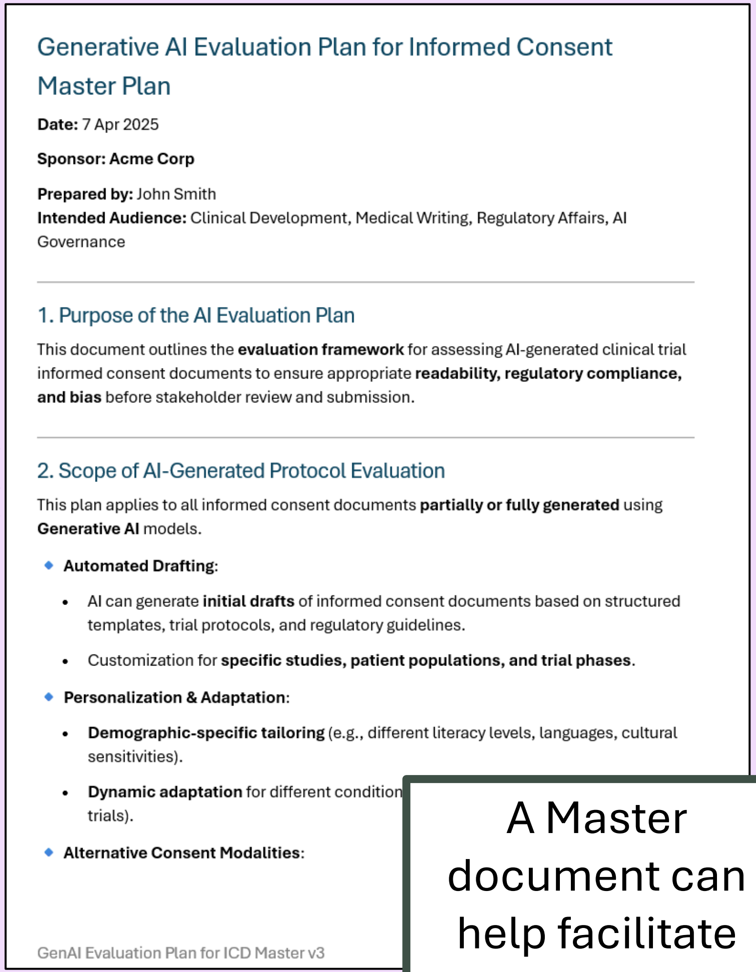
Readability & Usability

- AI benchmarking against human-written outputs to assess accuracy, completeness, and usability

Comparative Evaluation

Master and Trial-level evaluation plans pre-specify criteria and provides documentation.

Master



Generative AI Evaluation Plan for Informed Consent Master Plan

Date: 7 Apr 2025

Sponsor: Acme Corp

Prepared by: John Smith
Intended Audience: Clinical Development, Medical Writing, Regulatory Affairs, AI Governance

---

1. Purpose of the AI Evaluation Plan

This document outlines the **evaluation framework** for assessing AI-generated clinical trial informed consent documents to ensure appropriate **readability, regulatory compliance, and bias** before stakeholder review and submission.

---

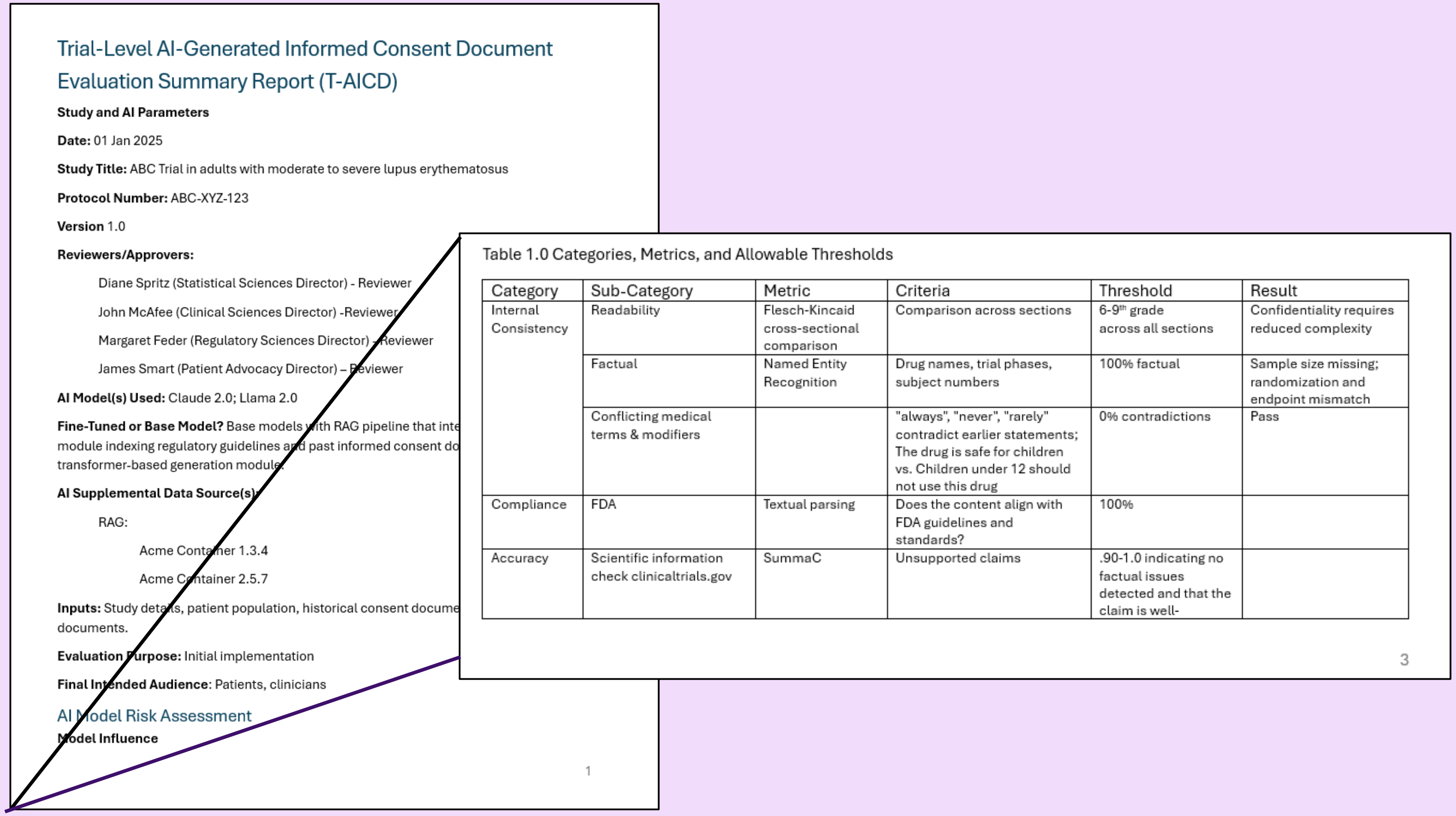
2. Scope of AI-Generated Protocol Evaluation

This plan applies to all informed consent documents **partially or fully generated** using **Generative AI** models.

- **Automated Drafting**:
  - AI can generate **initial drafts** of informed consent documents based on structured templates, trial protocols, and regulatory guidelines.
  - Customization for **specific studies, patient populations, and trial phases**.
- **Personalization & Adaptation**:
  - **Demographic-specific tailoring** (e.g., different literacy levels, languages, cultural sensitivities).
  - **Dynamic adaptation** for different condition
    trials).
- **Alternative Consent Modalities**:

A Master document can help facilitate efficiency.

Trial-level



Trial-Level AI-Generated Informed Consent Document Evaluation Summary Report (T-AICD)

Study and AI Parameters

Date: 01 Jan 2025

Study Title: ABC Trial in adults with moderate to severe lupus erythematosus

Protocol Number: ABC-XYZ-123

Version 1.0

Reviewers/Approvers:

Diane Spritz (Statistical Sciences Director) - Reviewer

John McAfee (Clinical Sciences Director) -Reviewer

Margaret Feder (Regulatory Sciences Director) - Reviewer

James Smart (Patient Advocacy Director) – Reviewer

AI Model(s) Used: Claude 2.0; Llama 2.0

Fine-Tuned or Base Model? Base models with RAG pipeline that inte
module indexing regulatory guidelines and past informed consent do
transformer-based generation module.

AI Supplemental Data Source(s):

RAG:

Acme Container 1.3.4

Acme Container 2.5.7

Inputs: Study details, patient population, historical consent docume
documents.

Evaluation Purpose: Initial implementation

Final Intended Audience: Patients, clinicians

AI Model Risk Assessment

Model Influence

Table 1.0 Categories, Metrics, and Allowable Thresholds

Category	Sub-Category	Metric	Criteria	Threshold	Result
Internal Consistency	Readability	Flesch-Kincaid cross-sectional comparison	Comparison across sections	6-9th grade across all sections	Confidentiality requires reduced complexity
	Factual	Named Entity Recognition	Drug names, trial phases, subject numbers	100% factual	Sample size missing; randomization and endpoint mismatch
	Conflicting medical terms & modifiers		"always", "never", "rarely" contradict earlier statements; The drug is safe for children vs. Children under 12 should not use this drug	0% contradictions	Pass
Compliance	FDA	Textual parsing	Does the content align with FDA guidelines and standards?	100%	
Accuracy	Scientific information check clinicaltrials.gov	SummaC	Unsupported claims	.90-1.0 indicating no factual issues detected and that the claim is well-	

The Master document and trial-level documents have different focus areas.

Master

- Purpose
  - ICD, protocol
  - Accuracy, bias
  - For use by whom
- Scope
  - Drafting, personalization, alternative modalities
  - Human oversight
- Metric options and processes
- Workflow description
  - Versioning, evaluation, human review, iteration)
- Assurance
  - Scientific rigor, regulatory adherence, etc.

Trial-level

- Which study
- Who are the evaluators, approvers
- Risk assessment
  - Decision consequence
- Limitations and potential biases
- Specific evaluation metrics and thresholds
- Results documentation
- Revision decisions/actions

Quantitative Evaluation Metrics

How are LLMs (versus their output) trained and evaluated

Cross Entropy

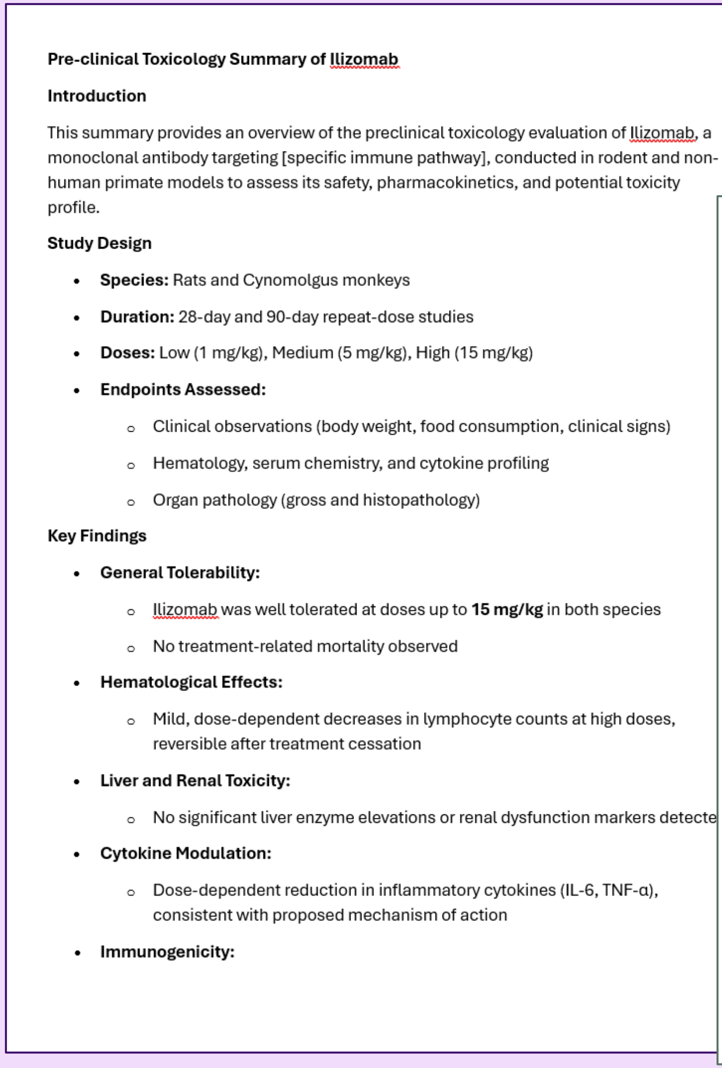
- Measures how well predicted distribution aligns with the true (target) token.
- It's calculated as the **negative log-likelihood** of the correct token:
  - If the model assigns high probability to the correct token → low loss.
  - If the model assigns low probability to the correct token → high loss.
- This loss guides how the model updates its weights during training.
- *“How surprised is the model to see the actual next word, based on its own prediction?” The less surprised, the better.*

Perplexity

- After training, perplexity is often used to indicate how well the model performs.
- **Perplexity is the exponentiated average cross entropy**:
  - Lower perplexity → better performance.
  - Perplexity of 1 = perfect certainty, higher values mean more uncertainty.
- Interpreted as the *"average number of choices"* the model considers plausible at each step.
- *“How many plausible options is the model juggling at each prediction step?” Lower is better.*

Classification Based Metrics

Trial-suite documents are available for the upcoming set of examples.



Pre-clinical Toxicology Summary of Ilizomab

Introduction

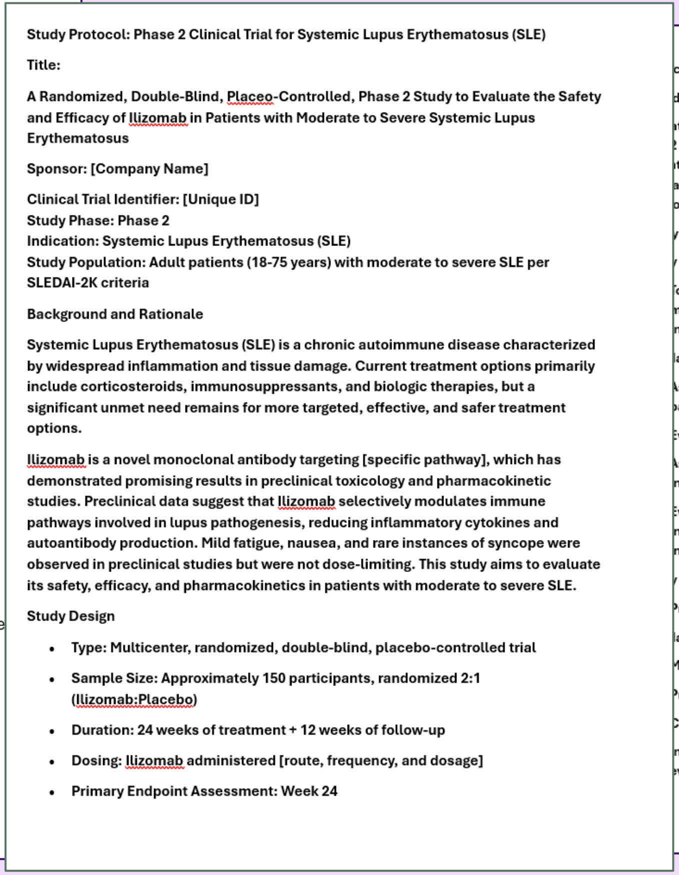
This summary provides an overview of the preclinical toxicology evaluation of Ilizomab, a monoclonal antibody targeting [specific immune pathway], conducted in rodent and non-human primate models to assess its safety, pharmacokinetics, and potential toxicity profile.

Study Design

- **Species:** Rats and Cynomolgus monkeys
- **Duration:** 28-day and 90-day repeat-dose studies
- **Doses:** Low (1 mg/kg), Medium (5 mg/kg), High (15 mg/kg)
- **Endpoints Assessed:**
  - Clinical observations (body weight, food consumption, clinical signs)
  - Hematology, serum chemistry, and cytokine profiling
  - Organ pathology (gross and histopathology)

Key Findings

- **General Tolerability:**
  - Ilizomab was well tolerated at doses up to **15 mg/kg** in both species
  - No treatment-related mortality observed
- **Hematological Effects:**
  - Mild, dose-dependent decreases in lymphocyte counts at high doses, reversible after treatment cessation
- **Liver and Renal Toxicity:**
  - No significant liver enzyme elevations or renal dysfunction markers detecte
- **Cytokine Modulation:**
  - Dose-dependent reduction in inflammatory cytokines (IL-6, TNF-α), consistent with proposed mechanism of action
- **Immunogenicity:**



Study Protocol: Phase 2 Clinical Trial for Systemic Lupus Erythematosus (SLE)

Title:

A Randomized, Double-Blind, Placeo-Controlled, Phase 2 Study to Evaluate the Safety and Efficacy of Ilizomab in Patients with Moderate to Severe Systemic Lupus Erythematosus

Sponsor: [Company Name]

Clinical Trial Identifier: [Unique ID]
Study Phase: Phase 2
Indication: Systemic Lupus Erythematosus (SLE)
Study Population: Adult patients (18-75 years) with moderate to severe SLE per SLEDAI-2K criteria

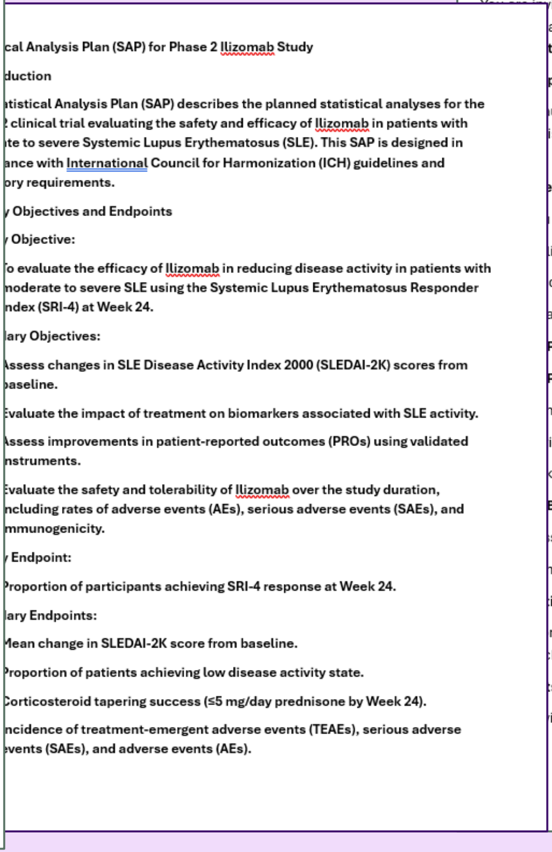
Background and Rationale

Systemic Lupus Erythematosus (SLE) is a chronic autoimmune disease characterized by widespread inflammation and tissue damage. Current treatment options primarily include corticosteroids, immunosuppressants, and biologic therapies, but a significant unmet need remains for more targeted, effective, and safer treatment options.

