



Evaluación del Modelo LLM

RGA YoCampo Prototipo

Análisis de métricas para evaluación
Modelos LLM-GPT 4o

Autores: Victor Manuel Mondragon

Maca – Laboratorio de Datos SNUIRA

Versión:1.0

Fecha: 28 de agosto de 2024

Unidad de Planificación Rural Agropecuaria (UPRA)

Calle 28 N° 13-22, Torre C, piso 3. Edif. Palma Real. Bogotá, Colombia.

+57(601) 552 9820, 245 7307

www.upra.gov.co



Este documento es propiedad intelectual de la Unidad de Planificación Rural Agropecuaria (UPRA). Solo se permite su reproducción parcial, cuando no se use con fines comerciales, citando este documento así: Apellido del autor, Inicial del nombre. (2024). Título del documento. Bogotá: UPRA. Recuperado de <URL de ubicación del documento>.

Unidad de Planificación Rural Agropecuaria (UPRA)

Calle 28 N° 13-22, Torre C, piso 3. Edif. Palma Real. Bogotá, Colombia.

+57(601) 552 9820, 245 7307

www.upra.gov.co



Resumen

Este documento proporciona un análisis de la evaluación del modelo LLM RGA YoCampo utilizando métricas como **coherencia**, **fluidez**, **fundamentación** y **relevancia** [1]. Este estudio evalúa el rendimiento del modelo LLM RGA YoCampo, basado en GPT-4o, diseñado para responder preguntas sobre agricultura, específicamente en el cultivo de cacao. La evaluación se centró en métricas clave como fluidez, coherencia, fundamentación, relevancia, similitud y F1 score para medir la calidad y precisión de las respuestas generadas. Los resultados mostraron un excelente rendimiento en fluidez (5/5) y coherencia (4.13/5), destacando la capacidad del modelo para generar respuestas claras y correctas. Sin embargo, se identificaron áreas de mejora en fundamentación (4.60/5) y relevancia (4.75/5), donde algunas respuestas no aprovecharon completamente el contexto o carecieron de especificidad. La similitud con las respuestas correctas fue moderada (4/5), lo que sugiere la necesidad de mejorar la alineación con el vocabulario técnico y la estructura esperada.



Tabla de contenido

Resumen.....	3
Índice de tablas	6
Índice de figuras	7
Lista de siglas y abreviaturas	8
Glosario	9
Introducción	11
Objetivo general.....	11
Objetivos	11
Alcance	12
1. Metodología de Evaluación	13
1.1. Definición del Escenario de Evaluación	13
1.2. Configuración de los Datos de Prueba	13
1.3. Selección de Métricas de Evaluación	14
1.4. Análisis de Resultados y Ajustes.....	15
1.5. Reevaluación y Ajuste Continuo	15
2. Evaluación	16
2.1. Preparación del Conjunto de Datos.	16
2.2. Datos de Prueba	17
2.2.1. Criterios para la Creación de las Preguntas	17
2.2.2. Selección del Contexto.....	18
2.2.3. Valores de latency y answer_length	19
2.3. Selección de las métricas a ser evaluadas.....	26
2.3.1. Groundedness (Fundamentación)	27
2.3.2. Relevance (Relevancia)	28
2.3.3. Coherence (Coherencia)	28
2.3.4. Fluency (Fluidez).....	29
2.3.5. GPT Similarity (Similitud GPT).....	29
2.3.6. F1 Score	30
2.3.7. Relación con los Archivos de Preguntas Creados	30

Unidad de Planificación Rural Agropecuaria (UPRA)

Calle 28 N° 13-22, Torre C, piso 3. Edif. Palma Real. Bogotá, Colombia.

+57(601) 552 9820, 245 7307

www.upra.gov.co



2.3.8. Justificación de las Métricas.	31
2.4. Análisis de los resultados de la evaluación del modelo GPT-4o YoCampo.	31
2.5. Reevaluación y Ajuste Continuo	40
2.5.1. Coherencia (Coherence)	40
2.5.2. Fluidez (Fluency)	41
2.5.3. Fundamentación (Groundedness)	42
2.5.4. Relevancia (Relevance)	43
2.5.5. Similitud (Similarity)	43
2.5.6. F1 Score	44
3. Conclusiones	46
4. Referencias	48



Índice de tablas

Tabla 1 Ejemplo pregunta de evaluación	19
Tabla 2 Ejemplo Pregunta Evaluación con latencia y longitud.	20
Tabla 3 Archivo de prueba qa_with_context_YoCAMPO.jsonl	21



Índice de figuras

Figura 1 Tipo de evaluación modelo RGA -YoCampo.	16
Figura 2 Carga y configuración de preguntas de evaluación al modelo.	26
Figura 3 Selección de las métricas a evaluar.....	27
Figura 4 Resultado métricas evaluación	31



Lista de siglas y abreviaturas

Aunap	Autoridad Nacional de Acuicultura y Pesca
UPRA	Unidad de Planificación Rural Agropecuaria

Unidad de Planificación Rural Agropecuaria (UPRA)

Calle 28 N° 13-22, Torre C, piso 3. Edif. Palma Real. Bogotá, Colombia.
+57(601) 552 9820, 245 7307

www.upra.gov.co



Glosario

Azure Cognitive Search: Un servicio de búsqueda en la nube con capacidades de inteligencia artificial, utilizado para realizar búsquedas avanzadas y recuperación de información.

GPT-4o: Modelo avanzado de procesamiento de lenguaje natural utilizado en YoCampo para generar respuestas automáticas basadas en preguntas agrícolas.

Fluidez: Capacidad del modelo para generar respuestas gramaticalmente correctas y naturales.

Coherencia: Estructura lógica y claridad en la respuesta generada por el modelo.

Fundamentación: Grado en que las respuestas están basadas en el contexto proporcionado.

Relevancia: Relación directa entre la respuesta generada y la pregunta formulada.

Similitud: Correspondencia entre la respuesta generada y la respuesta correcta esperada (ground truth).

F1 Score: Métrica que mide el equilibrio entre la precisión y la completitud de las respuestas.

Precision: Porcentaje de elementos correctos en la respuesta generada en comparación con los esperados.

Recall: Capacidad de la respuesta para incluir todos los elementos relevantes esperados.

Contexto: Información adicional proporcionada para ayudar al modelo a generar respuestas más precisas.

Unidad de Planificación Rural Agropecuaria (UPRA)

Calle 28 N° 13-22, Torre C, piso 3. Edif. Palma Real. Bogotá, Colombia.

+57(601) 552 9820, 245 7307

www.upra.gov.co



Ground Truth: Respuesta correcta utilizada como referencia para evaluar el rendimiento del modelo.

Validación Automática: Proceso de revisión automática para asegurar la precisión de las respuestas generadas.

Temperatura: Parámetro que controla la creatividad del modelo en la generación de texto.

Top-p: Parámetro que regula la probabilidad acumulada para seleccionar las respuestas más precisas.

Datos de Entrenamiento: Conjunto de ejemplos utilizados para enseñar al modelo a generar respuestas adecuadas.

Proceso Iterativo: Proceso continuo de mejora y ajuste del modelo basado en evaluaciones periódicas.

Optimización Continua: Estrategia de ajustes regulares para maximizar el rendimiento del modelo.



