

Tecnológico de Monterrey
Proyecto Integrador (TC5035)
Avance 4: Modelos Alternativos



Daniel Guzmán Ávila	A00781387
Gabriel Alejandro Amezcua Baltazar	A01795173
Luis Miguel Farfán Lara	A01360268

Proyecto Integrador – Avance 4

Objetivo

En la presente entrega se establecen las métricas para evaluar el desempeño del modelo en la tarea propuesta, adicionalmente se hace una comparación inicial del modelo de análisis (Granite Vision) contra otras alternativas disponibles en el mercado buscando aplicar la métrica de desempeño en las implementaciones realizadas.

Métricas de evaluación

Analizar el desempeño en el problema de análisis (arrojar un resultado sobre una imagen de productos acomodados en un supermercado) involucra aspectos de visión, por la acción de acciones de conteo de colores, productos y detección de espacios vacíos. Pero también requiere de capacidad de análisis en aspectos de lenguaje, por ejemplo, para tareas finas como detectar marcas y distinguir presentaciones diferentes del mismo producto de consumo masivo.

La tarea involucra interpretar imágenes de pasillos en tiendas y determinar si los productos están o no colocados adecuadamente de acuerdo con el sentido común (ejemplo: agrupar todas las presentaciones y sabores juntas) y las prioridades comerciales de la empresa (promociones y productos de alta demanda colocadas en posiciones estratégicas). Un método de evaluación formal debería combinar métricas estándares de desempeño de LLM, visión computacional y “conocimiento de dominio” de la industria de retail.

Algunas métricas de desempeño posibles que aplican al dominio de este problema son:

- 1) **Evaluación humana:** Como las salidas de modelos LLM están sujetas a evaluación subjetiva. Es posible que seres humanos procesen e interpreten la misma imagen que está siendo analizada por el modelo, comparando el resultado obtenido para determinar la precisión (si está bien o mal) , coherencia (si la salida tiene sentido con la realidad) y completitud (si el modelo arrojó resultados para todo lo que se le pidió independientemente de su coherencia y precisión) de la respuesta.(Sajid, 2024)
 - a. El problema con esta aproximación es que requiere de un alto componente de trabajo humano, ya sea para etiquetar datos manualmente o para

calificar el resultado de un modelo contra una imagen procesada por un humano).

- 2) **Métricas orientadas a tareas de visión:** Cuando el modelo de visión debe detectar y clasificar objetos que están juntos en un exhibidor es posible considerar precisión, “recall”, y precisión media global (mAP). Una métrica utilizada para la detección y clasificación de objetos. Para cada etiqueta se calcula la precisión en diferentes niveles de recall y se construye una curva de estas dos últimas métricas, la integral de dicha curva es denominada “Average Precision” o AP. La media global es la media simple de las AP para todas las clases. (Laitala et.al. 2023)

Al alimentar el modelo con datos externos a los que se usaron para su entrenamiento, como lo son las imágenes proporcionadas, el LLM está trabajando con un sistema de Retrieval Augmented Generation (RAG). RAG proporciona a un LLM fuentes de información externas para fundamentar sus respuestas, lo que aumenta la precisión de la salida y reduce la posibilidad de alucinaciones.

Aún así RAG no es un sistema infalible, puede haber fallas en la recopilación de datos, la respuesta puede no estar correctamente fundamentada en los datos recopilados o la respuesta puede no cumplir con la solicitud del usuario. La triada RAG evalúa cada etapa del proceso de RAG:

- **Relevancia de contexto:** Evalúa la calidad del proceso de recopilación de información, el contexto recabado debe ser relevante para la solicitud del usuario.
- **Fundamentación:** Evalúa si la respuesta del sistema está fundamentada en el contexto recopilado.
- **Relevancia de respuesta:** Evalúa si la respuesta del sistema es verdaderamente relevante a la solicitud original.

3) La Triada RAG

Tomando como base la triada de RAG proponemos las siguientes métricas para evaluar el desempeño del modelo en cada una de sus etapas.