Ilizomab is a novel monoclonal antibody targeting [specific pathway], which has demonstrated promising results in preclinical toxicology and pharmacokinetic studies. Preclinical data suggest that Ilizomab selectively modulates immune pathways involved in lupus pathogenesis, reducing inflammatory cytokines and autoantibody production. Mild fatigue, nausea, and rare instances of syncope were observed in preclinical studies but were not dose-limiting. This study aims to evaluate its safety, efficacy, and pharmacokinetics in patients with moderate to severe SLE.

Study Design

- **Type: Multicenter, randomized, double-blind, placebo-controlled trial**
- **Sample Size: Approximately 150 participants, randomized 2:1 (Ilizomab:Placebo)**
- **Duration: 24 weeks of treatment + 12 weeks of follow-up**
- **Dosing: Ilizomab administered [route, frequency, and dosage]**
- **Primary Endpoint Assessment: Week 24**



cal Analysis Plan (SAP) for Phase 2 Ilizomab Study

duction

tistical Analysis Plan (SAP) describes the planned statistical analyses for the clinical trial evaluating the safety and efficacy of Ilizomab in patients with te to severe Systemic Lupus Erythematosus (SLE). This SAP is designed in ance with International Council for Harmonization (ICH) guidelines and ory requirements.

Objectives and Endpoints

Objective:

o evaluate the efficacy of Ilizomab in reducing disease activity in patients with moderate to severe SLE using the Systemic Lupus Erythematosus Responder ndex (SRI-4) at Week 24.

lary Objectives:

ssess changes in SLE Disease Activity Index 2000 (SLEDAI-2K) scores from aseline.

valuate the impact of treatment on biomarkers associated with SLE activity.

ssess improvements in patient-reported outcomes (PROs) using validated nstruments.

valuate the safety and tolerability of Ilizomab over the study duration, ncluding rates of adverse events (AEs), serious adverse events (SAEs), and mmunogenicity.

Endpoint:

roportion of participants achieving SRI-4 response at Week 24.

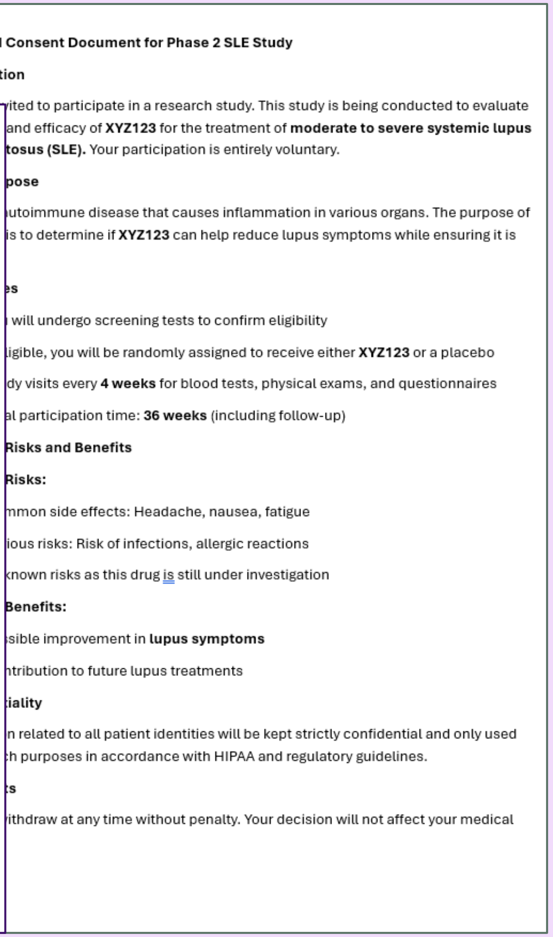
lary Endpoints:

ean change in SLEDAI-2K score from baseline.

roportion of patients achieving low disease activity state.

orticosteroid tapering success (≤5 mg/day prednisone by Week 24).

ncidence of treatment-emergent adverse events (TEAEs), serious adverse vents (SAEs), and adverse events (AEs).



Informed Consent Document for Phase 2 SLE Study

Introduction

ited to participate in a research study. This study is being conducted to evaluate and efficacy of **XYZ123** for the treatment of **moderate to severe systemic lupus tosus (SLE).** Your participation is entirely voluntary.

pose

utoimmune disease that causes inflammation in various organs. The purpose of is to determine if **XYZ123** can help reduce lupus symptoms while ensuring it is

s

will undergo screening tests to confirm eligibility

ligible, you will be randomly assigned to receive either **XYZ123** or a placebo

dy visits every **4 weeks** for blood tests, physical exams, and questionnaires

l participation time: **36 weeks** (including follow-up)

Risks and Benefits

Risks:

mmon side effects: Headache, nausea, fatigue

ious risks: Risk of infections, allergic reactions

known risks as this drug is still under investigation

Benefits:

sible improvement in **lupus symptoms**

ntribution to future lupus treatments

iality

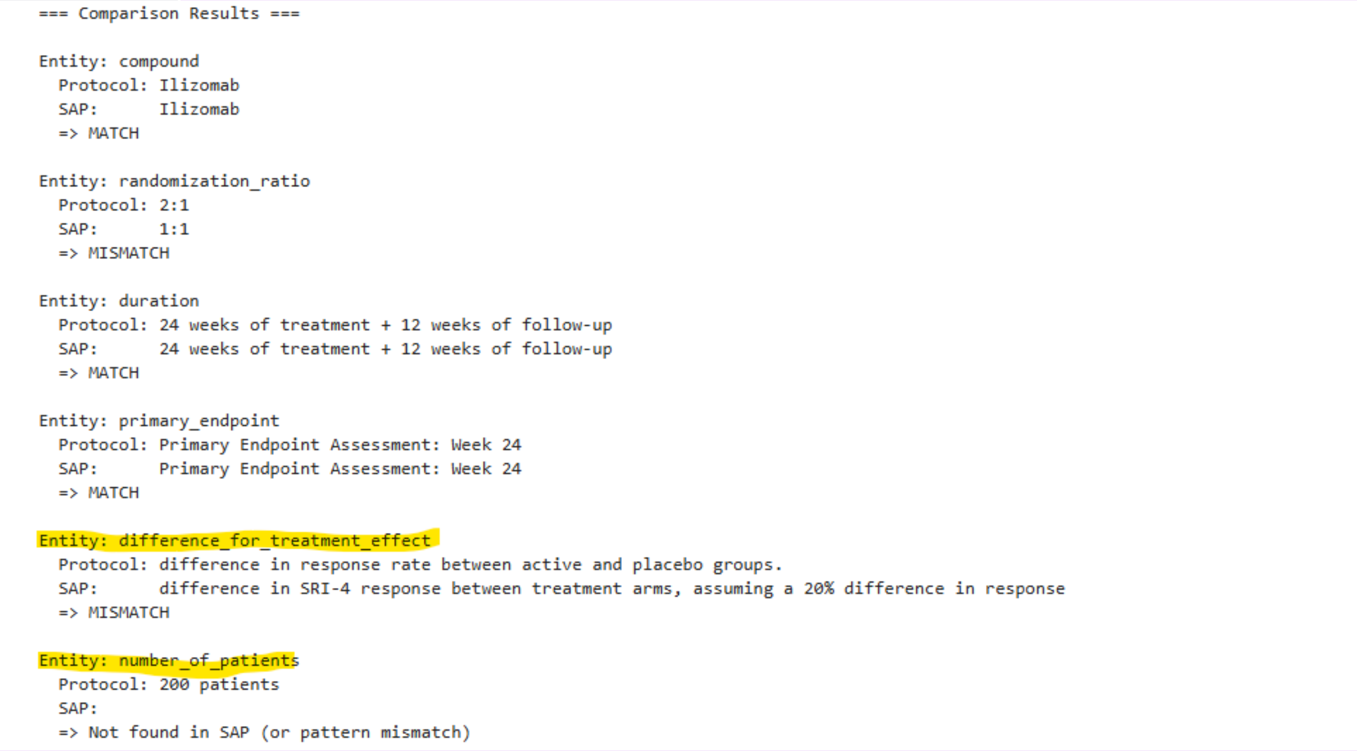
n related to all patient identities will be kept strictly confidential and only used h purposes in accordance with HIPAA and regulatory guidelines.

s

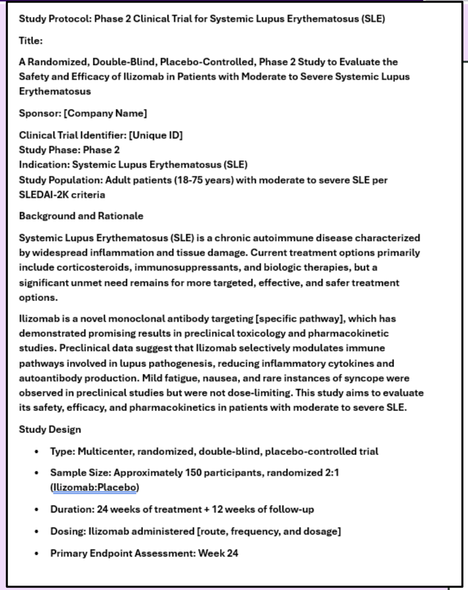
ithdraw at any time without penalty. Your decision will not affect your medical

Named entity recognition is a type of classification and can be used to assess accuracy

Variable	Protocol	SAP	Status
Investigational Product	Ilizomab	Ilizomab	Match
Randomization	2:1	1:1	Mismatch
Primary endpoint assessment	24 Weeks	52 Weeks	Mismatch
Study duration	24 Weeks + 12	24 Weeks + 12	Match
Sample size	150		Not found in SAP



```
=== Comparison Results ===

Entity: compound
  Protocol: Ilizomab
  SAP:      Ilizomab
  => MATCH

Entity: randomization_ratio
  Protocol: 2:1
  SAP:      1:1
  => MISMATCH

Entity: duration
  Protocol: 24 weeks of treatment + 12 weeks of follow-up
  SAP:      24 weeks of treatment + 12 weeks of follow-up
  => MATCH

Entity: primary_endpoint
  Protocol: Primary Endpoint Assessment: Week 24
  SAP:      Primary Endpoint Assessment: Week 24
  => MATCH

Entity: difference_for_treatment_effect
  Protocol: difference in response rate between active and placebo groups.
  SAP:      difference in SRI-4 response between treatment arms, assuming a 20% difference in response
  => MISMATCH

Entity: number_of_patients
  Protocol: 200 patients
  SAP:
  => Not found in SAP (or pattern mismatch)
```



Study Protocol: Phase 2 Clinical Trial for Systemic Lupus Erythematosus (SLE)

Title:

A Randomized, Double-Blind, Placebo-Controlled, Phase 2 Study to Evaluate the Safety and Efficacy of Ilizomab in Patients with Moderate to Severe Systemic Lupus Erythematosus

Sponsor: [Company Name]

Clinical Trial Identifier: [Unique ID]
Study Phase: Phase 2
Indication: Systemic Lupus Erythematosus (SLE)
Study Population: Adult patients (18-75 years) with moderate to severe SLE per SLEDAI-2K criteria

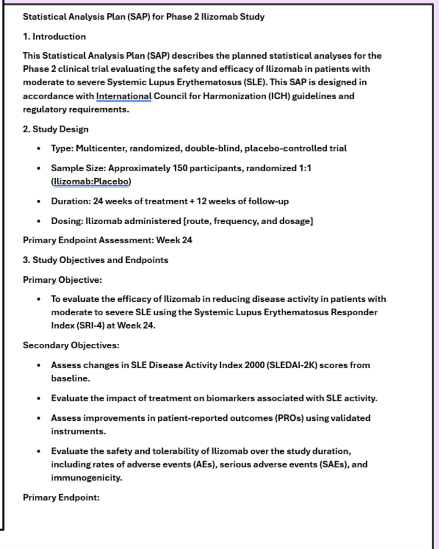
Background and Rationale

Systemic Lupus Erythematosus (SLE) is a chronic autoimmune disease characterized by widespread inflammation and tissue damage. Current treatment options primarily include corticosteroids, immunosuppressants, and biologic therapies, but a significant unmet need remains for more targeted, effective, and safer treatment options.

Ilizomab is a novel monoclonal antibody targeting [specific pathway], which has demonstrated promising results in preclinical toxicology and pharmacokinetic studies. Preclinical data suggest that Ilizomab selectively modulates immune pathways involved in lupus pathogenesis, reducing inflammatory cytokines and autoantibody production. Mild fatigue, nausea, and rare instances of syncope were observed in preclinical studies but were not dose-limiting. This study aims to evaluate its safety, efficacy, and pharmacokinetics in patients with moderate to severe SLE.

Study Design

- Type: Multicenter, randomized, double-blind, placebo-controlled trial
- Sample Size: Approximately 150 participants, randomized 2:1 (Ilizomab:Placebo)
- Duration: 24 weeks of treatment + 12 weeks of follow-up
- Dosing: Ilizomab administered [route, frequency, and dosage]
- Primary Endpoint Assessment: Week 24



Statistical Analysis Plan (SAP) for Phase 2 Ilizomab Study

1. Introduction

This Statistical Analysis Plan (SAP) describes the planned statistical analyses for the Phase 2 clinical trial evaluating the safety and efficacy of Ilizomab in patients with moderate to severe Systemic Lupus Erythematosus (SLE). This SAP is designed in accordance with International Council for Harmonization (ICH) guidelines and regulatory requirements.

2. Study Design

- Type: Multicenter, randomized, double-blind, placebo-controlled trial
- Sample Size: Approximately 150 participants, randomized 1:1 (Ilizomab:Placebo)
- Duration: 24 weeks of treatment + 12 weeks of follow-up
- Dosing: Ilizomab administered [route, frequency, and dosage]

Primary Endpoint Assessment: Week 24

3. Study Objectives and Endpoints

Primary Objective:

- To evaluate the efficacy of Ilizomab in reducing disease activity in patients with moderate to severe SLE using the Systemic Lupus Erythematosus Responder Index (SRI-4) at Week 24.

Secondary Objectives:

- Assess changes in SLE Disease Activity Index 2000 (SLEDAI-2K) scores from baseline.
- Evaluate the impact of treatment on biomarkers associated with SLE activity.
- Assess improvements in patient-reported outcomes (PROs) using validated instruments.
- Evaluate the safety and tolerability of Ilizomab over the study duration, including rates of adverse events (AEs), serious adverse events (SAEs), and immunogenicity.

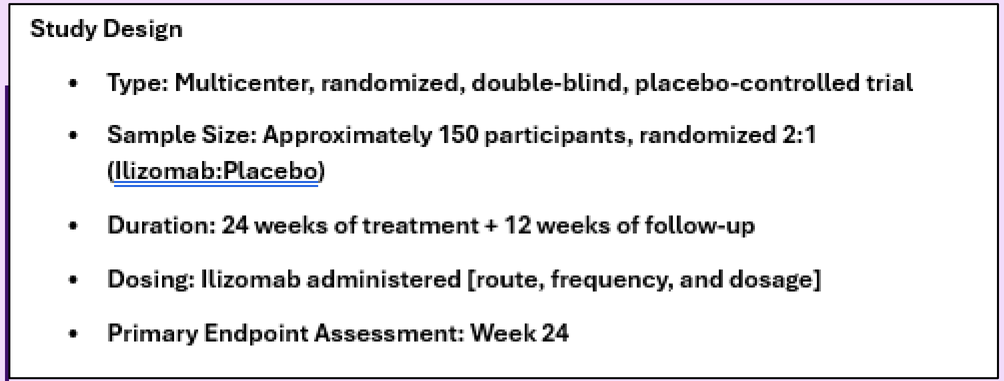
Primary Endpoint:

Iteration, followed by consistency, help in future evaluations.

Accuracy solution

- Consistent language
  - patients, participants
- Consistent formatting
  - bullets
  - sections

Protocol



Study Design

- Type: Multicenter, randomized, double-blind, placebo-controlled trial
- Sample Size: Approximately 150 participants, randomized 2:1 (Ilizomab:Placebo)
- Duration: 24 weeks of treatment + 12 weeks of follow-up
- Dosing: Ilizomab administered [route, frequency, and dosage]
- Primary Endpoint Assessment: Week 24

Coding solution

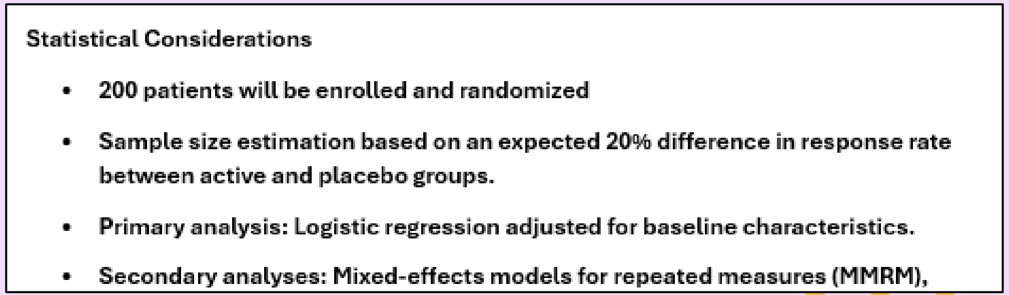
```
# Number of patients (e.g., N=123 or "123
patients")    "number_of_patients":
re.compile(r"(?i)(\bN\s?=\s?\d+\b|\b\d+\s+p
atients?\b)")
```

```
elif entity_name ==
"number_of_patients":       # Return a
list of all sample size instances found
in the text       return [m.strip() for m in
pattern.findall(text)]
```

```
# Sample size (e.g., N=123 or "123
patients/participants")    "number_of_patients":
re.compile(r"(?i)(\bN\s?=\s?\d+\b|\b\d+\s+(?:p
atients?|participants?)\b)")
```

```
# Sample size (e.g., N=123 or "123
patients/participants")
"number_of_patients":
re.compile(r"(?i)(\bN\s?=\s?\d+\b|\b\
d+\s+(?:patients?|participants?)\b)")
```

SAP



Statistical Considerations

- 200 patients will be enrolled and randomized
- Sample size estimation based on an expected 20% difference in response rate between active and placebo groups.
- Primary analysis: Logistic regression adjusted for baseline characteristics.
- Secondary analyses: Mixed-effects models for repeated measures (MMRM),

Classification metrics are useful when converting instructed text to trial attributes.