Introducción

El desarrollo de modelos de lenguaje de gran escala (LLM) [2] ha revolucionado la capacidad de generar respuestas automáticas basadas en consultas específicas. En este contexto, el modelo GPT-4o [3] fue implementado en la plataforma YoCampo para abordar preguntas relacionadas con la agricultura, específicamente enfocadas en el cultivo de cacao. La implementación de este modelo tiene como objetivo mejorar la interacción entre los usuarios, tales como agricultores, investigadores y extensionistas, y proporcionarles respuestas precisas y contextualizadas sobre prácticas agrícolas, enfermedades fitosanitarias y técnicas de cultivo.

Dado el impacto que estos modelos pueden tener en la toma de decisiones en el sector agrícola, es crucial evaluar su rendimiento utilizando métricas estandarizadas. En este estudio, se analizó el desempeño del modelo YoCampo a través de métricas como fluidez, coherencia, fundamentación, relevancia, similitud y F1 score [4].

Objetivo general

El objetivo principal de este trabajo es evaluar el rendimiento del modelo GPT-4o en YoCampo, identificando fortalezas y áreas de mejora mediante un análisis detallado de las respuestas generadas. Además, se busca proponer mejoras para optimizar el rendimiento del modelo

Objetivos

- Evaluar la calidad lingüística de las respuestas generadas por el modelo GPT-4o en YoCampo, mediante el análisis de métricas como fluidez y coherencia.

Unidad de Planificación Rural Agropecuaria (UPRA)

Calle 28 N° 13-22, Torre C, piso 3. Edif. Palma Real. Bogotá, Colombia.

+57(601) 552 9820, 245 7307

www.upra.gov.co



- Analizar la fundamentación de las respuestas proporcionadas, verificando qué tan bien el modelo utiliza el contexto disponible para generar respuestas precisas y relevantes.
- Medir la relevancia de las respuestas con respecto a las preguntas formuladas, asegurando que el modelo responda de manera directa y sin incluir información irrelevante.
- Evaluar la similitud entre las respuestas generadas y las respuestas correctas (ground truth) mediante el uso de métricas como similitud y F1 score, con el objetivo de identificar discrepancias en terminología o estructura.
- Proponer una estrategia de optimización continua del modelo [1], basada en los resultados de la evaluación, que incluya la mejora del conjunto de entrenamiento, la validación automática de respuestas y la reestructuración de respuestas en función de la terminología técnica.

Alcance

Este análisis abarca la revisión de métricas clave para evaluar la calidad de las respuestas del modelo en temas agrícolas. Se exploran métodos de mejora para ajustar el vocabulario técnico, la precisión y la capacidad de fundamentar las respuestas en el contexto proporcionado.



1. Metodología de Evaluación

Para evaluar el modelo **YoCampo** utilizando el escenario "Question and Answer with Context" en Azure AI Studio, a través de los siguientes pasos específicos para llevar a cabo una evaluación enfocada en pares de preguntas y respuestas con contexto. De acuerdo con la guía de Microsoft sobre la Evaluación de Aplicaciones de IA Generativa [1], se desglosa la metodología de evaluación aplicada a YoCampo en varios pasos. A continuación, se relacionan los pasos metodológicos clave con el proceso aplicado en los archivos de preguntas y respuestas creados.

1.1. Definición del Escenario de Evaluación

Escenario Evaluado: Preguntas y respuestas con contexto (Question and Answer with Context). [5]

Relación con el archivo: El archivo de preguntas creado contiene un conjunto de preguntas y respuestas que están vinculadas a un contexto. Esto implica que el modelo debe generar respuestas basadas en información contextual proporcionada, como sucede en la cadena de suministro del cacao.

1.2. Configuración de los Datos de Prueba

- **Datos de Prueba:** En este paso, se cargan las preguntas y respuestas en un formato estandarizado, en este caso, **JSONL**. En los archivos creados, las columnas clave incluyen [6]:
 - **Question (Pregunta):** La consulta realizada por el usuario.
 - **Context (Contexto):** Información relevante que se utiliza como base para generar la respuesta.
 - **Answer (Respuesta):** La respuesta generada por el modelo.



- **Ground Truth (Verdad esperada):** La respuesta correcta o esperada.

Relacionado con los archivos creados, este paso garantiza que las preguntas estén alineadas con el contexto y que la respuesta esperada esté claramente definida. Esto permite una evaluación adecuada del rendimiento del modelo.

1.3. Selección de Métricas de Evaluación

- Las métricas utilizadas en la evaluación [7] incluyen:
 - **Groundedness:** Evalúa si las respuestas del modelo están alineadas con la fuente de información (contexto).
 - **Relevance:** Mide qué tan relacionadas están las respuestas generadas con la pregunta.
 - **Coherence:** Mide la fluidez y cohesión de las respuestas.
 - **Fluency:** Evalúa si el lenguaje de la respuesta es natural y fácil de entender.
 - **GPT Similarity:** Compara la respuesta generada con la respuesta correcta, midiendo su similitud.
 - **F1 Score:** Evalúa la precisión y el recall de la respuesta comparada con la verdad esperada.

En los archivos generados, estas métricas se utilizan para comparar las respuestas generadas por el modelo con las respuestas correctas. Por ejemplo, en la pregunta "¿Cuáles son las principales zonas productoras de cacao en Colombia?", las métricas como **Groundedness** y **Relevance** [6] mostraron variaciones, lo que indica que el modelo no fundamentó bien sus respuestas en algunos casos.



1.4. Análisis de Resultados y Ajustes

Resultados: Al evaluar el modelo, se observa que algunas respuestas no estaban alineadas con el contexto o no eran lo suficientemente precisas. Por ejemplo, en la pregunta "¿Cómo se controla la moniliasis en el cacao?", la relevancia no fue suficiente, lo que indica que el modelo omitió detalles importantes.

Mejoras: Los resultados sugieren que se deben hacer ajustes en los datos de entrenamiento, mejorando la relevancia y la precisión del modelo para generar respuestas más alineadas con la verdad esperada. Esto se puede lograr ajustando los prompts y aumentando el número de ejemplos en el conjunto de datos de entrenamiento.

1.5. Reevaluación y Ajuste Continuo

Ciclo Iterativo [2]: La evaluación se repite iterativamente para ajustar el modelo en función de los resultados obtenidos. Cada ajuste mejora la precisión de las respuestas y asegura que el modelo se adapte mejor a las consultas con contexto.



2. Evaluación

A continuación, se realiza una descripción del resultado de la metodología de evaluación con los respectivos resultados.

2.1. Preparación del Conjunto de Datos.

A partir de los datos cargados de la Biblioteca Agropecuaria Colombiana BAC con un total de 1047 archivos no estructurados (PDF, Video, Excel), con los cuales fueron entrenado del modelo, fueron generadas 10 preguntas bajo el esquema propuesto por Microsoft Azure en el formato de pregunta a evaluar.

Con lo anterior se creo la evaluación el tipo de evaluación utilizando “contexto” en la pregunta a ser resultada por el modelo GPT-4º con los datos de entrenamientos indexados. Ver Figura 1

Figura 1 Tipo de evaluación modelo RGA -YoCampo.



2.2. Datos de Prueba

Preguntas, contexto, respuestas generadas y respuestas correctas alineadas. Creación del Conjunto de preguntas para evaluar el modelo. Creando el archivo JSON a partir de la siguiente estructura: question, ground_truth, answer, context, latency, answer_length.