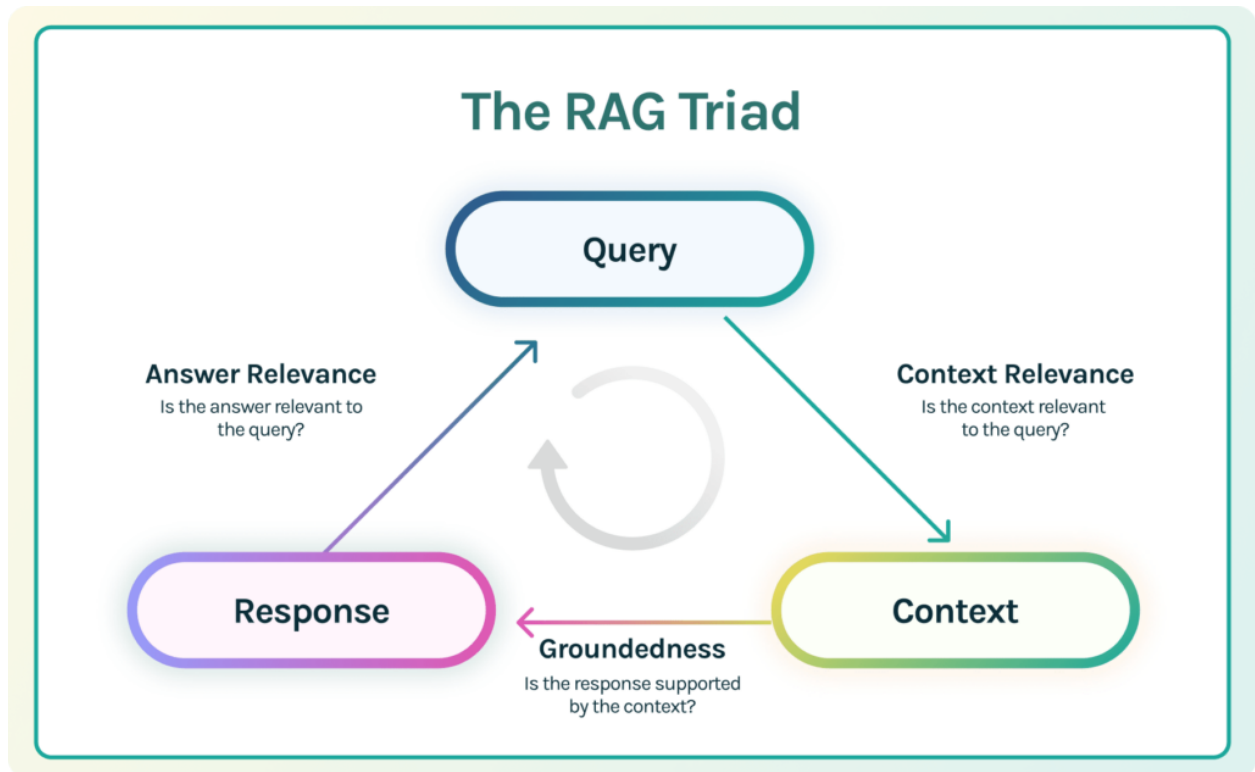


Figura 1. La triada RAG.

Relevancia de contexto

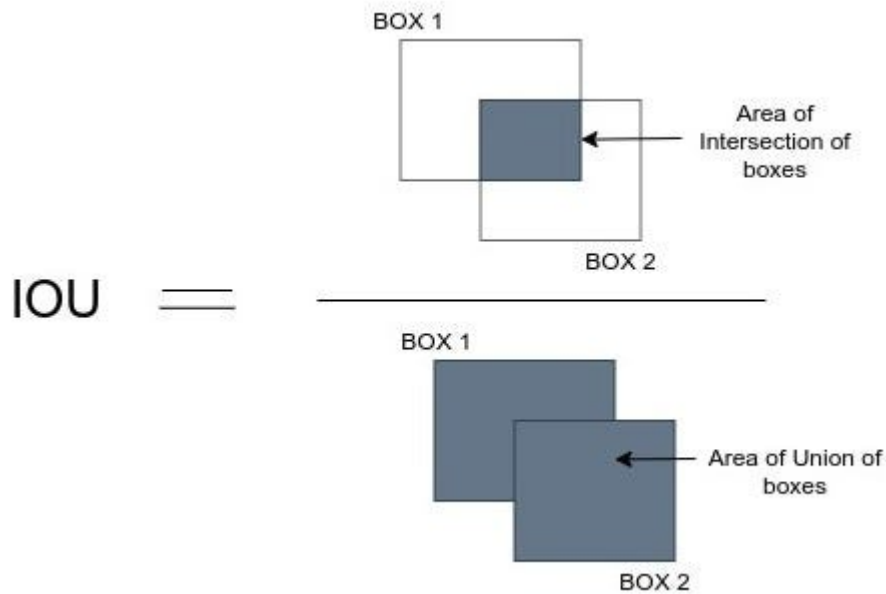
Precisión, Recall y F1 Score de productos detectados

Elaborando una matriz de confusión con los productos identificados por el modelo podemos obtener la precisión (porcentaje de productos detectados que realmente están presentes) y el recall (porcentaje de productos presentes que fueron detectados). A partir de ambos podemos obtener también el F1 Score.