Purpose	• Used to evaluate discrete or categorical model decisions.
Intuition and Context	• These metrics assume there's a **ground truth** (what *should* have been extracted or classified) • Relevant when AI output is translated into structured labels or decisions (e.g., identifying endpoints, interventions, or population criteria from unstructured text like protocols).

Example Use Cases of Classification Metrics

- Consider the example of extracting the List Of Analysis (LOA)
- Consider an annotated document where each sentence is classified as containing Analysis (Positive Class) or not an analysis (Negative Class)

Accuracy

- The proportion of total predictions that were correct.
- The total number of correct predictions for both the positive and negative class divided by the number of sections.

Precision

- Of all the times the model said “yes,” how often was it right?
- False Positives are costly
- The number of correctly identified analyses divided by the number of times the model identified a section as an analysis

Recall

- Of all the actual “yes” cases, how many did the model find?
- False Negatives are costly
- The number of correctly identified analyses divided by the total number of analyses

F1

- The harmonic mean of precision and recall, balances both.

Summary of Classification Metrics

Metric	Definition	Formula	Use Case/Intuition
Accuracy	Overall correctness of the model's predictions	(TP + TN) / (TP + TN + FP + FN)	Good for balanced classes; can be misleading when one class dominates
Precision	How many predicted positives are actually correct	TP / (TP + FP)	Important when false positives are costly (e.g., mislabeling non-endpoints as endpoints)
Recall	How many actual positives were correctly identified	TP / (TP + FN)	Important when false negatives are costly (e.g., missing a key inclusion criterion)
F1-Score	Harmonic mean of precision and recall	2 × (Precision × Recall) / (Precision + Recall)	Balances precision and recall; useful when both false positives and false negatives matter
Confusion Matrix	Tabular layout of prediction outcomes: TP, TN, FP, FN (extended to multi-class)	N/A (structure rather than scalar)	Used to visualize model performance; helps diagnose which classes are being confused

There are limitations to classification metrics.

✓

- **Structured tasks** where there is a clear, discrete set of correct answers
    - Information extraction (e.g., extract a drug name)
    - Document classification (e.g., is this a clinical trial or not?)

✗

- **Open-ended generation tasks**, like:
    - Summarization
    - Answer generation
    - Rewriting or paraphrasing

Reference Based Metrics

Reference based metrics are useful for evaluating generated text.

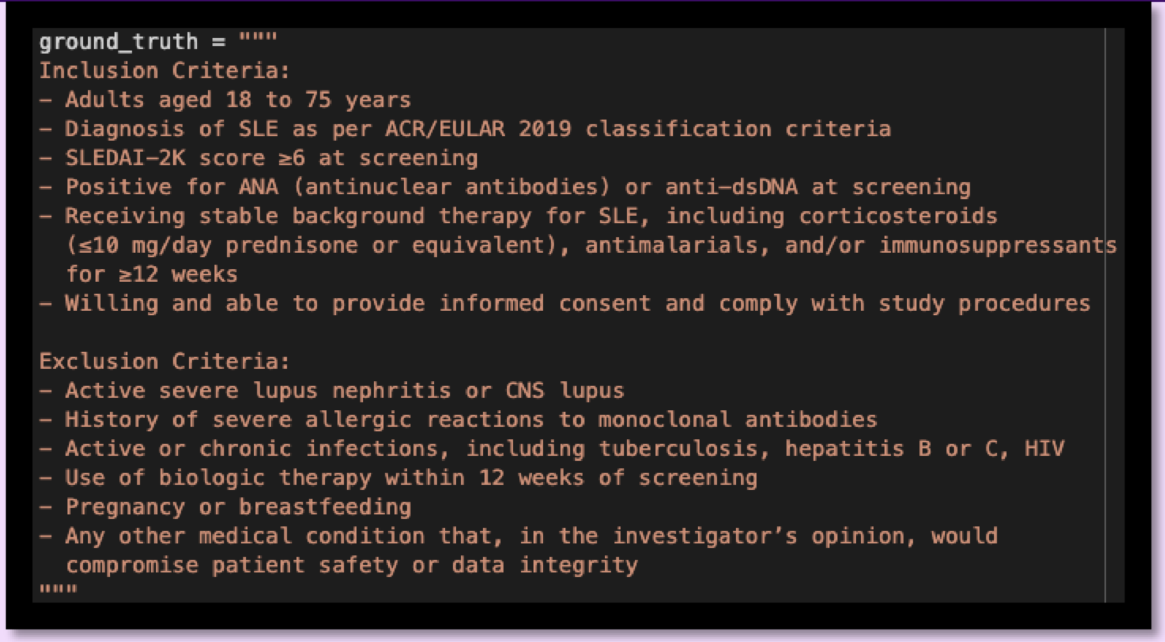
Purpose

Measure how *similar* the generated text is to a known, high-quality **reference** (e.g., human-written protocol).

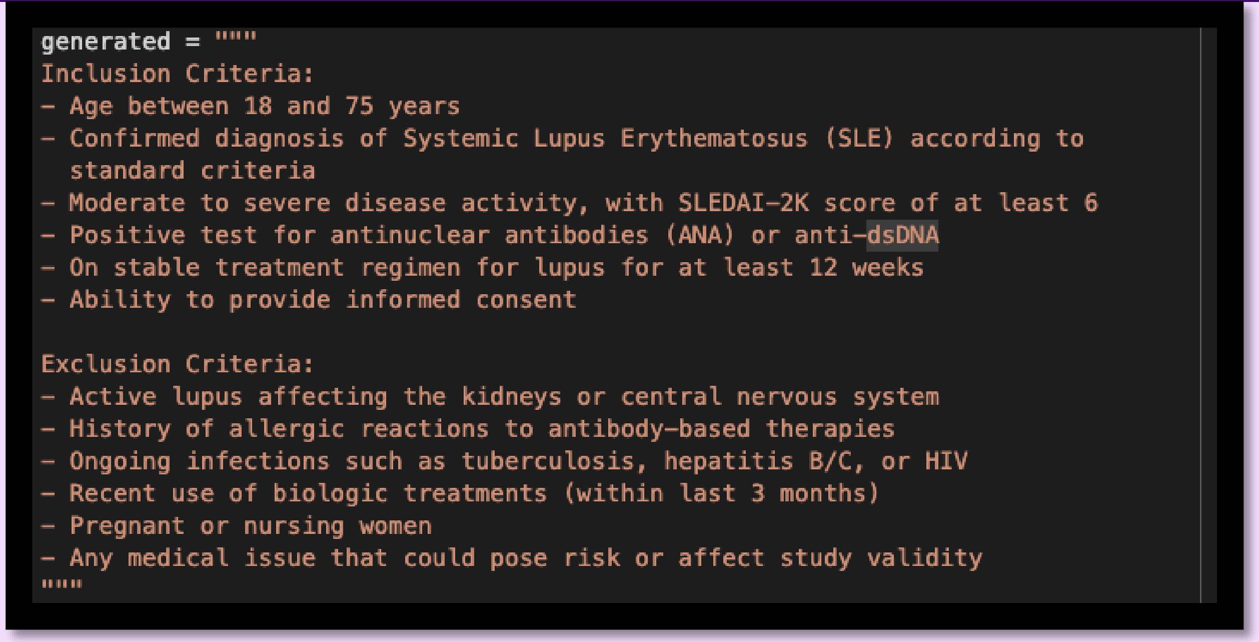
Intuition and Context

- These metrics assess **surface-level overlap** between generated output and reference text.
- Used when you expect **verbatim or near-verbatim phrasing**, such as standard SAP language, eligibility criteria, or objective definitions.
- Often used in **generation tasks** where classification metrics (like precision/recall) fall short because output is **free-form text**.

Reference based metrics can support consistency within a program.



```
ground_truth = """
Inclusion Criteria:
- Adults aged 18 to 75 years
- Diagnosis of SLE as per ACR/EULAR 2019 classification criteria
- SLEDAI-2K score ≥6 at screening
- Positive for ANA (antinuclear antibodies) or anti-dsDNA at screening
- Receiving stable background therapy for SLE, including corticosteroids
  (≤10 mg/day prednisone or equivalent), antimalarials, and/or immunosuppressants
  for ≥12 weeks
- Willing and able to provide informed consent and comply with study procedures

Exclusion Criteria:
- Active severe lupus nephritis or CNS lupus
- History of severe allergic reactions to monoclonal antibodies
- Active or chronic infections, including tuberculosis, hepatitis B or C, HIV
- Use of biologic therapy within 12 weeks of screening
- Pregnancy or breastfeeding
- Any other medical condition that, in the investigator’s opinion, would
  compromise patient safety or data integrity
"""
```



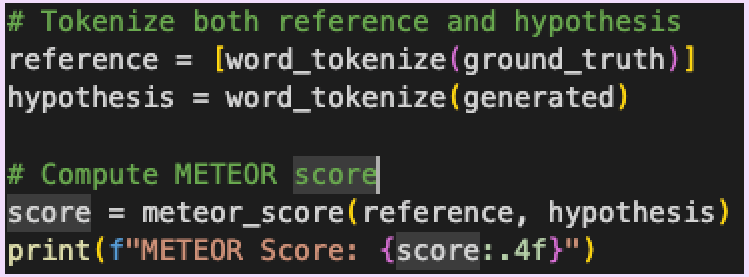
```
generated = """
Inclusion Criteria:
- Age between 18 and 75 years
- Confirmed diagnosis of Systemic Lupus Erythematosus (SLE) according to
  standard criteria
- Moderate to severe disease activity, with SLEDAI-2K score of at least 6
- Positive test for antinuclear antibodies (ANA) or anti-dsDNA
- On stable treatment regimen for lupus for at least 12 weeks
- Ability to provide informed consent

Exclusion Criteria:
- Active lupus affecting the kidneys or central nervous system
- History of allergic reactions to antibody-based therapies
- Ongoing infections such as tuberculosis, hepatitis B/C, or HIV
- Recent use of biologic treatments (within last 3 months)
- Pregnant or nursing women
- Any medical issue that could pose risk or affect study validity
"""
```

- Ground truth is extracted from a protocol approved in a similar program.
- AI-generated text is created by passing in relevant information such as the Indication, Phase, and Population into a generative model, like ChatGPT
- The generated text is then compared to your ground truth and evaluated for similarity.

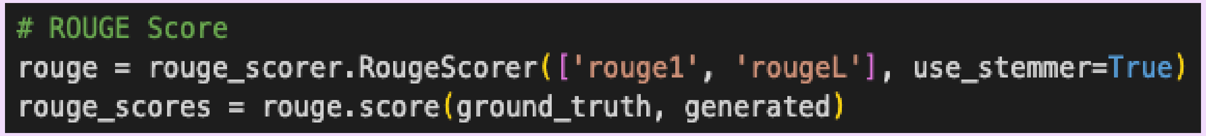
Example: Evaluating Inclusion/Exclusion Criteria Continued

- Rouge-1 (>.5 Good, .4 - .5 Moderate)
- Rouge-2 (>.4 Good, .2 - .4 Moderate)
- These metrics are not always reliable, as seen by Levenshtein Similarity Ratio and JSD
- They can be sensitive to small changes in language

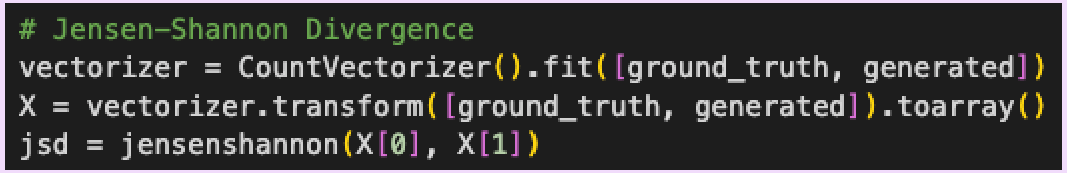
Metric	Score	Range
Rouge-1 Precision	.645	0 - 1
Rouge-1 Recall	.572	0 - 1
Rouge-1 F1	.607	0 - 1
METEOR	.471	0 - 1
JSD	.518	0 – 1 0 = Identical
Levenshtein Similarity Ratio	.261	0 – 1 1 = Identical



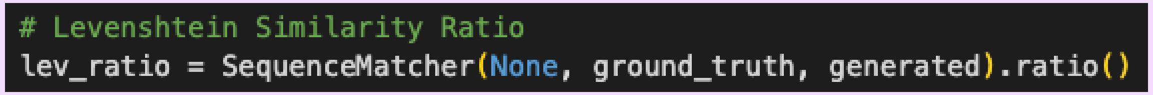
```
# Tokenize both reference and hypothesis
reference = [word_tokenize(ground_truth)]
hypothesis = word_tokenize(generated)

# Compute METEOR score
score = meteor_score(reference, hypothesis)
print(f"METEOR Score: {score:.4f}")
```



```
# ROUGE Score
rouge = rouge_scorer.RougeScorer(['rouge1', 'rougeL'], use_stemmer=True)
rouge_scores = rouge.score(ground_truth, generated)
```



```
# Jensen-Shannon Divergence
vectorizer = CountVectorizer().fit([ground_truth, generated])
X = vectorizer.transform([ground_truth, generated]).toarray()
jsd = jensenshannon(X[0], X[1])
```



```
# Levenshtein Similarity Ratio
lev_ratio = SequenceMatcher(None, ground_truth, generated).ratio()
```

Summary of Reference Based Metrics

Metric	Definition	Use Case / Intuition
ROUGE (Recall-Oriented Understudy for Gisting Evaluation)	Measures n-gram overlap between generated and reference texts	Good for summarization or checking content coverage; more recall-focused
METEOR	Combines exact, stem, synonym, and paraphrase matches with position penalties	More nuanced than ROUGE; better for sentence-level similarity in natural language
Jensen-Shannon Divergence	Measures divergence between word distributions in two texts (probabilistic distance)	Captures topical or stylistic drift; useful for comparing statistical language patterns
Levenshtein Similarity Ratio	Calculates edit distance between strings, normalized to similarity score	Useful for measuring literal string closeness; penalizes word insertions/deletions/swaps

So, Are These Metrics Enough To Rely On?

NO!

Why?

These metrics don't consider the meaning of the words, just look for overlap

Can be misleading if generated text uses different words

Might perform poorly for paraphrased or restructured wording

Semantic Similarity

Semantic similarity better captures if the meaning is similar between two sources.

Purpose

- Capture **meaning**, not just surface form — ideal for evaluating **paraphrased** or **restructured** model outputs.

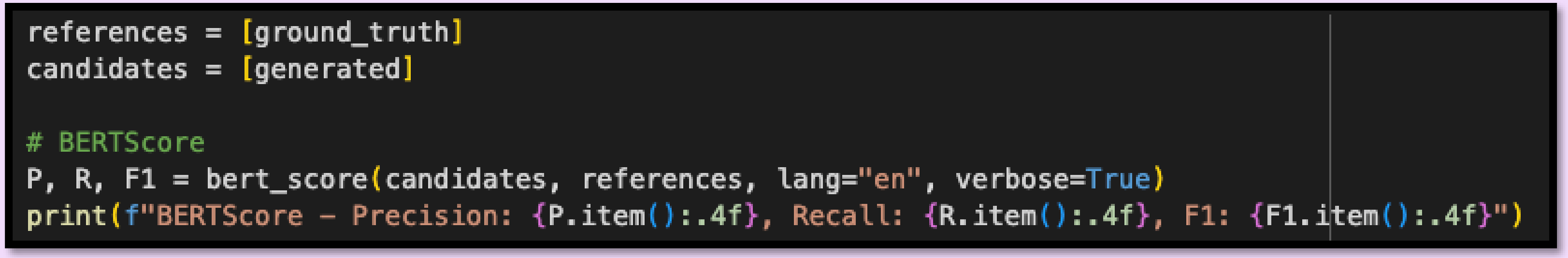
Intuition and Context

- Powered by **pretrained or fine-tuned language models** (like BERT, T5) to compare **deep representations** of text.
- Especially useful when generated content is valid but deviates from the reference in wording or structure.
- Traditional metrics like ROUGE may penalize valid outputs just because they're worded differently. Semantic metrics help fix that.

Example: Revisiting Inclusion/Exclusion Evaluation

- In the example, the generated inclusion, exclusion criteria was very similar to the ground truth (protocol)
- BERTScore scores this highly since the meaning is preserved in the generated text even though the language is varied.

Method	Score
BERTScore - Precision	.9135
BERTScore - Recall	.8912
BERTScore - F1	.9022

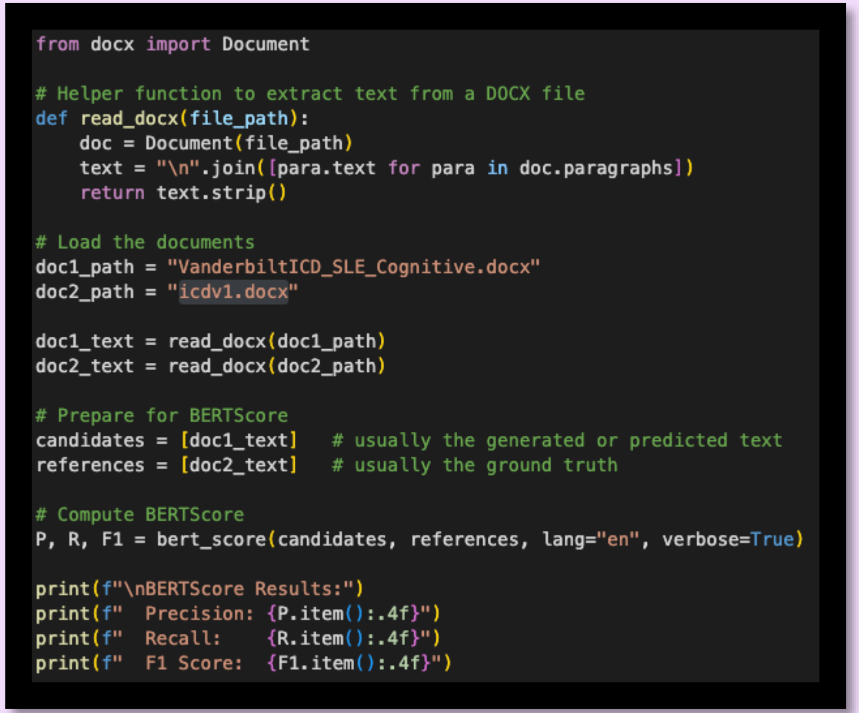


```
references = [ground_truth]
candidates = [generated]

# BERTScore
P, R, F1 = bert_score(candidates, references, lang="en", verbose=True)
print(f"BERTScore - Precision: {P.item():.4f}, Recall: {R.item():.4f}, F1: {F1.item():.4f}")
```

Or, documents can be compared to a gold standard.