En el desarrollo de los Datos de Prueba, se creó un conjunto de preguntas diseñado para evaluar el rendimiento del modelo GPT-4o en el contexto del proyecto YoCampo RGA. A continuación, se detallan los criterios utilizados para la creación de las preguntas, la selección del contexto, y cómo se obtuvieron los valores de context, latency y answer_length en el archivo JSON con la estructura: question, ground_truth, answer, context, latency, answer_length.

2.2.1. Criterios para la Creación de las Preguntas

Las preguntas fueron diseñadas siguiendo criterios específicos para asegurar una evaluación integral del modelo, utilizando información aleatorio temática:

2.2.1.1. Relevancia Temática (información de prueba aleatoria)

- Cultivo de Cacao en Colombia: Preguntas sobre zonas productoras, características de suelos, beneficios de sistemas agroforestales, y prácticas agrícolas.
- Enfermedades Fitosanitarias: Enfoque en enfermedades comunes como la moniliasis y la escoba de bruja, y sus métodos de control.
- Buenas Prácticas Agrícolas (BPA): Importancia de las BPA, trazabilidad, y técnicas de manejo integrado de plagas.

2.2.1.2. Cobertura de Contenido

- Se buscó cubrir una amplia gama de temas relevantes para agricultores, extensionistas y académicos.



- Incluyó preguntas tanto generales como específicas para evaluar la capacidad del modelo en distintos niveles de complejidad.

2.2.1.3. **Diversidad en la Complejidad**

- Preguntas directas de respuesta factual.
- Preguntas que requieren interpretación y síntesis de información del contexto.

2.2.1.4. **Diversidad en la Complejidad**

- Preguntas directas de respuesta factual.
- Preguntas que requieren interpretación y síntesis de información del contexto.

2.2.2. **Selección del Contexto**

El contexto proporcionado para cada pregunta fue cuidadosamente seleccionado o creado con base en los siguientes criterios:

2.2.2.1. **Relevancia Directa**

- El contexto debía contener la información necesaria para responder la pregunta de manera precisa.
- Incluye datos y detalles específicos que el modelo puede utilizar para generar la respuesta.

2.2.2.2. **Fuentes Confiables**

- Información extraída de documentos técnicos, investigaciones académicas y fuentes reconocidas en el ámbito agrícola y del cultivo de cacao.

2.2.2.3. **Equilibrio entre Detalle y Concisión**

- Contextos suficientemente detallados para permitir una respuesta completa.



- Evitar contextos excesivamente largos que puedan diluir la información relevante.

Una muestra de lo anterior es el ejemplo de la tabla X

Tabla 1 Ejemplo pregunta de evaluación

Pregunta: "¿Cuáles son las principales zonas productoras de cacao en Colombia?"

Contexto: "En Colombia, Santander, Antioquia y Huila son conocidas por sus condiciones favorables para el cultivo de cacao, contribuyendo significativamente a la producción nacional."

En este caso, el contexto proporciona información sobre las principales regiones, pero omite intencionalmente algunas (como Tolima) para evaluar si el modelo puede inferir o recordar información adicional.

2.2.3. Valores de latency y answer_length

2.2.3.1. Latency (Latencia)

- Definición: Tiempo en milisegundos que el modelo tarda en generar la respuesta desde que recibe la pregunta y el contexto.
- Obtención de Valores: Medición real durante el proceso de inferencia del modelo.
- Utilización de herramientas de cronometraje o funciones integradas que registran el tiempo de respuesta.
- Criterio de Importancia: Evaluar la eficiencia y rapidez del modelo.
- Latencias más bajas mejoran la experiencia del usuario en aplicaciones en tiempo real.



2.2.3.2. Answer Length (Longitud de la Respuesta)

Definición: Número de caracteres o palabras en la respuesta generada por el modelo.

- Obtención de Valores: Conteo automático de caracteres o palabras tras generar la respuesta.
- Criterio de Importancia: Asegurar que las respuestas sean concisas pero informativas. Evitar respuestas demasiado breves que carezcan de información o demasiado largas que puedan ser redundantes.

Análisis de ejemplo

A continuación se describe un ejemplo en la creación de la pregunta de evaluación al modelo GTP-4^o YoCampo. Ver Tabla 2

Tabla 2 Ejemplo Pregunta Evaluación con latencia y longitud.

```
{
  "question": "¿Qué técnicas de manejo integrado de plagas se utilizan en el cultivo de cacao?",
  "ground_truth": "El manejo integrado de plagas en cacao implica el uso de control biológico, rotación de cultivos y trampas.",
  "answer": "El control biológico, las trampas y la rotación de cultivos se utilizan en el manejo integrado de plagas en el cacao.",
  "context": "El manejo integrado de plagas en cacao involucra el uso de controles biológicos, rotación de cultivos y trampas para mantener las poblaciones de plagas en niveles manejables.",
  "latency": 140,
  "answer_length": 105
}
```

Análisis de la Pregunta



Se enfoca en las técnicas específicas aplicadas en el manejo integrado de plagas en el cacao [4].

Contexto: Proporciona información detallada sobre las prácticas utilizadas, incluyendo control biológico, rotación de cultivos y trampas.

Respuesta Generada: Reitera y organiza la información del contexto de manera coherente.

Ground Truth: La respuesta esperada que contiene los elementos clave.

Latency (140 ms): Indica que el modelo respondió en un tiempo razonable para aplicaciones interactivas.

Answer Length (105 caracteres): La respuesta es de longitud adecuada, cubriendo los puntos esenciales sin extenderse innecesariamente.

Como resultado de lo anterior, fue creado el archivo llamado "qa_with_context_YoCAMPO.jsonl" el cual describe su contenido en la siguiente Tabla 3.