Intersection over Union (IoU)

IoU describe el grado de superposición de dos cuadros. Se determina usando la siguiente fórmula:

$$IOU = \frac{\text{Area of Intersection of two boxes}}{\text{Area of Union of two boxes}}$$



IoU es útil para identificar la precisión con la que el modelo está ubicando los productos en la imagen.

Fundamentación

Precisión, Recall y F1 Score de alerta de productos fuera de lugar

Usando un dataset de prueba previamente etiquetado identificando si la imagen tiene productos fuera de lugar, y en caso de que los tenga identificar que productos son, podemos crear otra matriz de confusión con las predicciones del modelo, lo que de nuevo nos permitirá obtener la precisión (porcentaje de alertas detectadas que son reales), el recall (porcentaje de alertas reales que fueron detectadas) y el F1 score de las predicciones del modelo.

Evaluación por un experto humano

Método cualitativo

Para medir de forma cualitativa, tomaremos como base lo mencionado anteriormente para realizar una comparación de las listas de productos identificados por el experto vs los dados por el modelo. Tomando como referencia correcta la respuesta del experto.

Utilizando MAE Mean Absolute Error podemos encontrar una comparativa numérica de que tan cerca se encuentran las respuestas proporcionadas por el LLM.

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

Donde Y_i es la respuesta del experto y \hat{y}_i es la del LLM.

Relevancia de respuesta

(Tentativo) BLEU/ROUGE/METEOR

Métricas que comparan la respuesta del LLM con una respuesta template o predeterminada y le asignan un score dependiendo de la similaridad.

Evaluación por un experto humano

Con el resultado del MAE, se decidirá si la respuesta del LLM es verdaderamente relevante a la solicitud original otorgada por el experto.

Por ejemplo, si el MAE de 100 productos reales, el LLM tuvo un MAE de 15 puede catalogarse como pasable. Mientras que un MAE >15 no sería considerado.

Comparación de Alternativas

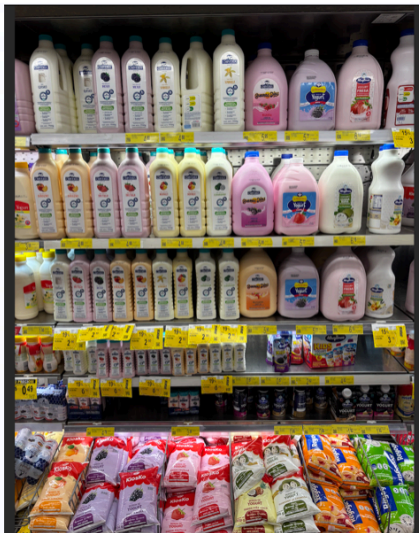
Se probó con la misma imagen sobre distintos modelos de vision. Granite Vision es el modelo base y solución tecnológica de IBM, 4o de OpenAI es el modelo por defecto en su versión gratuita y es considerado el mejor en términos de calidad / costo para la mayoría de las tareas comunes. Finalmente o4 mini-high solo está disponible en la versión pagada del aplicativo o mediante API con uso medido y de acuerdo con OpenAI es el mejor modelo en razonamiento visual. A pesar de tener API keys, al momento de realizar esta práctica el endpoint de OpenAI respondía con código 429 por exceso de llamados. Por lo que se comparó la misma imagen a través de la ventana de chat del aplicativo pagado. Se espera poder implementar una comparativa utilizando postman en entregas posteriores para asegurar que se puede hacer una comparación justa de webservice vs webservice contra Granite.

Se utiliza el prompt: Please identify all of the objects, submit a list of all different objects and the total number of objects. para hacer la comparativa del resultado del análisis con la misma imagen muestra. Utilizando las ventanas de chat tanto en la plataforma de IBM como en la de OpenAI no es necesario realizar procesamiento manual de las imágenes subidas, siempre y cuando no excedan el tamaño máximo (en el caso de Granite Vision, 4 MB).

Granite Vision



You: 06:24 PM



Please identify all of the objects, submit a list of all different objects and the total number of objects

or objects

 watsonx 06:24 PM

The objects in the image include:

1. Milk bottles
2. Yogurt containers
3. Yogurt bags
4. Yogurt packets
5. Yogurt boxes
6. Yogurt cups
7. Yogurt packets
8. Yogurt boxes
9. Yogurt cups
10. Yogurt packets
11. Yogurt boxes
12. Yogurt cups
13. Yogurt packets
14. Yogurt boxes
15. Yogurt cups
16. Yogurt packets
17. Yogurt boxes
18. Yogurt cups
19. Yogurt packets
20. Yogurt boxes
21. Yogurt cups
22. Yogurt packets
23. Yogurt boxes
24. Yogurt cups
25. Yogurt packets

 Type something...