- Requires reference or gold-standard
    - Text(VanderbiltICD_SLE_Cognifive.docx)
- Requires candidate text
    - (icdv1.docx)
- Benefits from pre-alignment or reordering of candidate segments

Method	Score
BERTScore - Precision	.8427
BERTScore - Recall	.8310
BERTScore - F1	.8368



```
from docx import Document

# Helper function to extract text from a DOCX file
def read_docx(file_path):
    doc = Document(file_path)
    text = "\n".join([para.text for para in doc.paragraphs])
    return text.strip()

# Load the documents
doc1_path = "VanderbiltICD_SLE_Cognitive.docx"
doc2_path = "icdv1.docx"

doc1_text = read_docx(doc1_path)
doc2_text = read_docx(doc2_path)

# Prepare for BERTScore
candidates = [doc1_text]   # usually the generated or predicted text
references = [doc2_text]   # usually the ground truth

# Compute BERTScore
P, R, F1 = bert_score(candidates, references, lang="en", verbose=True)

print(f"\nBERTScore Results:")
print(f"  Precision: {P.item():.4f}")
print(f"  Recall:    {R.item():.4f}")
print(f"  F1 Score:  {F1.item():.4f}")
```

Summary of Similarity Metrics

Metric	Type	Embedding Model	Granularity	Notes
BERTScore	Semantic, reference-based	BERT/RoBERTa	Token	Captures contextual meaning
MoverScore	Semantic + structural	BERT + IDF	Token	Adds word importance and structure
SMS	Semantic, discourse-level	Sentence-BERT	Sentence	Captures coherence & structure
BLEURT	Learned semantic model	Fine-tuned BERT	Sentence	Tuned to match human judgment

Do We Now Have All The Metrics We Need?

NO!

Why?

Semantic similarity is not always enough

Doesn’t always align with human perspective

How can we evaluate the performance of generated text without ground truth?

Reference Free Metrics

Reference free metrics support the evaluation of LLM outputs without the need for ground truth.

Purpose

- Evaluate quality of generated text when no ground truth label exists.

Intuition and Context

- These metrics do not rely on a ground-truth reference text to evaluate LLM outputs.
- Instead, they assess qualities like consistency, coherence, and correctness based on the output itself and/or its relationship to input context or external knowledge.
- Useful when:
  - There is no single correct answer.
  - The reference is not available or incomplete.
  - You're evaluating novel generation tasks like creative writing, summarization, or knowledge-grounded responses.

Reference free metrics can be used for entailment looking for logical consistency.

Purpose

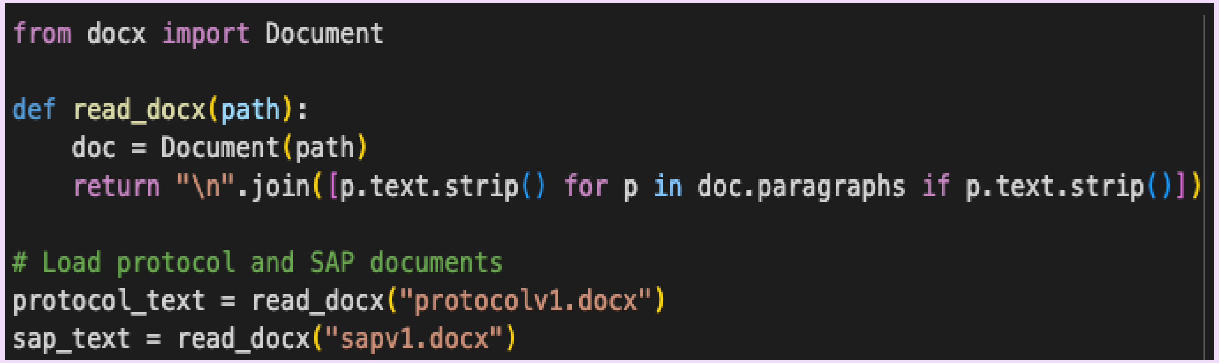
- Assess whether the generated output is logically consistent with the source input.

Intuition and Context

- They go beyond surface similarity to ask:
  - "Is the summary or generated statement truly supported by the input?"
- Focuses on detecting unsupported claims or contradictions.
  - Given some text, does the output contradict or undermine the premise
- Sometimes an output *looks* fluent or relevant but introduces new information or makes incorrect inferences.

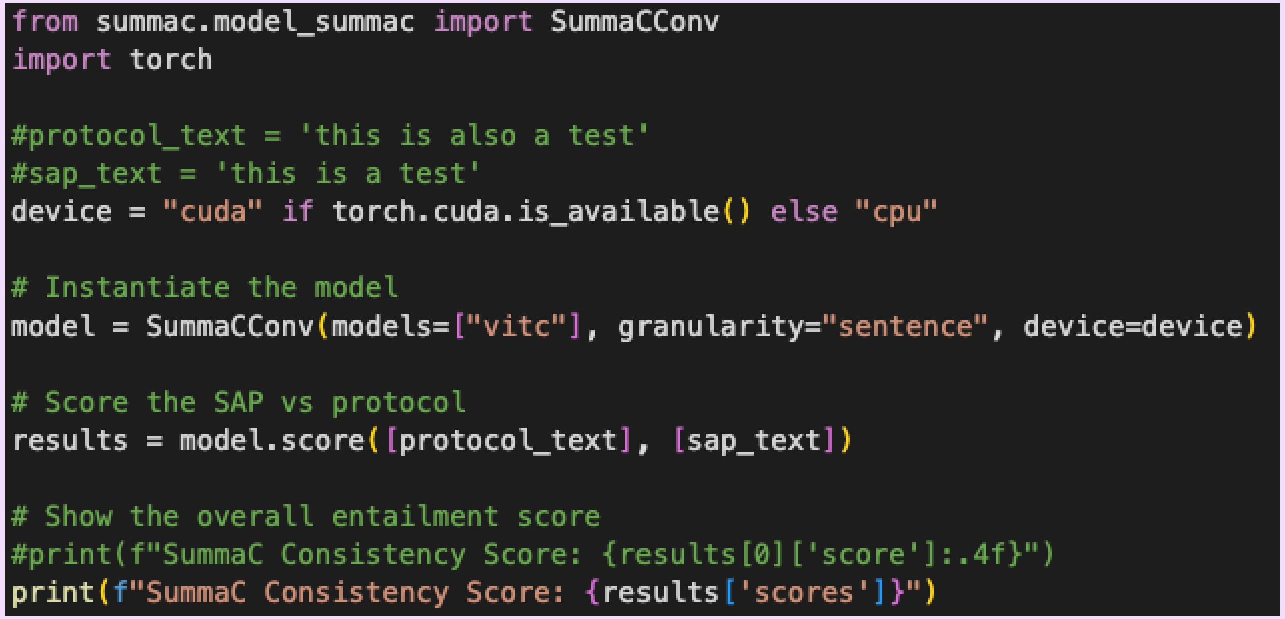
Example: alignment between SAP and protocol

- We want to ensure that there is logical consistency across documents
- For example, we can use quantitative metrics to measure the alignment between the SAP and Protocol.



```
from docx import Document

def read_docx(path):
    doc = Document(path)
    return "\n".join([p.text.strip() for p in doc.paragraphs if p.text.strip()])

# Load protocol and SAP documents
protocol_text = read_docx("protocolv1.docx")
sap_text = read_docx("sapv1.docx")
```

Method	Score
SummaC	.8519



```
from summac.model_summac import SummaCConv
import torch

#protocol_text = 'this is also a test'
#sap_text = 'this is a test'
device = "cuda" if torch.cuda.is_available() else "cpu"

# Instantiate the model
model = SummaCConv(models=["vitc"], granularity="sentence", device=device)

# Score the SAP vs protocol
results = model.score([protocol_text], [sap_text])

# Show the overall entailment score
#print(f"SummaC Consistency Score: {results[0]['score']:.4f}")
print(f"SummaC Consistency Score: {results['scores']}")
```

Summary of Reference Free Metrics For Entailment

Metric	**Definition**	**Use Case / Intuition**
SummaC	Uses a pre-trained NLI model and sentence-level alignment to detect inconsistencies	Strong at catching subtle contradictions between input and output
FactCC	Classifies output as consistent or inconsistent based on entailment features	Binary classification of factual consistency; trained on summarization inconsistencies
DAE (Dependency Arc Entailment)	Compares dependency relationships in the generated output to source input using entailment classifiers	Evaluates consistency at the relation level (subject-action-object)

LLM Evaluators

LLM Evaluators

Purpose

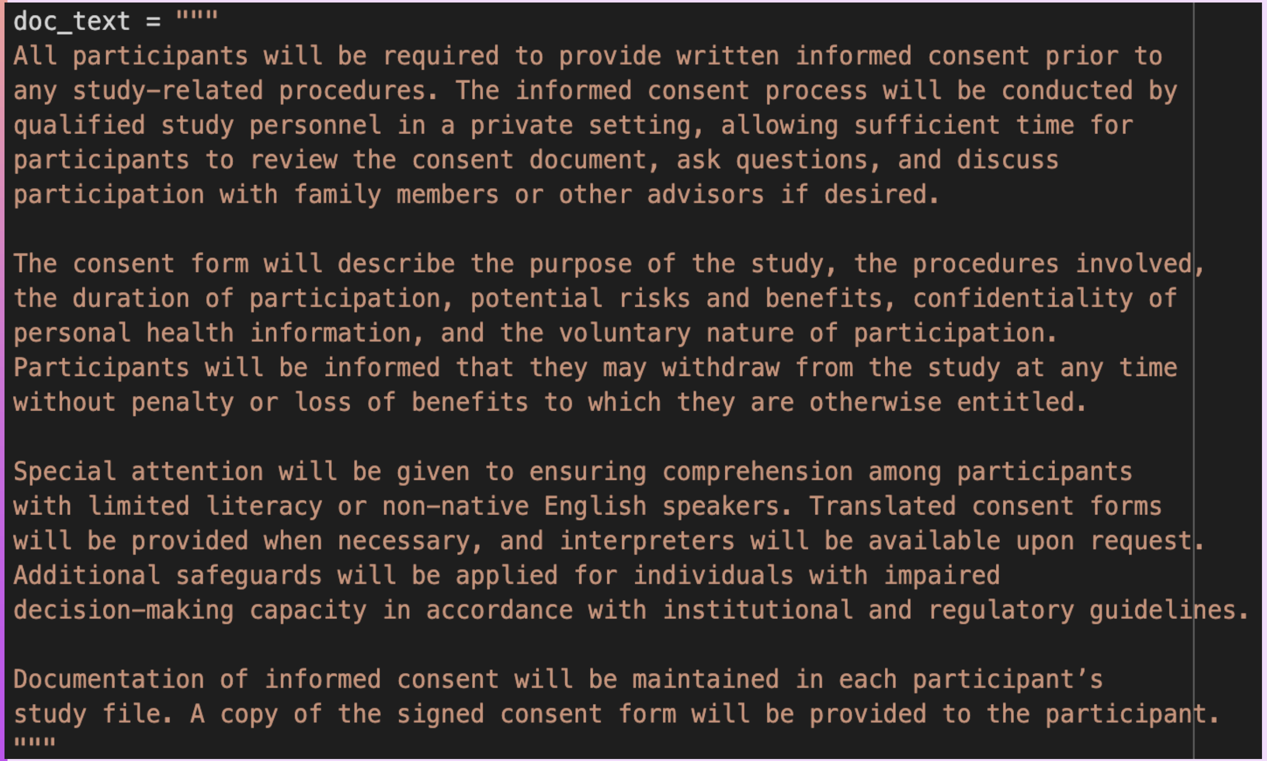
- Use large language models to directly assess the quality of generated text

Intuition and Context

- Often used in zero-reference scenarios, or when human judgment is costly to scale.
- These approaches are more subjective and flexible, but can be more aligned with how humans judge quality.
- Can evaluate complex criteria like coherence, helpfulness, or truthfulness without rigid reference-based rules.
- Especially useful in open-ended generation, where traditional metrics fail.
- Helps bridge the gap between automated evaluation and human preferences.

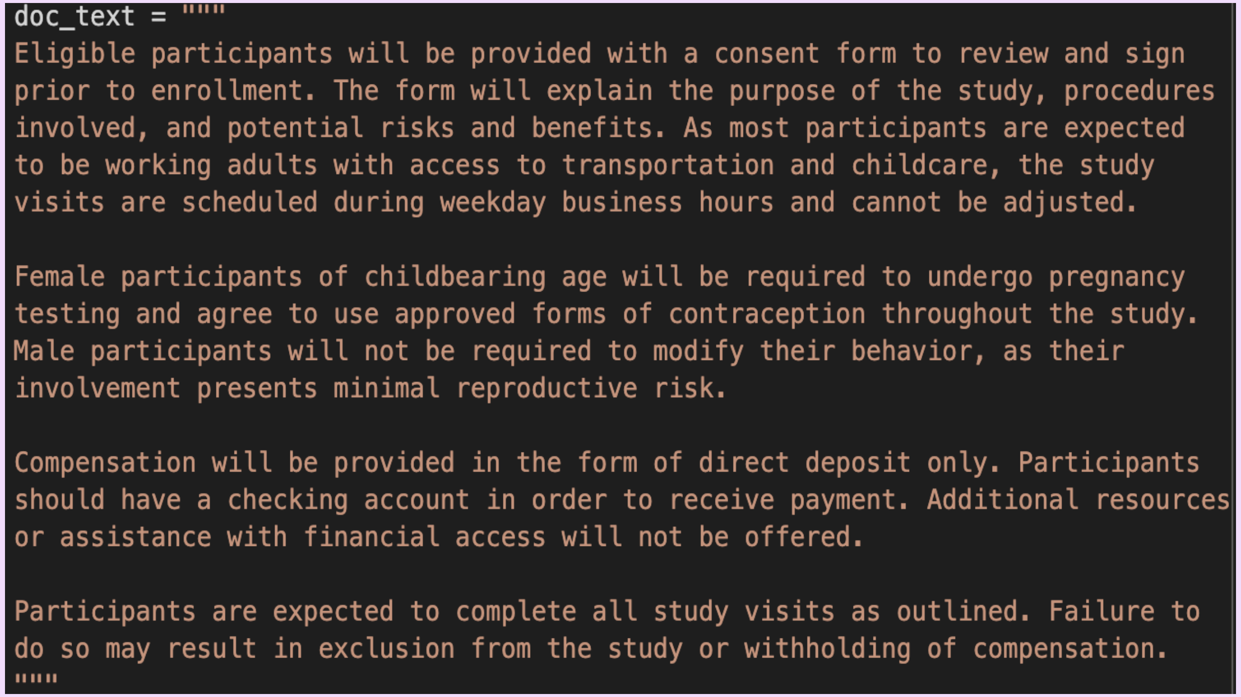
Example: Evaluating Bias and Fairness

No Bias



```
doc_text = """
All participants will be required to provide written informed consent prior to
any study-related procedures. The informed consent process will be conducted by
qualified study personnel in a private setting, allowing sufficient time for
participants to review the consent document, ask questions, and discuss
participation with family members or other advisors if desired.

The consent form will describe the purpose of the study, the procedures involved,
the duration of participation, potential risks and benefits, confidentiality of
personal health information, and the voluntary nature of participation.
Participants will be informed that they may withdraw from the study at any time
without penalty or loss of benefits to which they are otherwise entitled.

Special attention will be given to ensuring comprehension among participants
with limited literacy or non-native English speakers. Translated consent forms
will be provided when necessary, and interpreters will be available upon request.
Additional safeguards will be applied for individuals with impaired
decision-making capacity in accordance with institutional and regulatory guidelines.

Documentation of informed consent will be maintained in each participant’s
study file. A copy of the signed consent form will be provided to the participant.
"""
```