Tabla 3 Archivo de prueba qa_with_context_YoCAMPO.jsonl

```
{
  "question": "¿Cuáles son las principales zonas productoras de cacao en Colombia?",
  "ground_truth": "Las principales zonas productoras de cacao en Colombia incluyen Santander, Antioquia, Huila, Nariño y Tolima.",
  "answer": "Las principales regiones para la producción de cacao son Santander, Antioquia, Huila y Nariño.",
  "context": "En Colombia, Santander, Antioquia y Huila son conocidas por sus condiciones favorables para el cultivo de cacao, contribuyendo significativamente a la producción nacional.",
  "latency": 150,
  "answer_length": 85
}
```



```
}  
{  
  "question": "¿Qué características tienen los suelos del Nordeste  
Antioqueño para el cultivo de cacao?",  
  "ground_truth": "Los suelos del Nordeste Antioqueño son bien  
drenados, ácidos y ricos en materia orgánica, favoreciendo la  
producción de cacao.",  
  "answer": "Los suelos son bien drenados y ácidos, lo que apoya  
el cultivo de cacao.",  
  "context": "En el Nordeste Antioqueño, los suelos son bien  
drenados y ácidos, lo que los hace ideales para el cultivo de  
cacao, especialmente en sistemas con manejo adecuado del agua.",  
  "latency": 160,  
  "answer_length": 90  
}  
{  
  "question": "¿Cuáles son los beneficios de los sistemas  
agroforestales en el cultivo de cacao?",  
  "ground_truth": "Los sistemas agroforestales mejoran la  
biodiversidad, retienen la humedad del suelo y favorecen el  
desarrollo del cacao.",  
  "answer": "Los sistemas agroforestales mejoran la fertilidad del  
suelo y proporcionan sombra, reduciendo el estrés hídrico en las  
plantas de cacao.",  
  "context": "Los sistemas agroforestales se utilizan para mejorar  
la producción de cacao al mejorar la fertilidad del suelo y  
proporcionar sombra, lo que ayuda a reducir el estrés hídrico.",  
  "latency": 170,  
  "answer_length": 112  
}  
{  
  "question": "¿Cómo se controla la moniliasis en el cultivo de  
cacao?",
```



```
"ground_truth": "La moniliasis se controla mediante podas, fungicidas y métodos de control biológico.",
"answer": "La moniliasis se controla con podas, productos biológicos y monitoreo constante de la plantación.",
"context": "Moniliophthora roreri es un hongo que causa la moniliasis, una enfermedad importante del cacao en Colombia. Los métodos de control incluyen podas regulares, productos biológicos y monitoreo constante.",
"latency": 145,
"answer_length": 108
}
{
  "question": "¿Qué prácticas culturales se recomiendan para el manejo de Monalonion en el cacao?",
  "ground_truth": "Las prácticas recomendadas incluyen la eliminación de ramas infectadas y la aplicación de métodos de control biológico.",
  "answer": "Se recomienda eliminar las ramas infectadas y utilizar trampas para controlar la plaga.",
  "context": "Monalonion es una plaga en los cultivos de cacao, y las prácticas culturales eficaces incluyen la eliminación de ramas infectadas y el uso de trampas para controlar la población de plagas.",
  "latency": 135,
  "answer_length": 96
}
{
  "question": "¿Qué técnicas de manejo integrado de plagas se utilizan en el cultivo de cacao?",
  "ground_truth": "El manejo integrado de plagas en cacao implica el uso de control biológico, rotación de cultivos y trampas.",
  "answer": "El control biológico, las trampas y la rotación de cultivos se utilizan en el manejo integrado de plagas en el cacao.",
```



```
"context": "El manejo integrado de plagas en cacao involucra el
uso de controles biológicos, rotación de cultivos y trampas para
mantener las poblaciones de plagas en niveles manejables.",
"latency": 140,
"answer_length": 105
}
{
"question": "¿Qué es la escoba de bruja en el cacao y cómo se
previene?",
"ground_truth": "La escoba de bruja se controla mediante
fungicidas y la eliminación de partes infectadas de la planta.",
"answer": "La escoba de bruja es una enfermedad fúngica, se
maneja con podas y fungicidas.",
"context": "La escoba de bruja es una enfermedad fúngica que
afecta las plantas de cacao, generalmente se maneja con podas
regulares y la aplicación de fungicidas en las áreas infectadas.",
"latency": 130,
"answer_length": 80
}
{
"question": "¿Cuáles son las buenas prácticas agrícolas (BPA) en
el cultivo de cacao?",
"ground_truth": "Las buenas prácticas agrícolas incluyen la
rotación de cultivos, el manejo de plagas y la conservación del
agua.",
"answer": "Las BPA involucran la rotación de cultivos, el manejo
de plagas y el uso eficiente del agua.",
"context": "Las Buenas Prácticas Agrícolas (BPA) en cacao
incluyen la rotación de cultivos, el manejo adecuado de plagas y
técnicas de conservación del agua para sostener la productividad
y mejorar la calidad.",
"latency": 155,
"answer_length": 88
}
```




```
{
  "question": "¿Por qué es importante la trazabilidad en la cadena de suministro del cacao?",
  "ground_truth": "La trazabilidad garantiza la calidad y verifica el origen del cacao, facilitando el comercio internacional.",
  "answer": "La trazabilidad garantiza la calidad y la verificación del origen, importantes para los mercados de exportación.",
  "context": "Los sistemas de trazabilidad son cruciales en la cadena de suministro del cacao para garantizar la calidad del producto y verificar los orígenes de los granos de cacao, especialmente para los mercados internacionales.",
  "latency": 165,
  "answer_length": 91
}
{
  "question": "¿Cómo afecta la sombra de los árboles a la productividad del cacao?",
  "ground_truth": "La sombra de los árboles ayuda a regular la temperatura y reduce el estrés hídrico, mejorando la productividad del cacao.",
  "answer": "La sombra reduce el estrés hídrico y regula la temperatura, lo que aumenta la productividad del cacao.",
  "context": "La productividad del cacao se beneficia de la sombra proporcionada por los árboles, lo que ayuda a regular la temperatura y reducir la necesidad de riego.",
  "latency": 150,
  "answer_length": 97
}
```

Fuente: Propia (2024).

Una vez se obtiene los archivos de preguntas, son cargados a la plataforma de evaluación del modelo como se ilustra en la Figura 2.



Figura 2 Carga y configuración de preguntas de evaluación al modelo.

al.azure.com/build/createEvaluation?wsid=/subscriptions/775882f-7eef-4c2d-be51-f7db61a359d4/resourcegroups/rg-ai-lab-snuira/providers/Microsoft.MachineLearningServices/workspaces/yocampo00d-44e45df-aa2...

AI hubs / yocampo00d / All projects / YoCampo / Evaluation

Create a new evaluation

- Basic information
- Configure test data**
- Select metrics
- Review and finish

Configure test data

Select the data you want to evaluate *

Use existing dataset
Choose from your uploaded dataset collection

Add your dataset
Upload a file

Download data template for selected scenario

Drag and drop CSV or JSON file here

Preview of top 3 rows from your dataset

question	ground truth	answer	context	latency	answer length
¿Cuáles son las principales zonas prod...	Las principales zonas productoras de...	Las principales regiones para la produc...	En Colombia, Santander, Antioquia y H...	150	85
¿Qué características tienen los suelos d...	Los suelos del Nordeste Antioqueño...	Los suelos con buen drenaje y ácidos...	En el Nordeste Antioqueño, los suelos...	160	90
¿Cuáles son los beneficios de los siste...	Los sistemas agroforestales mejoran la...	Los sistemas agroforestales mejoran la...	Los sistemas agroforestales se utilizan...	170	112

Dataset mapping for prompt flow

Flow	Input	Type	Dataset column
Flow-to-Campo_V012	chat_history	list	\$(data.correct)
	query	string	\$(data.question)

Back Next Cancel

Help

What data formats are supported?
We only support csv and json file formats.
How can I map my dataset to my flow?

2.3. Selección de las métricas a ser evaluadas

Para evaluar de manera efectiva el rendimiento del modelo GPT-4o en el proyecto YoCampo, es esencial seleccionar las métricas más adecuadas para medir diferentes aspectos de las respuestas generadas por el modelo. Estas métricas deben reflejar tanto la calidad de las respuestas como la alineación con el contexto, lo que permitirá realizar ajustes precisos y optimizaciones del modelo.

La Figura 3 muestra la configuración realizada para evaluar a través de Azure Visual AI en el modelo de evaluación del modelo YoCampo RAG.



Figura 3 Selección de las métricas a evaluar.

Create a new evaluation PREVIEW

Basic information

Configure test data

Select metrics

Review and finish

Select metrics

Learn more about Microsoft-curated metrics

Performance and quality metrics curated by Microsoft

☒ Groundedness

Measures how well the generative AI application's generated answers align with information from the input source.

☒ Fluency

Measure the language proficiency of a generative AI application's predicted answer.