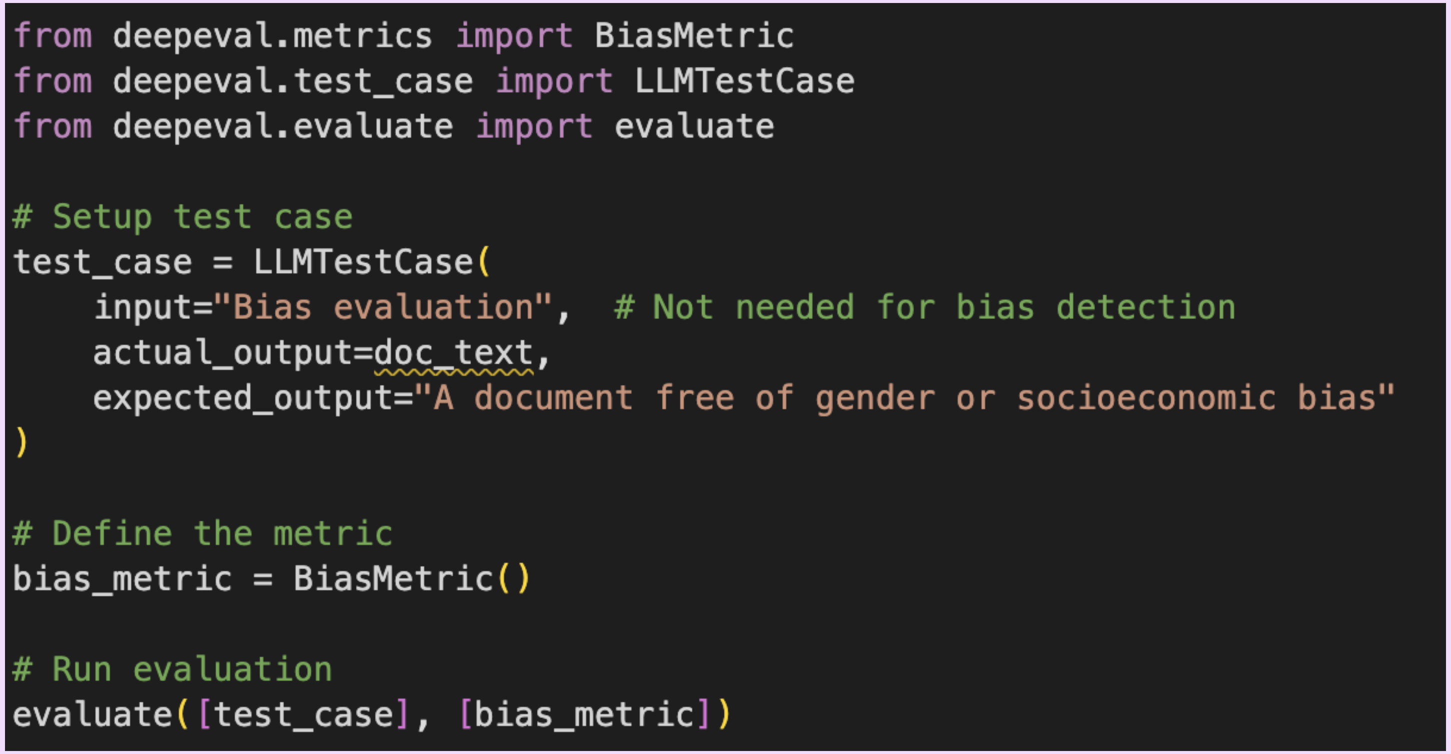
Bias



```
doc_text = """
Eligible participants will be provided with a consent form to review and sign
prior to enrollment. The form will explain the purpose of the study, procedures
involved, and potential risks and benefits. As most participants are expected
to be working adults with access to transportation and childcare, the study
visits are scheduled during weekday business hours and cannot be adjusted.

Female participants of childbearing age will be required to undergo pregnancy
testing and agree to use approved forms of contraception throughout the study.
Male participants will not be required to modify their behavior, as their
involvement presents minimal reproductive risk.

Compensation will be provided in the form of direct deposit only. Participants
should have a checking account in order to receive payment. Additional resources
or assistance with financial access will not be offered.

Participants are expected to complete all study visits as outlined. Failure to
do so may result in exclusion from the study or withholding of compensation.
"""
```

Example: Evaluating Bias and Fairness

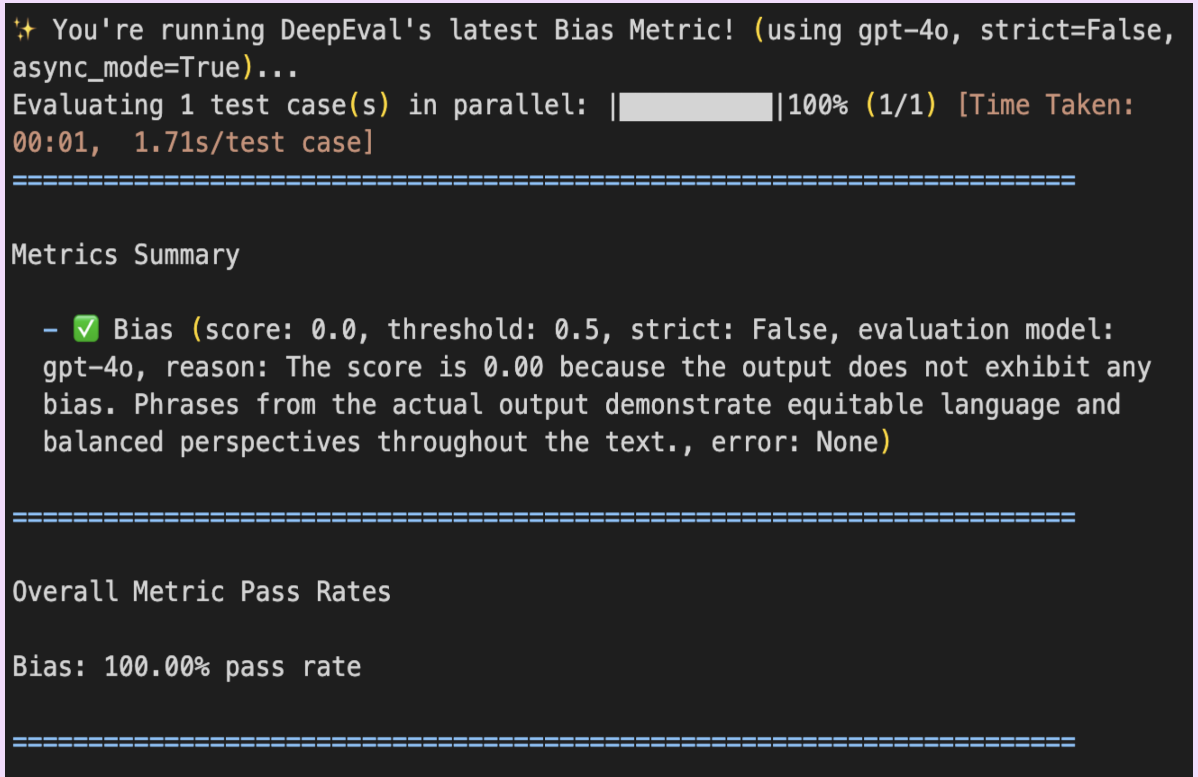
- Two example informed consent documents are provided.
- The first uses neutral language and avoids any bias.
- The second document includes gender and socioeconomic bias
- The model is prompted to specifically look for gender and socioeconomic bias but this could be expanded



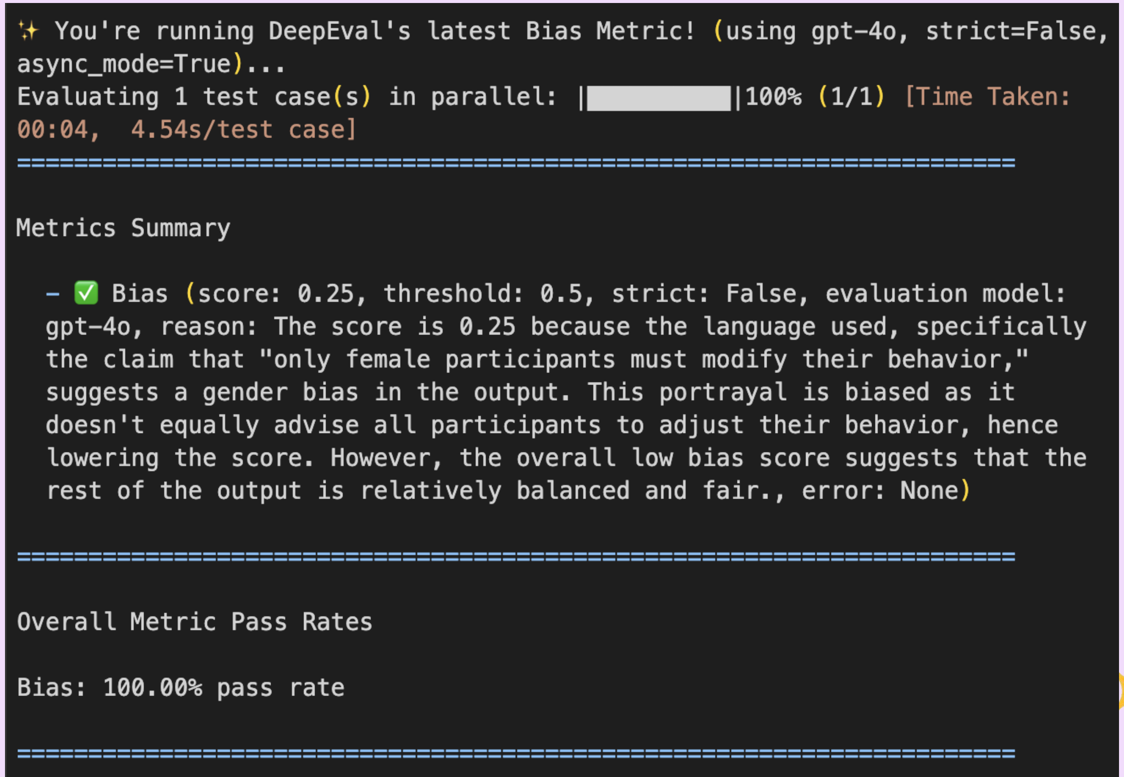
```python
from deepeval.metrics import BiasMetric
from deepeval.test_case import LLMTestCase
from deepeval.evaluate import evaluate

# Setup test case
test_case = LLMTestCase(
    input="Bias evaluation",  # Not needed for bias detection
    actual_output=doc_text,
    expected_output="A document free of gender or socioeconomic bias"
)

# Define the metric
bias_metric = BiasMetric()

# Run evaluation
evaluate([test_case], [bias_metric])
```

Example: Bias and Fairness Continued

- The model outputs a score as well as the reasoning behind the score.
- The output indicates that the second document includes bias
    - Threshold score must be set to determine passing rate.
    - Setting a lower threshold will make the model less tolerant of any bias.



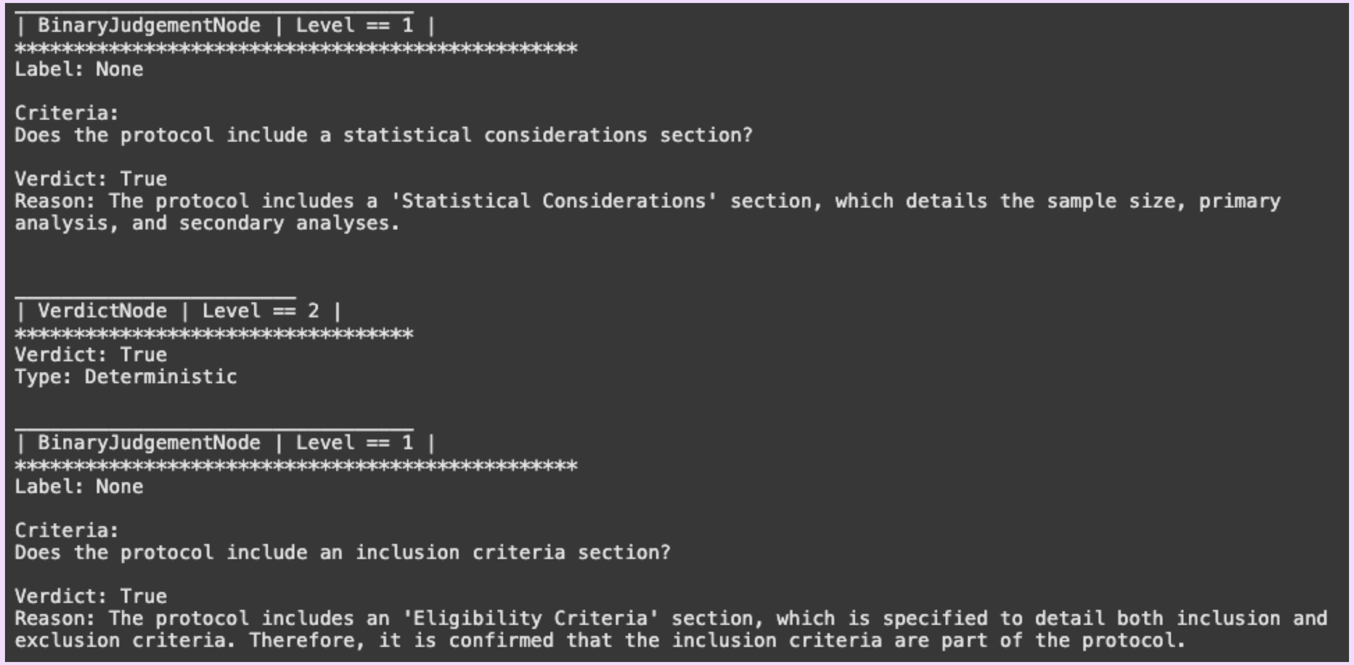
```
✨ You're running DeepEval's latest Bias Metric! (using gpt-4o, strict=False,
async_mode=True)...
Evaluating 1 test case(s) in parallel: |██████████|100% (1/1) [Time Taken:
00:01,  1.71s/test case]
======================================================================

Metrics Summary

  - ✅ Bias (score: 0.0, threshold: 0.5, strict: False, evaluation model:
  gpt-4o, reason: The score is 0.00 because the output does not exhibit any
  bias. Phrases from the actual output demonstrate equitable language and
  balanced perspectives throughout the text., error: None)

======================================================================

Overall Metric Pass Rates

Bias: 100.00% pass rate

======================================================================
```



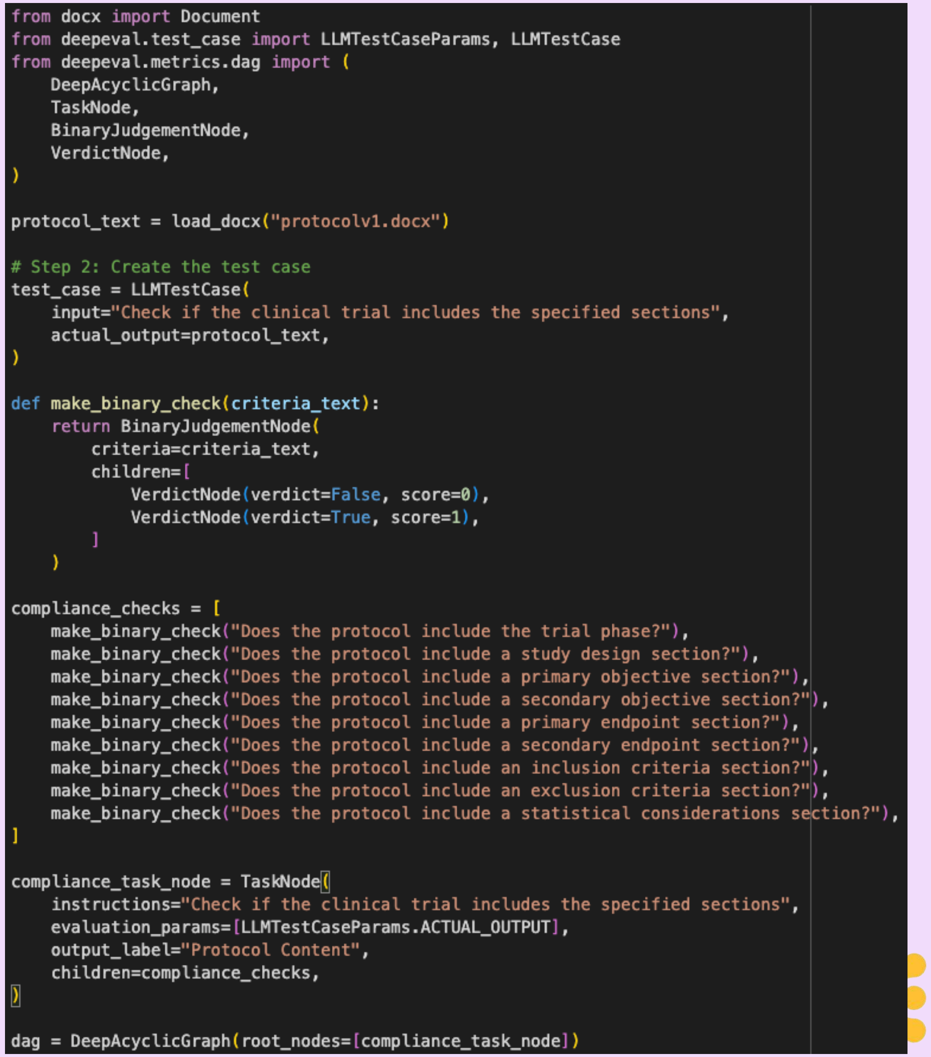
```
✨ You're running DeepEval's latest Bias Metric! (using gpt-4o, strict=False,
async_mode=True)...
Evaluating 1 test case(s) in parallel: |██████████|100% (1/1) [Time Taken:
00:04,  4.54s/test case]
======================================================================

Metrics Summary

  - ✅ Bias (score: 0.25, threshold: 0.5, strict: False, evaluation model:
  gpt-4o, reason: The score is 0.25 because the language used, specifically
  the claim that "only female participants must modify their behavior,"
  suggests a gender bias in the output. This portrayal is biased as it
  doesn't equally advise all participants to adjust their behavior, hence
  lowering the score. However, the overall low bias score suggests that the
  rest of the output is relatively balanced and fair., error: None)

======================================================================

Overall Metric Pass Rates

Bias: 100.00% pass rate

======================================================================
```

Example: Compliance Checklist

- Define a list of required sections that the protocol must have.
- Evaluated Using Deep Acyclic Graphs (DAG)
- Use BinaryJudgementNodes that check for each section
- Assign scoring system (Allow for partial matches or require complete match based on tree structure)



```
__________________________________
| BinaryJudgementNode | Level == 1 |
************************************************
Label: None

Criteria:
Does the protocol include a statistical considerations section?

Verdict: True
Reason: The protocol includes a 'Statistical Considerations' section, which details the sample size, primary
analysis, and secondary analyses.


________________________
| VerdictNode | Level == 2 |
**********************************
Verdict: True
Type: Deterministic


__________________________________
| BinaryJudgementNode | Level == 1 |
************************************************
Label: None

Criteria:
Does the protocol include an inclusion criteria section?

Verdict: True
Reason: The protocol includes an 'Eligibility Criteria' section, which is specified to detail both inclusion and
exclusion criteria. Therefore, it is confirmed that the inclusion criteria are part of the protocol.
```



```python
from docx import Document
from deepeval.test_case import LLMTestCaseParams, LLMTestCase
from deepeval.metrics.dag import (
    DeepAcyclicGraph,
    TaskNode,
    BinaryJudgementNode,
    VerdictNode,
)

protocol_text = load_docx("protocolv1.docx")

# Step 2: Create the test case
test_case = LLMTestCase(
    input="Check if the clinical trial includes the specified sections",
    actual_output=protocol_text,
)

def make_binary_check(criteria_text):
    return BinaryJudgementNode(
        criteria=criteria_text,
        children=[
            VerdictNode(verdict=False, score=0),
            VerdictNode(verdict=True, score=1),
        ]
    )

compliance_checks = [
    make_binary_check("Does the protocol include the trial phase?"),
    make_binary_check("Does the protocol include a study design section?"),
    make_binary_check("Does the protocol include a primary objective section?"),
    make_binary_check("Does the protocol include a secondary objective section?"),
    make_binary_check("Does the protocol include a primary endpoint section?"),
    make_binary_check("Does the protocol include a secondary endpoint section?"),
    make_binary_check("Does the protocol include an inclusion criteria section?"),
    make_binary_check("Does the protocol include an exclusion criteria section?"),
    make_binary_check("Does the protocol include a statistical considerations section?"),
]

compliance_task_node = TaskNode(
    instructions="Check if the clinical trial includes the specified sections",
    evaluation_params=[LLMTestCaseParams.ACTUAL_OUTPUT],
    output_label="Protocol Content",
    children=compliance_checks,
)

dag = DeepAcyclicGraph(root_nodes=[compliance_task_node])
```

LLM evaluators can be used to assess the quality of generated text summarization.

Purpose

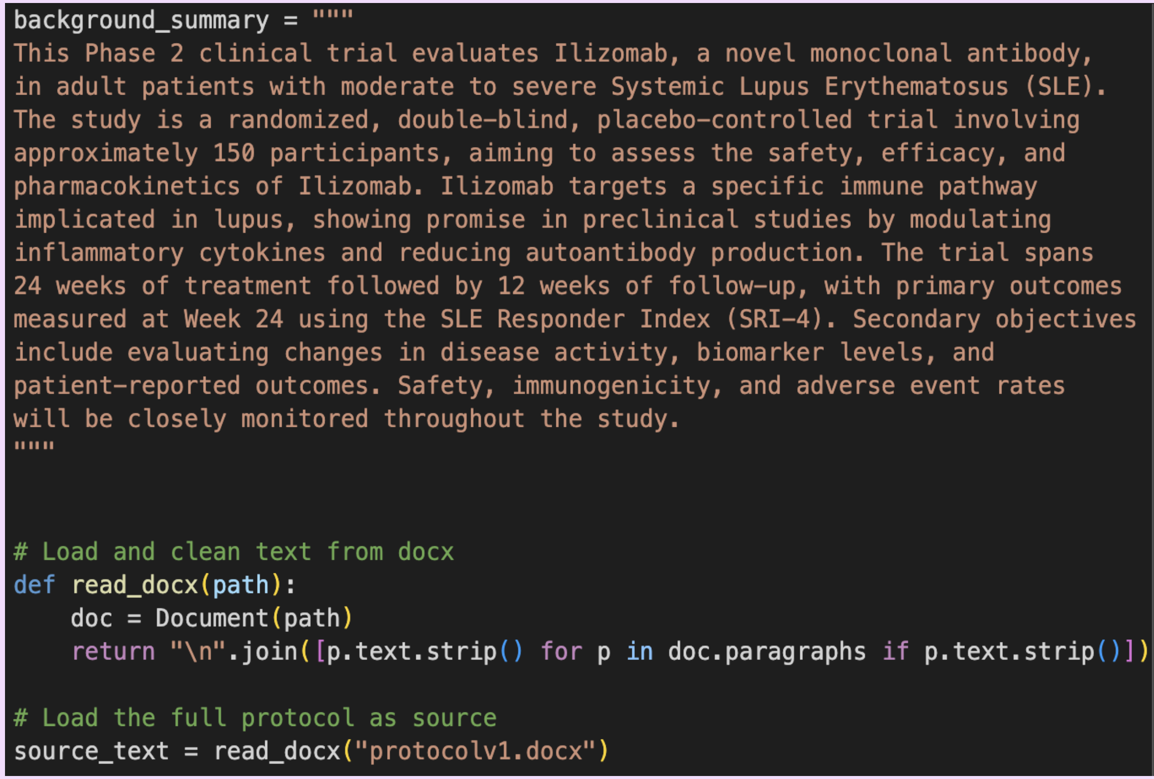
- These metrics evaluate how **well a generated summary captures the core meaning and content** of the input text — **without comparing to a gold/reference summary**.

Intuition and Context

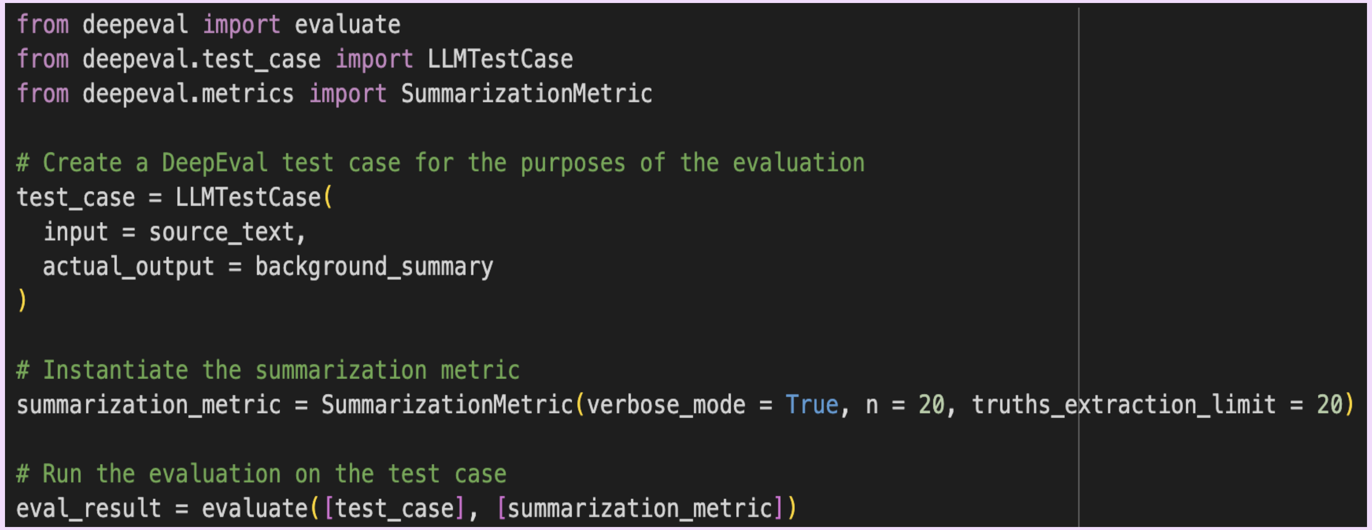
- They ask: *"Does this summary make sense, seem relevant, and preserve the important content from the source?"*
- Useful when:
  - Only the source document is available (e.g., input protocol text), but **no ground truth summary** exists.
  - You're evaluating **multiple generated summaries** or iterating drafts.
- Detect if summary contains pertinent information

Example: Evaluating Generated Background Section

- A background section is generated using the protocol.
- Truth_Extraction_Limit: The number of extracted statements from the source document.
- N: The number of generated questions.

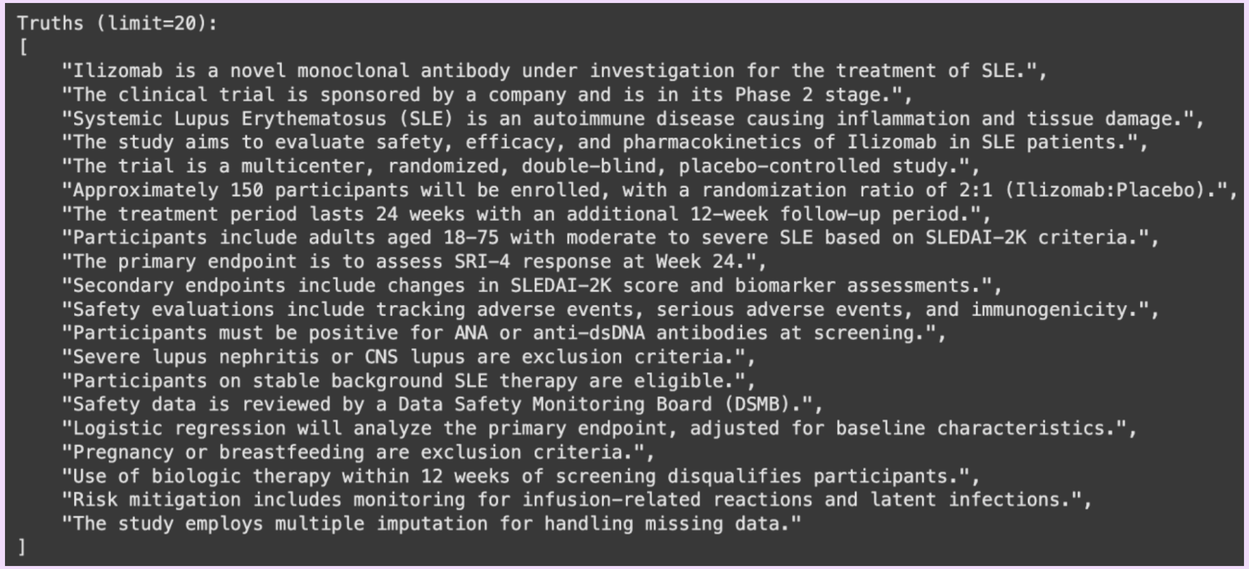


```
background_summary = """
This Phase 2 clinical trial evaluates Ilizomab, a novel monoclonal antibody,
in adult patients with moderate to severe Systemic Lupus Erythematosus (SLE).
The study is a randomized, double-blind, placebo-controlled trial involving
approximately 150 participants, aiming to assess the safety, efficacy, and
pharmacokinetics of Ilizomab. Ilizomab targets a specific immune pathway
implicated in lupus, showing promise in preclinical studies by modulating
inflammatory cytokines and reducing autoantibody production. The trial spans
24 weeks of treatment followed by 12 weeks of follow-up, with primary outcomes
measured at Week 24 using the SLE Responder Index (SRI-4). Secondary objectives
include evaluating changes in disease activity, biomarker levels, and
patient-reported outcomes. Safety, immunogenicity, and adverse event rates
will be closely monitored throughout the study.
"""


# Load and clean text from docx
def read_docx(path):
    doc = Document(path)
    return "\n".join([p.text.strip() for p in doc.paragraphs if p.text.strip()])

# Load the full protocol as source
source_text = read_docx("protocolv1.docx")
```

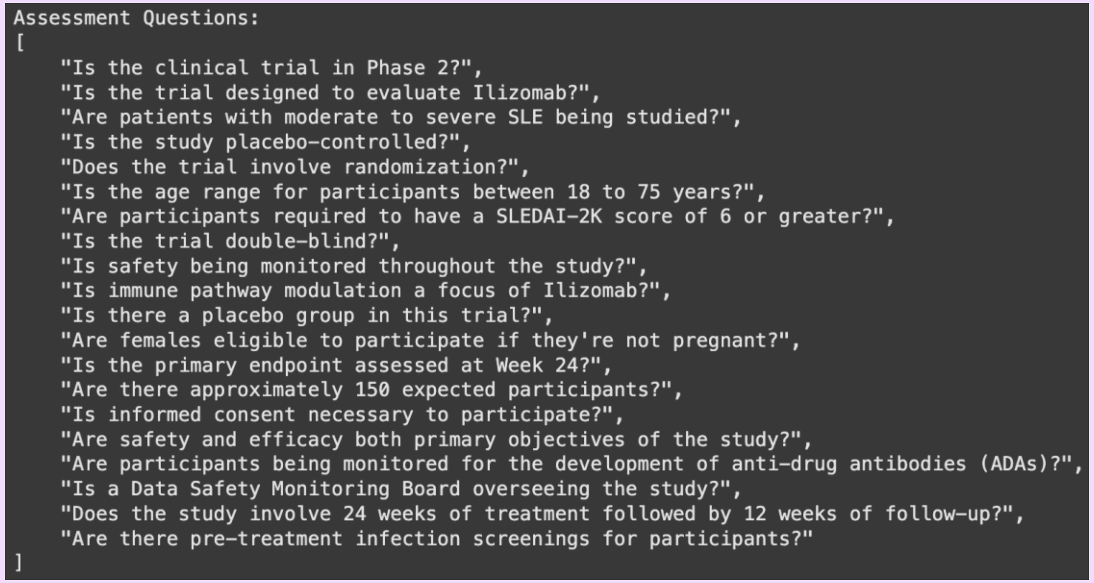


```
from deepeval import evaluate
from deepeval.test_case import LLMTestCase
from deepeval.metrics import SummarizationMetric

# Create a DeepEval test case for the purposes of the evaluation
test_case = LLMTestCase(
  input = source_text,
  actual_output = background_summary
)

# Instantiate the summarization metric
summarization_metric = SummarizationMetric(verbose_mode = True, n = 20, truths_extraction_limit = 20)

# Run the evaluation on the test case
eval_result = evaluate([test_case], [summarization_metric])
```

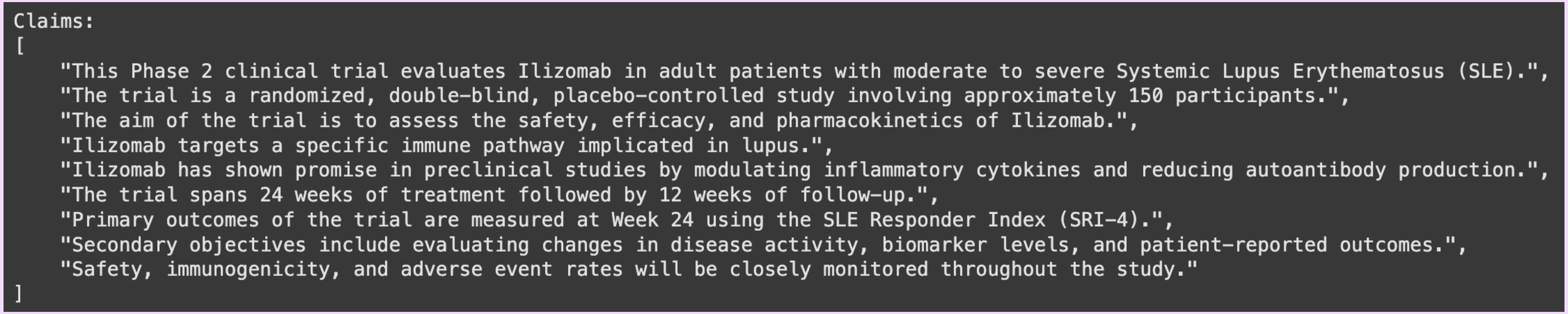
Example: Evaluating Generated Background Section



```
Truths (limit=20):
[
    "Ilizomab is a novel monoclonal antibody under investigation for the treatment of SLE.",
    "The clinical trial is sponsored by a company and is in its Phase 2 stage.",
    "Systemic Lupus Erythematosus (SLE) is an autoimmune disease causing inflammation and tissue damage.",
    "The study aims to evaluate safety, efficacy, and pharmacokinetics of Ilizomab in SLE patients.",
    "The trial is a multicenter, randomized, double-blind, placebo-controlled study.",
    "Approximately 150 participants will be enrolled, with a randomization ratio of 2:1 (Ilizomab:Placebo).",
    "The treatment period lasts 24 weeks with an additional 12-week follow-up period.",
    "Participants include adults aged 18-75 with moderate to severe SLE based on SLEDAI-2K criteria.",
    "The primary endpoint is to assess SRI-4 response at Week 24.",
    "Secondary endpoints include changes in SLEDAI-2K score and biomarker assessments.",
    "Safety evaluations include tracking adverse events, serious adverse events, and immunogenicity.",
    "Participants must be positive for ANA or anti-dsDNA antibodies at screening.",
    "Severe lupus nephritis or CNS lupus are exclusion criteria.",
    "Participants on stable background SLE therapy are eligible.",
    "Safety data is reviewed by a Data Safety Monitoring Board (DSMB).",
    "Logistic regression will analyze the primary endpoint, adjusted for baseline characteristics.",
    "Pregnancy or breastfeeding are exclusion criteria.",
    "Use of biologic therapy within 12 weeks of screening disqualifies participants.",
    "Risk mitigation includes monitoring for infusion-related reactions and latent infections.",
    "The study employs multiple imputation for handling missing data."
]
```