☒ Relevance

Measures the extent to which the generative AI application's generated responses are pertinent and directly related to the given questions.

☒ GPT similarity

Measures the similarity between a source data (ground truth) sentence and the generated response by a generative AI application.

☒ Coherence

Measures how well the generative AI application can produce output that flows smoothly, reads naturally, and resembles human-like language.

☒ F1 score

Measures the ratio of the number of shared words between the generative AI application prediction and the source data (ground truth).

Connection

Provider

Deployment name/Model

YocampoAIServiceOpen...

AzureOpenAI

gpt-4o

How does your dataset map to your evaluation input? ⓘ

Name	Description	Type	Data source
context	The source that response is generated with respect to	string	<input type="text" value="\$data.context"/>
answer	The response to question generated by the model as answer	string	<input type="text" value="\$data.answer"/>
question	A query seeking specific information	string	<input type="text" value="\$data.question"/>
ground_truth	The response to question generated by user/human as the true answer	string	<input type="text" value="\$data.ground_truth"/>

Back

Next

Cancel

Help ⓘ

- > What kinds of metrics are supported?
- > Why do I need to specify a connection and deployment?
- > How can I map my dataset to the evaluation input?

De la anterior figura se describen las métricas utilizadas en la evaluación YoCAMPO que incluyen:

2.3.1. Groundedness (Fundamentación)

Esta métrica mide qué tan bien las respuestas generadas por el modelo están alineadas con el contexto proporcionado. La fundamentación es crítica cuando las respuestas dependen de información externa, como documentos, fuentes de datos o contexto adicional. Importancia para YoCampo: En el caso de YoCampo, el contexto es esencial, ya que las preguntas están basadas en información concreta sobre el cultivo de cacao, buenas prácticas agrícolas, y control de enfermedades. El modelo debe generar respuestas que estén directamente fundamentadas en este contexto.

- Ejemplo: En la pregunta "¿Cuáles son las principales zonas productoras de cacao en Colombia?", la respuesta debe estar alineada con los datos proporcionados en el contexto sobre las regiones productoras.
- Criterio de Evaluación: Si el modelo genera una respuesta que no está directamente relacionada con el contexto proporcionado, recibirá un



puntaje bajo en groundedness. Por ejemplo, si omite una región importante mencionada en el contexto.

2.3.2. Relevance (Relevancia)

Esta métrica mide qué tan bien la respuesta generada está alineada con la pregunta formulada. La relevancia asegura que el modelo no solo responda correctamente, sino que lo haga en el marco adecuado de la pregunta planteada.

- Importancia para YoCampo: En el contexto de YoCampo, las preguntas relacionadas con prácticas agrícolas o enfermedades requieren respuestas precisas que aborden directamente la consulta. La irrelevancia puede conducir a respuestas ambiguas o poco útiles para el usuario.
- Ejemplo: Si se pregunta "¿Cómo se controla la moniliasis en el cultivo de cacao?" y el modelo responde con una explicación sobre otra enfermedad o prácticas no relacionadas, el puntaje de relevancia será bajo.
- Criterio de Evaluación: Si la respuesta no aborda el tema de la pregunta o divaga en otros temas, la métrica de relevancia se verá afectada.

2.3.3. Coherence (Coherencia)

Esta métrica mide qué tan coherente es la respuesta generada por el modelo en términos de estructura y claridad. Una respuesta coherente es aquella que tiene sentido y está bien estructurada.

- Importancia para YoCampo: En el caso de YoCampo, las respuestas deben ser claras y comprensibles para un público compuesto por agricultores, científicos e investigadores. Las respuestas incoherentes pueden generar confusión y disminuir la utilidad del sistema.



- Ejemplo: En la respuesta a "¿Qué técnicas de manejo integrado de plagas se utilizan en el cultivo de cacao?", el modelo debe proporcionar una respuesta estructurada con técnicas específicas como control biológico, rotación de cultivos, etc.
- Criterio de Evaluación: Si la respuesta es gramaticalmente correcta pero carece de sentido lógico o es difícil de entender, su puntuación de coherencia será baja.

2.3.4. Fluency (Fluidez)

La fluidez evalúa la calidad lingüística de la respuesta. Esto incluye el uso adecuado del lenguaje, la gramática y la sintaxis.

- Importancia para YoCampo: Las respuestas del modelo deben ser fáciles de leer y estar bien escritas para todos los usuarios del sistema YoCampo. Una buena fluidez asegura que las respuestas sean claras y accesibles.
- Ejemplo: Si se pregunta "¿Cuáles son las buenas prácticas agrícolas (BPA) en el cultivo de cacao?" y la respuesta contiene errores gramaticales o está mal redactada, el puntaje de fluidez será bajo.
- Criterio de Evaluación: Las respuestas con errores gramaticales o con una estructura de frase pobre tendrán una baja calificación de fluidez.

2.3.5. GPT Similarity (Similitud GPT)

Esta métrica compara la respuesta generada por el modelo con la respuesta correcta esperada, midiendo la similitud en términos de contenido y precisión.

- Importancia para YoCampo: La similitud GPT es crucial para medir si las respuestas generadas por el modelo se alinean con las respuestas correctas o ground truth, especialmente en temas técnicos o específicos como las prácticas agrícolas o las enfermedades del cacao.
- Ejemplo: Si se pregunta "¿Qué características tienen los suelos del Nordeste Antioqueño para el cultivo de cacao?" y la respuesta generada omite detalles clave, el puntaje de similitud GPT será bajo.



- Criterio de Evaluación: Si la respuesta es similar en estructura, pero omite información crítica, la similitud será más baja.

2.3.6. F1 Score

El F1 score mide el equilibrio entre la precisión (qué tan correcta es la respuesta) y el recall (cuántos elementos relevantes se incluyen en la respuesta). Es una métrica importante cuando se busca exactitud en la generación de respuestas.

- Importancia para YoCampo: En el contexto de YoCampo, es fundamental que el modelo no solo genere respuestas correctas, sino que incluya toda la información relevante. Un bajo F1 score indicaría que la respuesta no es lo suficientemente completa o precisa.
- Ejemplo: Si la pregunta es "¿Qué prácticas culturales se recomiendan para el manejo de Monalonion en el cacao?" y el modelo solo menciona el control biológico, pero omite la poda, el F1 score sería bajo.
- Criterio de Evaluación: Una respuesta completa que incluya todos los elementos correctos y relevantes tendrá un puntaje alto en el F1 score.

2.3.7. Relación con los Archivos de Preguntas Creados

Los archivos de preguntas creados siguen la estructura basada en preguntas, contexto, respuestas generadas y ground truth, lo que facilita la evaluación con las métricas seleccionadas:

- Groundedness y Relevance: Evaluarán si las respuestas generadas están fundamentadas en el contexto y si son relevantes para las preguntas formuladas.
- Coherence y Fluency: Analizarán la calidad lingüística y la estructura de las respuestas, asegurando que las respuestas sean claras y bien escritas.
- GPT Similarity y F1 Score: Compararán las respuestas generadas con las respuestas correctas esperadas, midiendo la precisión y exhaustividad de las respuestas.

2.3.8. Justificación de las Métricas.

- Groundedness y Relevance son cruciales en este contexto porque el modelo de YoCampo se basa en información agrícola técnica y necesita proporcionar respuestas basadas en hechos, como datos sobre las zonas productoras de cacao o prácticas fitosanitarias.
- Fluency y Coherence aseguran que las respuestas sean comprensibles para un público variado.
- GPT Similarity y F1 Score miden el éxito del modelo en generar respuestas alineadas con los estándares esperados en términos de contenido y precisión.