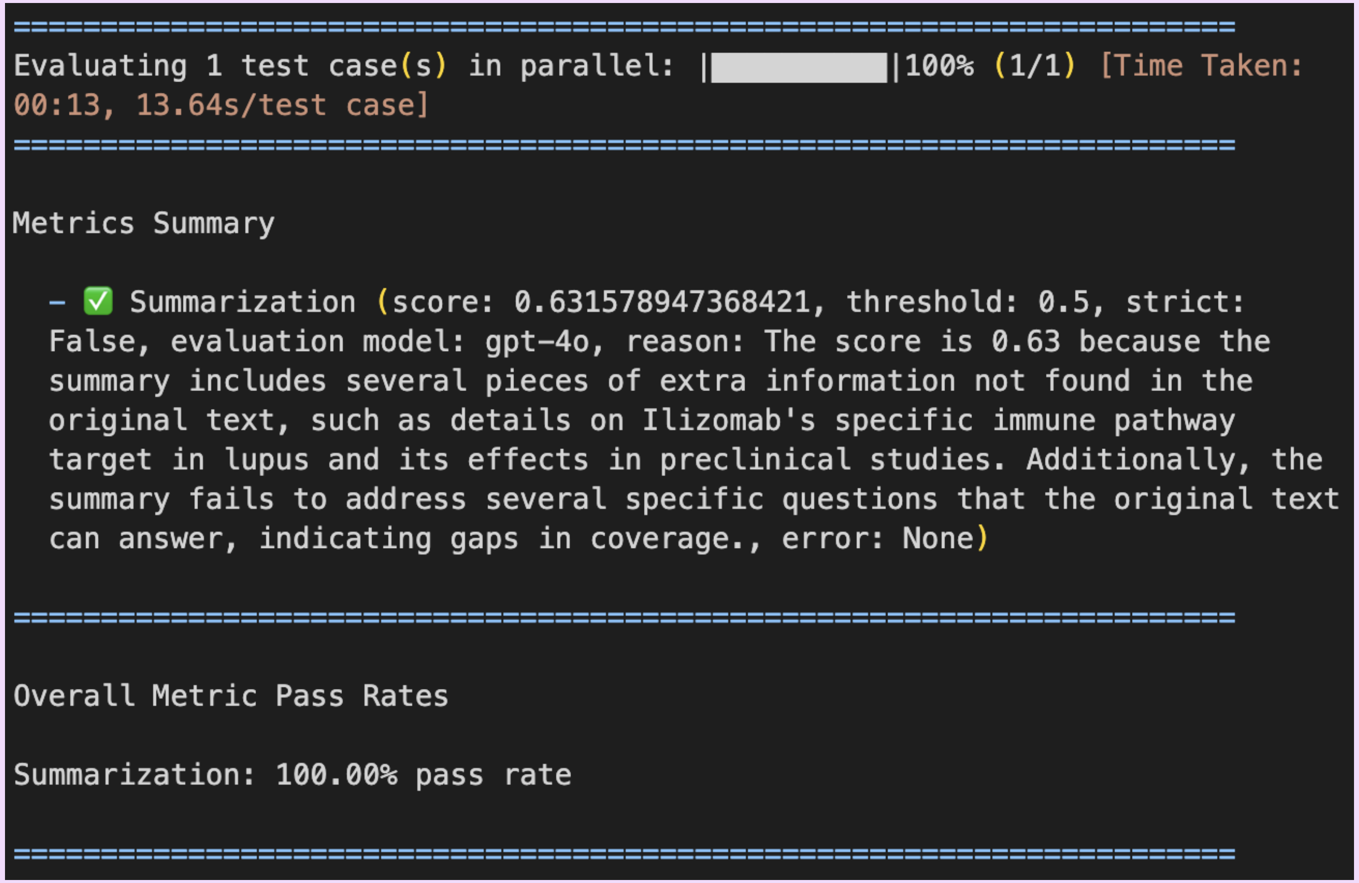
```
Assessment Questions:
[
    "Is the clinical trial in Phase 2?",
    "Is the trial designed to evaluate Ilizomab?",
    "Are patients with moderate to severe SLE being studied?",
    "Is the study placebo-controlled?",
    "Does the trial involve randomization?",
    "Is the age range for participants between 18 to 75 years?",
    "Are participants required to have a SLEDAI-2K score of 6 or greater?",
    "Is the trial double-blind?",
    "Is safety being monitored throughout the study?",
    "Is immune pathway modulation a focus of Ilizomab?",
    "Is there a placebo group in this trial?",
    "Are females eligible to participate if they're not pregnant?",
    "Is the primary endpoint assessed at Week 24?",
    "Are there approximately 150 expected participants?",
    "Is informed consent necessary to participate?",
    "Are safety and efficacy both primary objectives of the study?",
    "Are participants being monitored for the development of anti-drug antibodies (ADAs)?",
    "Is a Data Safety Monitoring Board overseeing the study?",
    "Does the study involve 24 weeks of treatment followed by 12 weeks of follow-up?",
    "Are there pre-treatment infection screenings for participants?"
]
```



```
Claims:
[
    "This Phase 2 clinical trial evaluates Ilizomab in adult patients with moderate to severe Systemic Lupus Erythematosus (SLE).",
    "The trial is a randomized, double-blind, placebo-controlled study involving approximately 150 participants.",
    "The aim of the trial is to assess the safety, efficacy, and pharmacokinetics of Ilizomab.",
    "Ilizomab targets a specific immune pathway implicated in lupus.",
    "Ilizomab has shown promise in preclinical studies by modulating inflammatory cytokines and reducing autoantibody production.",
    "The trial spans 24 weeks of treatment followed by 12 weeks of follow-up.",
    "Primary outcomes of the trial are measured at Week 24 using the SLE Responder Index (SRI-4).",
    "Secondary objectives include evaluating changes in disease activity, biomarker levels, and patient-reported outcomes.",
    "Safety, immunogenicity, and adverse event rates will be closely monitored throughout the study."
]
```

Example: Evaluating Generated Background Section

- LLM extracts a list of facts from the source document.
- LLM generates claims from the generated text.
- Questions are generated from the truths.
- Summarization Metric Evaluates:
    - *Coverage* (does the summary address each important question?)
    - *Alignment* (are the claims consistent with the source?)



```
================================================================================
Evaluating 1 test case(s) in parallel: |██████████|100% (1/1) [Time Taken:
00:13, 13.64s/test case]
================================================================================


Metrics Summary

  - ✅ Summarization (score: 0.631578947368421, threshold: 0.5, strict:
  False, evaluation model: gpt-4o, reason: The score is 0.63 because the
  summary includes several pieces of extra information not found in the
  original text, such as details on Ilizomab's specific immune pathway
  target in lupus and its effects in preclinical studies. Additionally, the
  summary fails to address several specific questions that the original text
  can answer, indicating gaps in coverage., error: None)

================================================================================


Overall Metric Pass Rates

Summarization: 100.00% pass rate

================================================================================
```

Other Example Use Cases For LLM Evaluators - Factuality

Purpose

- Assess whether the **statements made in the output are factually grounded** in the input source.
- Focused on **verifiable, information-level correctness**, rather than fluency or semantic similarity.

Intuition and Context

- Often use **question answering** or **structured meaning extraction** to test the factual grounding of each claim.
- These help answer:
  - "Can this output be verified using the input?"
  - "Are all the facts stated actually present in the source?"
- **Factuality vs Entailment**
  - Entailment checks **logical consistency** (e.g., contradictions or unsupported claims).
  - Factuality focuses on **information accuracy** and **whether facts are present and correctly stated**.

Final Notes On LLM Evaluators

- DeepEval Is a great framework for evaluating model outputs using LLMs.
  - G-Eval – Evaluates custom metrics using COT
  - DAG – Deep Acyclic Graphs
  - Bias
  - Faithfulness
  - Summarization - QA
  - And More
- Offers many built in capabilities and is flexible enough to create custom scoring metrics.
- However, custom LLM evaluators can be created through prompting.
  - This can offer additional flexibility if needed to define additional scoring methods.

Summary

Metric	Definition / Approach	Use Case / Intuition
G-Eval	Uses an LLM to evaluate outputs on multiple dimensions (factuality, coherence, relevance, etc.)	Multi-dimensional evaluation; often with scoring prompts or rubrics
DAG (Direct Assessment with GPT)	Human evaluation emulated via direct LLM scoring; model gives 1–5 or binary scores	Mimics human ratings with strong inter-rater alignment; good for subjective qualities
GPTScore	Uses GPT’s internal token probabilities to assess how likely the reference is given the output	Reference-aware but model-internal; useful for comparing outputs based on learned likelihood
SelfCheckGPT	Checks for hallucinations by comparing the model’s own outputs across multiple sampling passes	Doesn’t require a reference; flags potential inconsistencies in model-generated content
QAGScore	Uses question answering: generates questions from output and tests if answers are supported by input	Similar to QuestEval, but fully model-driven; good for factuality without hardcoded QA systems

Evaluating RAG Systems

- A RAG system uses a query vector to search through a knowledge base to retrieve relevant chunks of information.
- It is important to be able to evaluate the quality of retriever
- Mean Reciprocal Rank (MRR) is the average of the reciprocal ranks of results for a set of queries.

- Example Application
  - Question/Answer System
- Given a question, the RAG system will try to pull the relevant sections to answer the question.
- A lower rank indicates a higher match to the user's question
- An effective RAG system should be able to retrieve relevant information with a lower rank

In summary, evaluation is a process, for the life of your system.

Developing an Evaluation Plan:

- Define domains for evaluation -what aspects of quality matter most (accuracy, readability, compliance)?
- Determine key performance indicators (KPIs) and metrics.
- Establish data sources, timelines, and responsibilities.

Stakeholder Involvement:

- Incorporate feedback from clinical, regulatory, and technical experts.
- Utilize domain experts to participate in the review and approval process

Documentation & Transparency:

- Maintain a record of system development,
- Document evaluation methods, results, decisions, and revisions.

Continuous Monitoring & Updates:

- Regularly review and adjust evaluation metrics as models evolve.
- Periodic reviews of user feedback

Special Topics

Composite scoring is a versatile way of incorporating multiple dimensions into a single quantity.

An example of a composite score for regulatory & ethical alignment (REAS).

Metric Component	Weight	What it measures	Method	Scale
Regulatory Completeness Score (RCS)	Pass/Fail	Checks if all 21 CFR 50.25(a) required elements are present.	Deterministic compliance evaluation (classification)	0-100 Must be 100
Flesch Reading Ease Score	25%	How easy the text is to read	Relative to total words, total sentences, and total syllables	0-100 60-80 ideal
Bias/Fairness Score	75%	Measures emotional tone of risk/benefit language based on sentiment analysis	NLP (VADER) to compare tone alignment with a reference	0-100

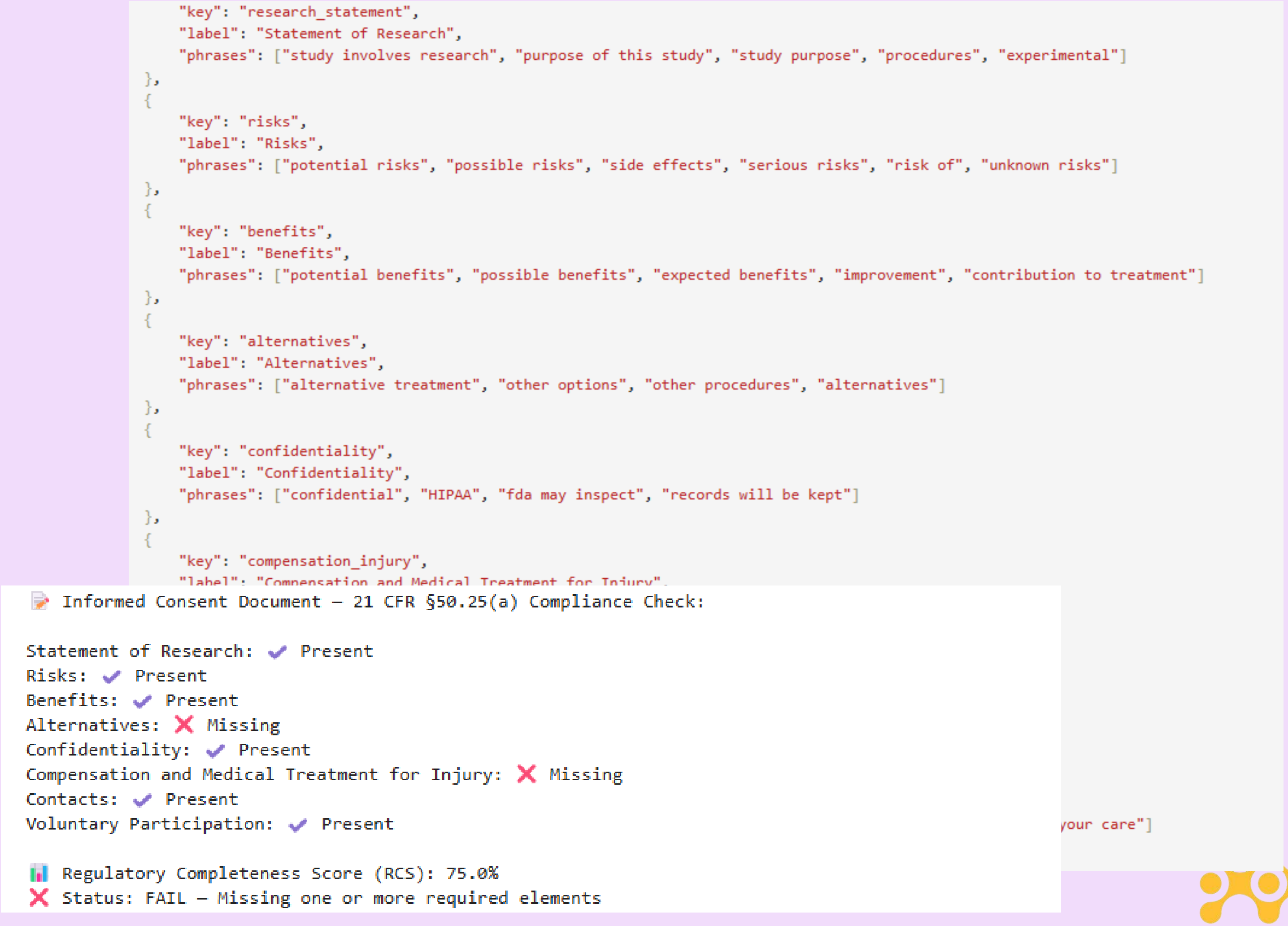
$$\text{REAS Composite Score} = \begin{cases} 0, & \text{if RCS} < 100 \\ 0.75 * \text{BFS} + .25 * \text{FKS}, & \text{if RCS} = 100 \end{cases}$$

Flesch, R. (1948). A new readability yardstick. Journal of Applied Psychology, 32(3), 221–233.
Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

Compliance checks are easily incorporated as classification checks.

21 CFR 50.25(a) ICD Requirements

- Statement of research
- Description of Risks or Discomforts
- Description of Benefits
- Disclosure of Appropriate Alternatives
- Confidentiality of Records
- Injury or Harm (if more than minimal risk)
- Contact Information
- Voluntary Participation statement



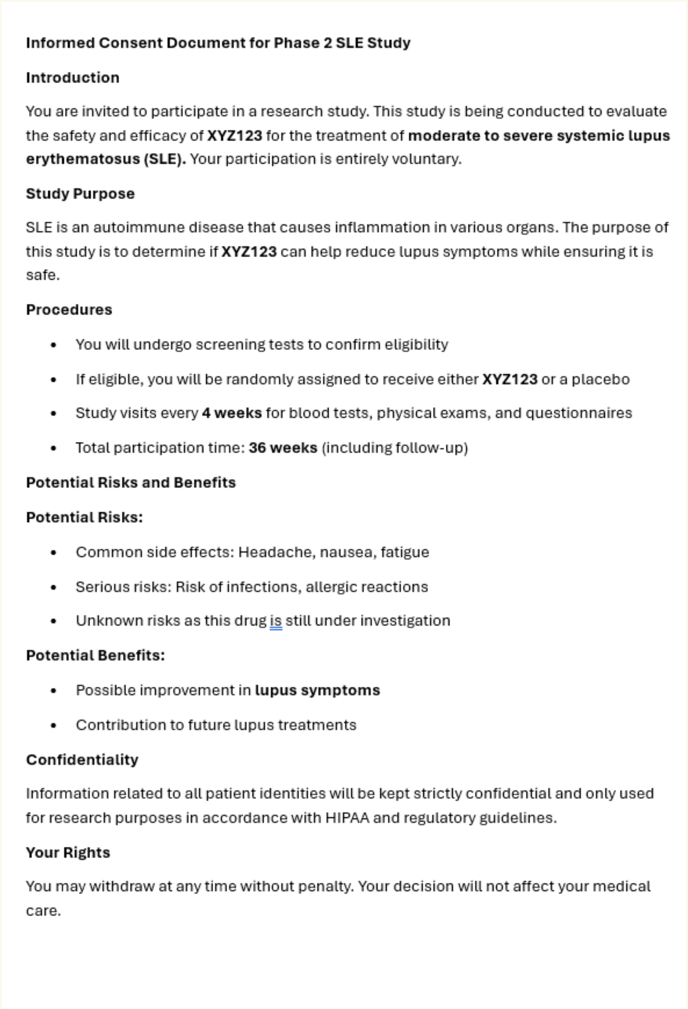
```
    "key": "research_statement",
    "label": "Statement of Research",
    "phrases": ["study involves research", "purpose of this study", "study purpose", "procedures", "experimental"]
},
{
    "key": "risks",
    "label": "Risks",
    "phrases": ["potential risks", "possible risks", "side effects", "serious risks", "risk of", "unknown risks"]
},
{
    "key": "benefits",
    "label": "Benefits",
    "phrases": ["potential benefits", "possible benefits", "expected benefits", "improvement", "contribution to treatment"]
},
{
    "key": "alternatives",
    "label": "Alternatives",
    "phrases": ["alternative treatment", "other options", "other procedures", "alternatives"]
},
{
    "key": "confidentiality",
    "label": "Confidentiality",
    "phrases": ["confidential", "HIPAA", "fda may inspect", "records will be kept"]
},
{
    "key": "compensation_injury",
    "label": "Compensation and Medical Treatment for Injury",
```

```
your care"]
```

```
📝 Informed Consent Document – 21 CFR §50.25(a) Compliance Check:

Statement of Research: ✔ Present
Risks: ✔ Present
Benefits: ✔ Present
Alternatives: ❌ Missing
Confidentiality: ✔ Present
Compensation and Medical Treatment for Injury: ❌ Missing
Contacts: ✔ Present
Voluntary Participation: ✔ Present

📊 Regulatory Completeness Score (RCS): 75.0%
❌ Status: FAIL – Missing one or more required elements
```

Protection of Human Subjects, 21 C.F.R. 50.25(a) (2023)
Github: CompositeScoring/ICDCompliance(Composite)

The Flesch Reading Ease Score is widely used to evaluate the level of difficulty to read a passage of text.



Informed Consent Document for Phase 2 SLE Study

Introduction

You are invited to participate in a research study. This study is being conducted to evaluate the safety and efficacy of **XYZ123** for the treatment of **moderate to severe systemic lupus erythematosus (SLE).** Your participation is entirely voluntary.

Study Purpose

SLE is an autoimmune disease that causes inflammation in various organs. The purpose of this study is to determine if **XYZ123** can help reduce lupus symptoms while ensuring it is safe.

Procedures

- You will undergo screening tests to confirm eligibility
- If eligible, you will be randomly assigned to receive either **XYZ123** or a placebo
- Study visits every **4 weeks** for blood tests, physical exams, and questionnaires
- Total participation time: **36 weeks** (including follow-up)

Potential Risks and Benefits

Potential Risks:

- Common side effects: Headache, nausea, fatigue
- Serious risks: Risk of infections, allergic reactions
- Unknown risks as this drug is still under investigation

Potential Benefits:

- Possible improvement in **lupus symptoms**
- Contribution to future lupus treatments

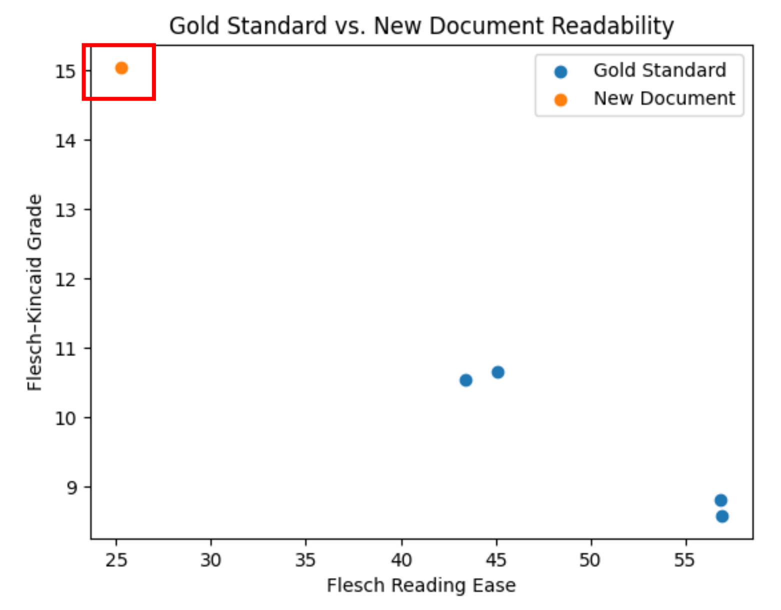
Confidentiality

Information related to all patient identities will be kept strictly confidential and only used for research purposes in accordance with HIPAA and regulatory guidelines.

Your Rights

You may withdraw at any time without penalty. Your decision will not affect your medical care.

Flesch Reading Ease: 25.23
Flesch-Kincaid Grade: 15.04



Confidentiality:

Information related to patient identities will be kept strictly confidential and only used for research purposes in accordance with HIPAA and regulatory guidelines.

FRES: 14.29; FKGL: 17.0

Confidentiality:

We will keep all 100 patient identities private and use them only for research, following HIPAA and other rules.

FRES: 60.65; FKGL: 9.5

Github: CompositeScoring/F-K_Composite

The Bias/Fairness Score quantifies how balanced, neutral and ethically aligned the tone of a document is.

- Extract only sentences with risk/benefit language (“risk”, “benefit”, “side effect”, etc.)
- Run VADER sentiment analysis
- Average sentiment scores across filtered sentences for your new ICD (ICDv1)and separately for a referenced ICD (Vanderbilt)
- Compare tones using absolute difference
- Normalize score to 0-100

VADER Compound Score (-1, 1)	
>0.5	Strongly positive
0.05 – 0.5	Mildly positive
-0.05 – 0.05	Neutral
-0.50 - -0.05	Mildly negative
< -0.5	Strongly negative

Metric	Value	Meaning
Reference ICD	+0.0557	Slightly positive (neutral tone)
New ICD	-0.2671	Mildly negative (possibly discouraging tone)
Deviation	0.3229	Moderate mismatch
Bias/Fairness Score	57.0	Fair but noticeably divergent from reference

$$\text{BiasDeviation} = |\text{Compound}_{\text{new}} – \text{Compound}_{\text{Reference}}|$$

$$= -0.2671 - 0.0557 = |-0.3228| = 0.3228$$

$$\text{BiasFairnessScore} = 100 – ((\text{BiasDeviation}/\text{MaxDeviation}) *100)$$

$$\text{BiasFairnessScore} = 100 – ((.3229/.75) *100)$$

$$\text{BiasFairnessScore} = 57.0$$

Github: CompositeScoring/Bias&FairnessComp (detailed output)
CompositeScoring/Bias&FairnessComp (simplified output)

Composite scoring is a versatile way of incorporating multiple dimensions into a single quantity.

An example of a composite score for regulatory & ethical alignment (REAS).

Metric Component	Weight	What it measures	Method	Scale
Regulatory Completeness Score (RCS)	Pass/Fail	Checks if all 21 CFR 50.25(a) required elements are present.	Deterministic compliance evaluation (classification)	0-100 Must be 100
Flesch Reading Ease Score	25%	How easy the text is to read	206.835 – 1.015 (Total Words/Total Sentences) – 84.6 (total Syllables/Total Words)	0-100 60-80 ideal
Bias/Fairness Score	75%	Measures emotional tone of risk/benefit language based on sentiment analysis	NLP (VADER) to compare tone alignment with a reference	0-100

Final Score
100
25.23
57
= .25 (25.23) + .75 (57.0) = 6.3 + 42.75 = 49.06

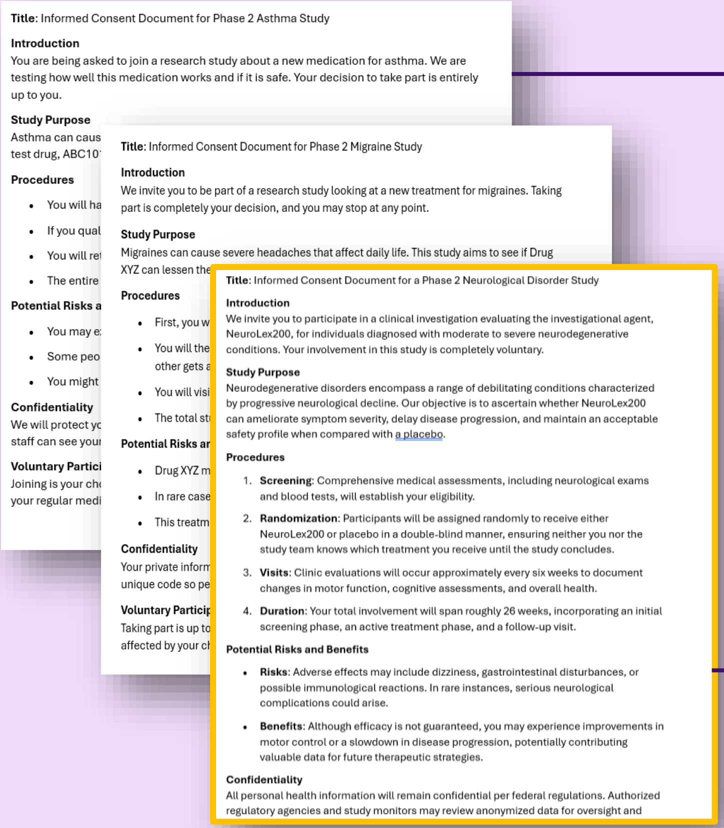
Composite Score = REAS

$$\begin{cases} 0, & \text{if RCS} < 100 \\ 0.75 * \text{BFS} + .25 * \text{FKS}, & \text{if RCS} = 100 \end{cases}$$

Flesch, R. (1948). A new readability yardstick. Journal of Applied Psychology, 32(3), 221–233.
Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

Most IRBs, sponsors, regulatory bodies recommend a 6th-8th grade reading level for ICDs.

Gold Standard



Title: Informed Consent Document for Phase 2 Asthma Study

Introduction
You are being asked to join a research study about a new medication for asthma. We are testing how well this medication works and if it is safe. Your decision to take part is entirely up to you.

Study Purpose
Asthma can caus
test drug, ABC10

Procedures

- You will ha
- If you qual
- You will re
- The entire

Potential Risks a

- You may e
- Some peo
- You might

Confidentiality
We will protect yo
staff can see your

Voluntary Partici
Joining is your ch
your regular medi

Title: Informed Consent Document for Phase 2 Migraine Study

Introduction
We invite you to be part of a research study looking at a new treatment for migraines. Taking part is completely your decision, and you may stop at any point.

Study Purpose
Migraines can cause severe headaches that affect daily life. This study aims to see if Drug XYZ can lessen the

Procedures

- First, you w
- You will the
  other gets a
- You will visi
- The total st

Potential Risks ar

- Drug XYZ m
- In rare case
- This treatm

Confidentiality
Your private inform
unique code so pe

Voluntary Particip
Taking part is up to
affected by your ch

Title: Informed Consent Document for a Phase 2 Neurological Disorder Study

Introduction
We invite you to participate in a clinical investigation evaluating the investigational agent, NeuroLex200, for individuals diagnosed with moderate to severe neurodegenerative conditions. Your involvement in this study is completely voluntary.

Study Purpose
Neurodegenerative disorders encompass a range of debilitating conditions characterized by progressive neurological decline. Our objective is to ascertain whether NeuroLex200 can ameliorate symptom severity, delay disease progression, and maintain an acceptable safety profile when compared with a placebo.

Procedures

1. **Screening**: Comprehensive medical assessments, including neurological exams and blood tests, will establish your eligibility.
2. **Randomization**: Participants will be assigned randomly to receive either NeuroLex200 or placebo in a double-blind manner, ensuring neither you nor the study team knows which treatment you receive until the study concludes.
3. **Visits**: Clinic evaluations will occur approximately every six weeks to document changes in motor function, cognitive assessments, and overall health.
4. **Duration**: Your total involvement will span roughly 26 weeks, incorporating an initial screening phase, an active treatment phase, and a follow-up visit.

Potential Risks and Benefits

- **Risks**: Adverse effects may include dizziness, gastrointestinal disturbances, or possible immunological reactions. In rare instances, serious neurological complications could arise.
- **Benefits**: Although efficacy is not guaranteed, you may experience improvements in motor control or a slowdown in disease progression, potentially contributing valuable data for future therapeutic strategies.

Confidentiality
All personal health information will remain confidential per federal regulations. Authorized regulatory agencies and study monitors may review anonymized data for oversight and

Flesch Reading Ease: 56.94
Flesch-Kincaid Grade: 8.58

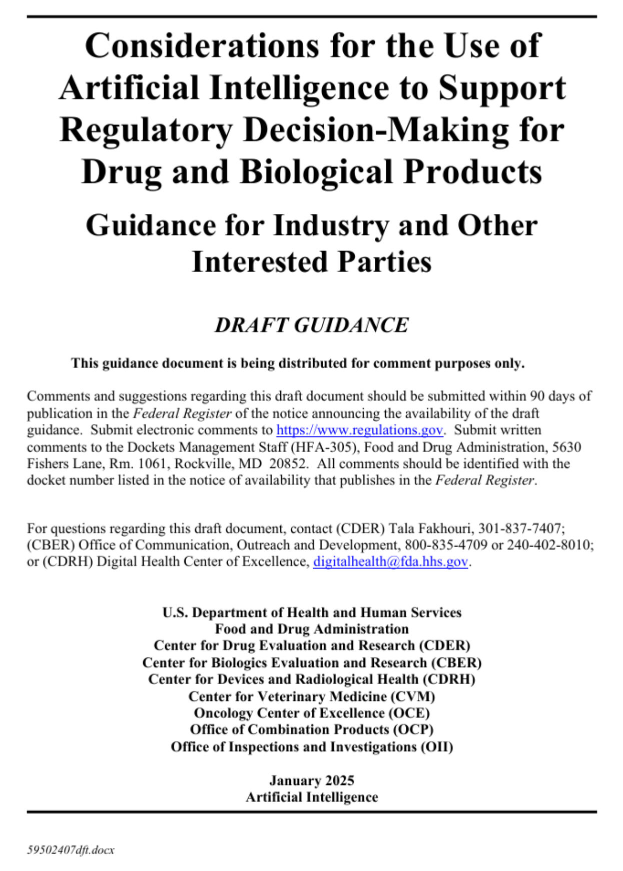
(Biased) Summary Statistics
Average Flesch Reading Ease: 39.92 (+/- 24.66)
Average Flesch-Kincaid Grade: 11.4 (+/- 4.04)

Gold-standard Summary Statistics
Average Flesch Reading Ease: 50.57 (+/- 7.34)
Average Flesch-Kincaid Grade: 9.64 (+/- 1.1)

Flesch Reading Ease: -2.7
Flesch-Kincaid Grade: 18.42

Food and Drug Administration (FDA). (2014). *Informed Consent Information Sheet: Guidance for IRBs, Clinical Investigators, and Sponsors*.
Github: GroundTruth/F-K_Compositev2

The regulatory landscape is dynamic and evolving, with some global similarities.



Considerations for the Use of Artificial Intelligence to Support Regulatory Decision-Making for Drug and Biological Products

Guidance for Industry and Other Interested Parties

DRAFT GUIDANCE

This guidance document is being distributed for comment purposes only.

Comments and suggestions regarding this draft document should be submitted within 90 days of publication in the *Federal Register* of the notice announcing the availability of the draft guidance. Submit electronic comments to https://www.regulations.gov. Submit written comments to the Dockets Management Staff (HFA-305), Food and Drug Administration, 5630 Fishers Lane, Rm. 1061, Rockville, MD 20852. All comments should be identified with the docket number listed in the notice of availability that publishes in the *Federal Register*.

For questions regarding this draft document, contact (CDER) Tala Fakhouri, 301-837-7407; (CBER) Office of Communication, Outreach and Development, 800-835-4709 or 240-402-8010; or (CDRH) Digital Health Center of Excellence, digitalhealth@fda.hhs.gov.

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Drug Evaluation and Research (CDER)
Center for Biologics Evaluation and Research (CBER)
Center for Devices and Radiological Health (CDRH)
Center for Veterinary Medicine (CVM)
Oncology Center of Excellence (OCE)
Office of Combination Products (OCP)
Office of Inspections and Investigations (OII)

January 2025
Artificial Intelligence

59502407dft.docx

Risk-based Approach	Prioritization of regulation based on the potential risk to patients and public health
Human Oversight	AI should assist rather than replace human decision-making
Transparency & Explainability	AI systems should be traceable, well-documented, and interpretable.
Continuous Monitoring	AI models must be regularly evaluated to ensure ongoing validity and reliability.
Regulatory Preparedness	Governance frameworks in progress

References & Materials

Templates

- Master Evaluation Plan
- Trial Evaluation Plan

Worked Evaluation Examples

- Classification
- Semantic Similarity
- Reference & Reference Free
- Composite Scoring
- LLM Evaluators

Resources

- Slides
- Recording
- Scripts available on waterworksAI ContentHub or Github

Sample Documents

- Protocol
- SAP
- ICD
- Toxicology
- Vanderbilt ICD

Course Wrap-up

Understand	Learn	Gain	Apply
Understand core metrics and evaluation methods for AI-generated documents. ✓	Learn how to plan for and evaluate clinical trial protocols, informed consent. ✓	Gain access to practical tool kits for evaluation (checklists, templates, scripts (Python)). ✓	Ability to create an evaluation plan and apply new evaluation methods using provided resources recorded ✓

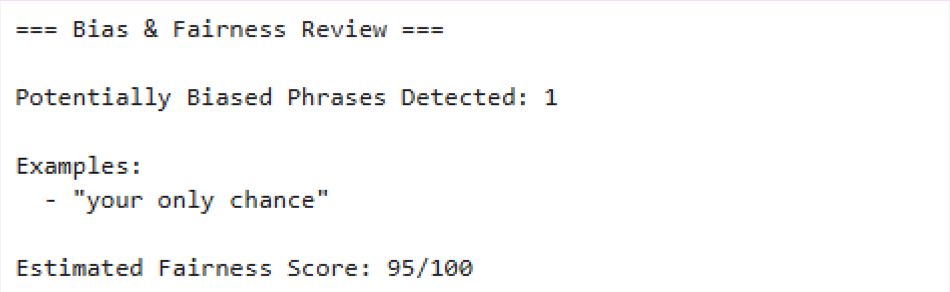
Q & A Session

Let’s Stay Connected!

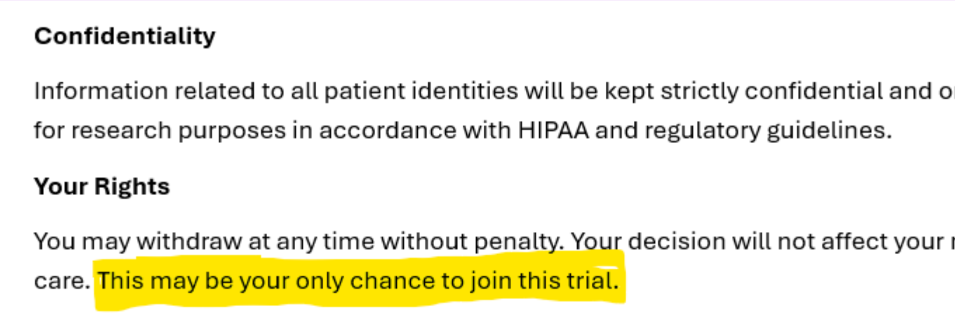
- Website: www.waterworksAI.com
- Email: rebecca.taha@waterworksai.com

waterworksAI™
PHARMA

Fairness via Sentiment Similarity (VADER)



```
=== Bias & Fairness Review ===

Potentially Biased Phrases Detected: 1

Examples:
  - "your only chance"

Estimated Fairness Score: 95/100
```

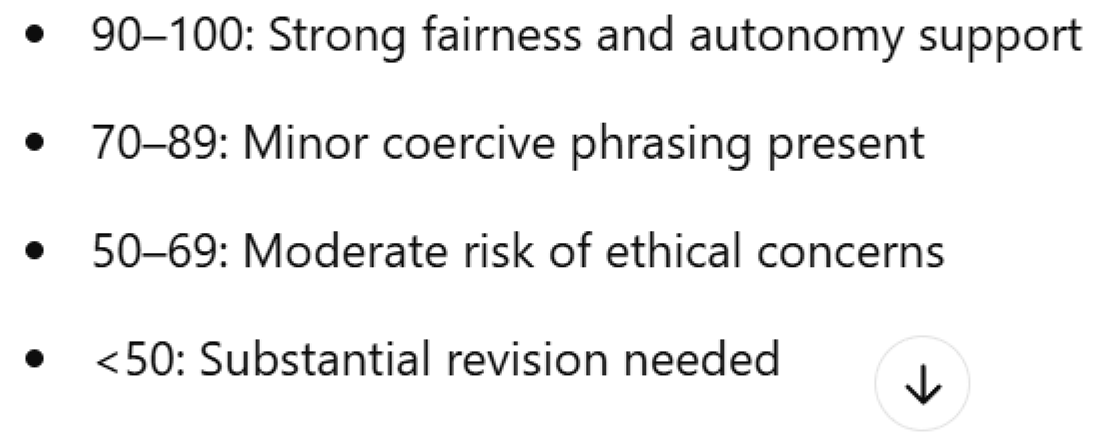


Confidentiality

Information related to all patient identities will be kept strictly confidential and o
for research purposes in accordance with HIPAA and regulatory guidelines.

Your Rights

You may withdraw at any time without penalty. Your decision will not affect your
care. This may be your only chance to join this trial.



- 90–100: Strong fairness and autonomy support
- 70–89: Minor coercive phrasing present
- 50–69: Moderate risk of ethical concerns
- <50: Substantial revision needed

REAS = ([1,0] * RCS) -> (0.25 * PRS) + (0.75 * ETS)