2.4. Análisis de los resultados de la evaluación del modelo GPT-4o YoCampo.

A continuación, se proporciona un análisis detallado de los resultados obtenidos en la evaluación de preguntas y respuestas con contexto:

A partir del modelo de evaluación [5] con las 10 preguntas generadas como datos de prueba, se obtuvo la siguiente gráfica que corresponde a la matriz de resultados contenida en el archivo anexo: evaluation_YocampoV01_Output_Table_16-8-2024-16-40.csv.

Figura 4 Resultado métricas evaluación





A partir de las métricas proporcionadas en la evaluación, se puede hacer un análisis en profundidad de cada una de las métricas clave que miden la calidad del modelo:

- **Coherence (Coherencia):** La coherencia mide qué tan coherentes y estructuradas son las respuestas generadas por el modelo. Con un puntaje promedio de 4.13/5, el modelo genera respuestas que en su mayoría son claras y lógicas. Sin embargo, algunas respuestas obtuvieron un puntaje de 3, lo que indica que en algunos casos, la estructura de las respuestas puede mejorarse.
- **Fluency (Fluidez):** La fluidez fue perfecta, con una puntuación de 5/5 en todas las respuestas. Esto indica que el modelo genera respuestas gramaticalmente correctas y fluidas, sin errores lingüísticos.
- **Groundedness (Fundamentación):** La fundamentación mide qué tan bien las respuestas están alineadas con el contexto proporcionado. El puntaje promedio fue de 4.60/5, lo que indica que el modelo utiliza bien el contexto en la mayoría de los casos. Sin embargo, algunas respuestas carecieron de detalles importantes, lo que explica las puntuaciones más bajas en esta métrica.
- **Relevance (Relevancia):** Con una puntuación promedio de 4.75/5, la relevancia del modelo es alta, indicando que el modelo responde correctamente a las preguntas en la mayoría de los casos. No obstante, en casos donde el modelo omite información clave o da una respuesta incompleta, la relevancia se ve afectada.
- **Similarity (Similitud):** La similitud promedia un 4/5, lo que indica que las respuestas generadas por el modelo son en su mayoría similares a las respuestas esperadas, aunque con variaciones en ciertos términos o detalles. Esta métrica es la que más variabilidad presenta, lo que sugiere que hay margen de mejora en la precisión de las respuestas generadas.



Análisis Específico de Preguntas del CSV:

Pregunta 1: ¿Cuáles son las principales zonas productoras de cacao en Colombia? La respuesta generada omite la región de Tolima, lo que afectó las métricas de fundamentación (1/5), relevancia (3/5) y similitud (3/5). Aunque la coherencia y fluidez son perfectas (5/5), la falta de completitud en la respuesta afecta su precisión.

Pregunta 2: ¿Qué características tienen los suelos del Nordeste Antioqueño para el cultivo de cacao? La respuesta es bastante precisa, con buenas puntuaciones en fundamentación (5/5) y relevancia (5/5). Sin embargo, la similitud es moderada (3/5), debido a que faltó mencionar la materia orgánica como una característica clave.

Pregunta 3: ¿Cuáles son los beneficios de los sistemas agroforestales en el cultivo de cacao? La respuesta generada fue coherente y fluida, pero la similitud es baja (3/5) debido a la falta de información sobre la biodiversidad, lo que también afectó el F1 score (0.54).

Pregunta 4: ¿Cómo se controla la moniliasis en el cultivo de cacao?

Respuesta generada: La moniliasis se controla con podas, productos biológicos y monitoreo constante de la plantación.

Contexto: *Moniliophthora roreri* es un hongo que causa la moniliasis, una enfermedad importante del cacao en Colombia. Los métodos de control incluyen podas regulares, productos biológicos y monitoreo constante.

Ground Truth: La moniliasis se controla mediante podas, fungicidas y métodos de control biológico.

**Métricas:**

- Coherencia: NaN
- Fluidez: NaN
- Fundamentación: 5/5
- Relevancia: NaN
- Similitud: NaN
- F1 Score: 0.54

Análisis:

- Fundamentación perfecta (5/5) indica que la respuesta utiliza bien el contexto proporcionado.
- Sin embargo, los valores faltantes en coherencia, fluidez, relevancia, y similitud sugieren un problema con la evaluación de esta respuesta en particular.
- Acción sugerida: Completar la evaluación de todas las métricas para obtener un análisis más completo. Además, proporcionar más detalles en el contexto puede mejorar la precisión de las respuestas.

Pregunta 5: ¿Qué prácticas culturales se recomiendan para el manejo de Monalonion en el cacao?

- Respuesta generada: Se recomienda eliminar las ramas infectadas y utilizar trampas para controlar la plaga.
- Contexto: Monalonion es una plaga en los cultivos de cacao, y las prácticas culturales eficaces incluyen la eliminación de ramas infectadas y el uso de trampas para controlar la población de plagas.



- Ground Truth: Las prácticas recomendadas incluyen la eliminación de ramas infectadas y la aplicación de métodos de control biológico.

Métricas:

- Coherencia: NaN
- Fluidez: NaN
- Fundamentación: 5/5
- Relevancia: NaN
- Similitud: NaN
- F1 Score: 0.33

Análisis:

- La fundamentación es alta (5/5), lo que indica que la respuesta fue bien fundamentada en el contexto proporcionado.
- Los valores faltantes en coherencia, fluidez, relevancia, y similitud sugieren que la respuesta no fue completamente evaluada.
- F1 score bajo (0.33) indica que la respuesta podría haber sido incompleta en cuanto a la cantidad de detalles proporcionados.
- Acción sugerida: Evaluar completamente las métricas y agregar más contexto sobre el manejo biológico para aumentar la precisión.

Pregunta 6: ¿Qué técnicas de manejo integrado de plagas se utilizan en el cultivo de cacao?



- Respuesta generada: El control biológico, las trampas y la rotación de cultivos se utilizan en el manejo integrado de plagas en el cacao.
- Contexto: El manejo integrado de plagas en cacao involucra el uso de controles biológicos, rotación de cultivos y trampas para mantener las poblaciones de plagas en niveles manejables.
- Ground Truth: El manejo integrado de plagas en cacao implica el uso de control biológico, rotación de cultivos y trampas.

Métricas:

- Coherencia: 4/5
- Fluidez: 5/5
- Fundamentación: 5/5
- Relevancia: 5/5
- Similitud: 5/5
- F1 Score: 0.77

Análisis:

- Coherencia ligeramente baja (4/5), pero las demás métricas muestran un desempeño excelente (5/5), lo que indica que la respuesta fue precisa y relevante.
- Acción sugerida: Revisar casos donde la coherencia sea baja para entender si hay problemas en la redacción o en la estructura de las respuestas.

Pregunta 7: ¿Qué es la escoba de bruja en el cacao y cómo se previene?



- Respuesta generada: La escoba de bruja es una enfermedad fúngica, se maneja con podas y fungicidas.
- Contexto: La escoba de bruja es una enfermedad fúngica que afecta las plantas de cacao, generalmente se maneja con podas regulares y la aplicación de fungicidas en las áreas infectadas.
- Ground Truth: La escoba de bruja se controla mediante fungicidas y la eliminación de partes infectadas de la planta.

Métricas:

- Coherencia: 4/5
- Fluidez: 5/5
- Fundamentación: 5/5
- Relevancia: 5/5
- Similitud: 4/5
- F1 Score: 0.45

Análisis:

- Las métricas de fundamentación, relevancia y fluidez son excelentes, pero la similitud es moderada (4/5) y el F1 score relativamente bajo (0.45), sugiriendo que faltan algunos detalles.
- Acción sugerida: Mejorar la precisión de las respuestas asegurando que se incluya toda la información relevante, como el uso de fungicidas y podas.

Pregunta 8: ¿Cuáles son las buenas prácticas agrícolas (BPA) en el cultivo de cacao?

- Respuesta generada: Las BPA involucran la rotación de cultivos, el manejo de plagas y el uso eficiente del agua.



- Contexto: Las Buenas Prácticas Agrícolas (BPA) en cacao incluyen la rotación de cultivos, el manejo adecuado de plagas y técnicas de conservación del agua para sostener la productividad y mejorar la calidad.
- Ground Truth: Las buenas prácticas agrícolas incluyen la rotación de cultivos, el manejo de plagas y la conservación del agua.

Métricas:

- Coherencia: 4/5
- Fluidez: 5/5
- Fundamentación: 5/5
- Relevancia: 5/5
- Similitud: 4/5
- F1 Score: 0.69

Análisis:

- Aunque las métricas son bastante buenas en general, la similitud (4/5) y el F1 score (0.69) sugieren que la respuesta podría incluir más detalles.
- Acción sugerida: Incluir más información específica sobre el manejo del agua y la conservación para mejorar la precisión.

Pregunta 9: ¿Por qué es importante la trazabilidad en la cadena de suministro del cacao?

- Respuesta generada: La trazabilidad garantiza la calidad y la verificación del origen, importantes para los mercados de exportación.



- Contexto: Los sistemas de trazabilidad son cruciales en la cadena de suministro del cacao para garantizar la calidad del producto y verificar los orígenes de los granos de cacao, especialmente para los mercados internacionales.
- Ground Truth: La trazabilidad garantiza la calidad y verifica el origen del cacao, facilitando el comercio internacional.

Métricas:

- Coherencia: 4/5
- Fluidez: 5/5
- Fundamentación: 5/5
- Relevancia: 5/5
- Similitud: 5/5
- F1 Score: 0.52

Análisis:

- La fundamentación y similitud son altas, pero el F1 score es relativamente bajo (0.52), lo que sugiere que la respuesta carece de algunos detalles adicionales.
- Acción sugerida: Incluir más detalles sobre los beneficios comerciales de la trazabilidad podría mejorar el F1 score.

Pregunta 10: ¿Cómo afecta la sombra de los árboles a la productividad del cacao?

- Respuesta generada: La sombra reduce el estrés hídrico y regula la temperatura, lo que aumenta la productividad del cacao.



- Contexto: La productividad del cacao se beneficia de la sombra proporcionada por los árboles, lo que ayuda a regular la temperatura y reducir la necesidad de riego.
- Ground Truth: La sombra de los árboles ayuda a regular la temperatura y reduce el estrés hídrico, mejorando la productividad del cacao.

Métricas:

- Coherencia: 4/5
- Fluidez: 5/5
- Fundamentación: 5/5
- Relevancia: 5/5
- Similitud: 5/5
- F1 Score: 0.52

2.5. Reevaluación y Ajuste Continuo

El análisis de los resultados de la evaluación del modelo GPT-4o en YoCampo muestra que el modelo funciona bien en la mayoría de las métricas, pero hay áreas clave que requieren ajustes para optimizar su rendimiento. A continuación, se realiza un análisis detallado de cómo mejorar las métricas obtenidas y se proponen acciones concretas para cada una.

2.5.1. Coherencia (Coherence)

Puntuación promedio: 4.13/5



Descripción: La coherencia mide qué tan bien estructuradas y lógicas son las respuestas del modelo.

Problema Identificado:

Aunque la mayoría de las respuestas son coherentes, algunas han recibido un puntaje más bajo, lo que indica que en ciertos casos, la estructura o el flujo de la información es deficiente.

Propuesta de Mejora:

Optimización de la estructuración de respuestas: Revisar ejemplos de respuestas que tengan un bajo puntaje de coherencia y ajustar el entrenamiento para que el modelo preste más atención a la construcción lógica de las respuestas.

Incorporar más ejemplos con estructura compleja: Ampliar el conjunto de datos de entrenamiento con ejemplos que incluyan preguntas que requieran respuestas estructuradas en varios pasos o con múltiples argumentos.

Resultado Esperado:

Mejor flujo en la estructura de las respuestas, con puntuaciones de coherencia más consistentes y cercanas a 5.

2.5.2. **Fluidez (Fluency)**

Puntuación promedio: 5/5

Descripción: La fluidez mide la calidad lingüística de las respuestas generadas por el modelo.

Problema Identificado:



Ninguno, la puntuación de fluidez es perfecta.

Propuesta de Mejora:

Mantener la calidad del lenguaje: No se requieren ajustes aquí. Sin embargo, es importante mantener la calidad de las respuestas asegurándose de que cualquier mejora en otras áreas no comprometa la fluidez.

Resultado Esperado:

Mantener la fluidez perfecta en futuras evaluaciones.

2.5.3. Fundamentación (Groundedness)

Puntuación promedio: 4.60/5

Descripción: La fundamentación mide qué tan bien están alineadas las respuestas con el contexto proporcionado.

Problema Identificado: Algunas respuestas carecen de detalles importantes o no utilizan adecuadamente el contexto proporcionado, lo que reduce su precisión y fundamentación.

Propuesta de Mejora: Mejorar la selección de contexto: Revisar el contexto proporcionado para asegurarse de que sea lo suficientemente rico y completo para generar respuestas precisas.

Refinamiento del modelo en preguntas de detalle: Entrenar al modelo para que preste atención a los detalles más pequeños del contexto, como nombres, ubicaciones específicas o técnicas agrícolas precisas.

Validación de contexto: Implementar un paso de validación automática del contexto para verificar si toda la información relevante está siendo utilizada.

Resultado Esperado:



Mejor alineación entre la respuesta generada y el contexto proporcionado, lo que aumentará la fundamentación a un promedio más cercano a 5.

2.5.4. Relevancia (Relevance)

Puntuación promedio: 4.75/5

Descripción: La relevancia mide si la respuesta del modelo está directamente relacionada con la pregunta planteada.

Problema Identificado: Algunas respuestas, aunque coherentes y bien estructuradas, no son completamente relevantes para la pregunta, lo que indica que la respuesta puede incluir información no directamente relacionada con la consulta.

Propuesta de Mejora: Optimización de la detección de relevancia: Ajustar el entrenamiento para que el modelo se enfoque únicamente en responder a la pregunta específica, evitando divagaciones o la inclusión de información irrelevante.

Incluir más ejemplos de preguntas ambiguas: Incluir más ejemplos de preguntas ambiguas en el entrenamiento para que el modelo aprenda a filtrar la información y ofrecer la respuesta más directa y relevante.

Resultado Esperado: Mejora en la capacidad del modelo para responder de manera concisa y enfocada en la pregunta, alcanzando una relevancia más consistente.

2.5.5. Similitud (Similarity)

Puntuación promedio: 4/5

Descripción: La similitud mide qué tan cercana es la respuesta generada por el modelo a la respuesta correcta esperada (ground truth).



Problema Identificado: Las respuestas generadas por el modelo no siempre coinciden completamente con las respuestas esperadas en términos de terminología o detalles específicos.

Propuesta de Mejora: Ajustar los datos de entrenamiento para mejorar la precisión: Incluir más ejemplos específicos con respuestas esperadas claras y bien definidas en el entrenamiento para ayudar al modelo a generar respuestas más precisas y similares al ground truth.

Refinamiento de términos técnicos y vocabulario: Enfocar el entrenamiento en la correcta terminología y vocabulario técnico relevante para temas como enfermedades fitosanitarias y buenas prácticas agrícolas.

Validación automática de términos clave: Implementar una validación automática de las respuestas generadas, comparando palabras clave en la respuesta con las que deberían estar presentes en la verdad esperada.

Resultado Esperado: Respuestas más precisas y alineadas con el ground truth, aumentando el puntaje de similitud a 4.5 o superior.

2.5.6. F1 Score

Puntuación variable (por pregunta): Entre 0.33 y 0.77

Descripción: El F1 score mide el equilibrio entre precisión y recall, es decir, si las respuestas generadas son completas y correctas.

Problema Identificado: Algunas respuestas carecen de detalles o incluyen información irrelevante, lo que afecta la precisión general y el recall.

Propuesta de Mejora: Mejorar el recall en preguntas técnicas: Entrenar al modelo para que no omita detalles importantes, especialmente en preguntas técnicas o con múltiples partes.



Implementar validación automática de precisión: Realizar una verificación automática de la completitud de las respuestas antes de ser generadas, asegurando que todos los puntos clave sean abordados.

Ajustar el enfoque de la generación de respuestas: Incluir un proceso de reescritura automática en el modelo para que, cuando se genere una respuesta, el modelo revise si hay elementos clave que podrían faltar.

Resultado Esperado: Aumento del F1 score con respuestas más completas y precisas, lo que elevará la puntuación a un rango más alto (0.7-0.8).

El modelo GPT-4o de YoCampo muestra un buen desempeño general en términos de fluidez y coherencia, pero hay áreas clave, como la fundamentación, relevancia y similitud, donde se pueden realizar ajustes específicos para mejorar el rendimiento. Las siguientes acciones se recomiendan para optimizar el modelo:

- Mejorar la precisión en la selección de contexto: Revisar los datos proporcionados y asegurarse de que el contexto sea lo suficientemente rico y detallado.
- Refinamiento del conjunto de entrenamiento: Incluir más ejemplos complejos y con detalles técnicos para mejorar la precisión de las respuestas en preguntas difíciles.
- Optimizar las respuestas técnicas: Aumentar la capacidad del modelo para abordar preguntas sobre temas técnicos y especializados.
- Validación automática: Incluir un proceso de validación automática para verificar si la respuesta contiene todos los elementos clave antes de ser enviada.



3. Conclusiones

- El modelo YoCampo GPT-4o demuestra un desempeño superior en fluidez y coherencia lingüística. Los resultados de la evaluación indicaron una alta calidad en la generación de texto, con puntuaciones máximas en fluidez (5/5) y una coherencia general de 4.13/5. Estas métricas aseguran que las respuestas sean legibles y estructuralmente sólidas, lo que es crucial para garantizar que los usuarios de diversos niveles educativos puedan comprender la información fácilmente. Este rendimiento destaca la capacidad del modelo para mantener un lenguaje técnico claro, necesario en la aplicación agrícola.
- Existen áreas de mejora en la fundamentación de respuestas en contexto agrícola. Si bien la fundamentación obtuvo una puntuación alta (4.60/5), en algunos casos, las respuestas generadas no aprovecharon completamente la información contextual proporcionada. Esto sugiere que el modelo puede beneficiarse de una mayor exposición a ejemplos más complejos y ricos en datos, especialmente en dominios específicos como la fitosanidad y las prácticas agrícolas sostenibles. Incrementar la precisión del contexto y afinar los datos de entrenamiento son claves para mejorar la precisión de las respuestas.
- La similitud con las respuestas esperadas es moderada y requiere ajustes técnicos. Aunque el modelo muestra una buena capacidad para generar respuestas relevantes, la similitud con las respuestas correctas fue moderada (4/5). La discrepancia se debe principalmente a la variabilidad en la terminología y estructura de las respuestas, lo que sugiere la necesidad de una mejor alineación con el vocabulario técnico específico de la agricultura. Mejorar la similitud requiere ajustes en los datos de entrenamiento, particularmente en las áreas de terminología especializada y estructuras de respuesta estándar.



- El F1 Score refleja una variabilidad que indica problemas de completitud y exactitud. El F1 Score [4] mostró fluctuaciones importantes, con puntuaciones entre 0.33 y 0.77, lo que implica que las respuestas del modelo tienden a ser incompletas o imprecisas en preguntas técnicas específicas. La introducción de más ejemplos de preguntas técnicas en el conjunto de datos de entrenamiento, junto con validaciones automáticas para asegurar la completitud de las respuestas, permitirá mejorar el equilibrio entre precisión y recall, y generar respuestas más robustas.
- Es necesaria una estrategia de optimización continua para maximizar el rendimiento del modelo.
- Los resultados de la evaluación sugieren que una estrategia de optimización iterativa será crucial para mejorar el desempeño del modelo. Esto incluye la ampliación del conjunto de datos de entrenamiento, el ajuste de parámetros como la temperatura y el top-p, y la implementación de validaciones automáticas de respuestas. Estas mejoras no solo incrementarán la precisión, sino que también asegurarán que el modelo responda de manera más coherente y relevante en entornos agrícolas específicos, maximizando así su utilidad en el contexto de la agricultura sostenible.



4. Referencias

- [1] Microsoft, «Evaluación de aplicaciones de IA generativas con Azure AI Studio,» [En línea]. Available: <https://learn.microsoft.com/es-es/azure/ai-studio/how-to/evaluate-generative-ai-app>.
- [2] Microsoft, «Azure Machine Learning prompt flow,» 01 mayo 2024. [En línea]. Available: <https://learn.microsoft.com/en-us/azure/machine-learning/prompt-flow/overview-what-is-prompt-flow?view=azureml-api-2>.
- [3] Microsoft Azure AI, «github - Vector-Search-AI-Assistant,» github , Marzo 2024. [En línea]. Available: <https://github.com/Azure/Vector-Search-AI-Assistant/tree/cognitive-search-vector>. [Último acceso: mayo 2024].
- [4] J. C. M. W. L. K. & T. K. (. Devlin, «BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE,» IEEE, 2019. [En línea].
- [5] B. Peng y C. Li, «Instruction Tuning with GPT-4,» *arXiv*, 2023.
- [6] P. J. M. I. & K. D. Liang, «Learning to Learn in Natural Language Processing.,» *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(1), 234-248., 2020.
- [7] Mondragon, V.M y García-Díaz, V, «Adaptive contents for interactive TV guided by machine learning based on predictive sentiment analysis of data,» *Springer*.



[8] Microsoft-architecture, «Architecting applications on Azure,» 20 06 2023.
[En línea]. Available: <https://learn.microsoft.com/en-us/azure/architecture/>.