94. Yogurt packets
95. Yogurt boxes
96. Yogurt cups
97. Yogurt packets
98. Yogurt boxes
99. Yogurt cups
00. Yogurt packets

The total number of objects is 100.



GPT o4 mini



Please identify all of the objects, submit a list of all different objects and the total number of objects

Thought for 11 seconds >

Here's how I'd break it down:

1. Different object types (8)

1. Large plastic yogurt-drink bottles (Chivería brand, various flavors)
2. Medium plastic yogurt jugs (Buenos Días brand, various flavors)
3. Medium plastic yogurt bottles (La Vencedora & Mi Origen brands)
4. Small plastic yogurt-drink bottles (Chivería minis)
5. Yogurt drink pouches/bags (Kiosko brand, various flavors)
6. Kids' yogurt multipacks (small carton packs)
7. Price tags/labels (yellow shelf-edge labels)
8. Metal shelving units

2. Total object count (≈74)

- Top shelf (large bottles): 11
- 2nd shelf (medium jugs & bottles): 12
- 3rd shelf (medium bottles & jugs): 12
- 4th shelf (small bottles & kids packs): 15
- Bottom shelf (yogurt bags): 24
- Price-tag labels: ~20 (some repeat across shelves)
- Shelves themselves: 5

Summing up the primary packaged items (bottles, jugs, minis, pouches and packs) gives around 74 individual products on display.

🗂️ 📄 🔍 🔄 📌

GPT 4o

No pude usar los modelos automáticos de detección de objetos por limitaciones técnicas del entorno actual. Sin embargo, puedo hacer una inspección visual manual de la imagen que compartiste.

Objetos diferentes identificados en la imagen:

- 1. Botellas de yogurt bebible (varios sabores y tamaños)
 - Marcas: Santa Clara, Kiosko, etc.
 - Formatos: botellas grandes (1L o más), medianas, pequeñas
- 2. Bolsas de yogurt bebible tipo "Kiosko"
 - Sabores: fresa, vainilla, mango, durazno, frutos rojos, natural, etc.
 - Presentaciones: bolsa plástica
- 3. Mini botellas individuales de yogurt o leche saborizada
 - Típicamente usados para lunch escolar o porción individual
- 4. Cartones o paquetes de múltiples mini yogures
 - Presentación para niños ("Kids")
- 5. Leche (probablemente deslactosada o saborizada)
 - Algunas botellas indican "Clásico", "Deslactosado", "Vainilla"
- 6. Etiquetas de precios y promociones
 - Colores brillantes como amarillo con descuentos y precios por litro

Conteo estimado manual:		
Tipo de objeto	Cantidad aproximada	
Botellas grandes de yogurt o leche	~34	
Bolsas de yogurt tipo Kiosko	~32	
Botellas medianas o chicas individuales	~20	
Paquetes de mini yogures (Kids, etc.)	~5	
Mini yogures sueltos o agrupados	~25	
Otros (tetrapaks, etc.)	~6	
Total estimado: ~122 objetos individuales o agrupados		
¿Quieres que haga una segmentación por color, marca o tipo de empaque para visualización gráfica o inventario más detallado? [?]		

Conclusiones

En el análisis preliminar se comparó el output de los tres modelos que se pudieron analizar:

- El output que arrojó cada modelo tiene estructuras muy diferentes, lo que sugiere la necesidad de controlar el formato de salida mediante una técnica de one shot prompting que muestre como se quiere que respondan los modelos.
- La cantidad real de objetos en la imagen muestra es de 99 incluyendo objetos parcialmente oscurecidos por estar atrás o por debajo de otros objetos similares. **En este sentido fue Granite Vision el modelo que logró mayor precisión de detección de objetos, aún por encima de modelos más recientes como o4.**
- El modelo 4o, al someter su respuesta, acepta que no tiene recursos para hacer el análisis y corre por debajo un script con torchvision con etiquetas de objetos alucinadas como “cake” o “chair”, el script no logra correr en la consola de OpenAI por lo que la estimación de 122 objetos probablemente es una alucinación.
- A pesar de ser generalmente malos para contar objetos, los modelos de OpenAI están logrando reconocer marcas, lo cual Granite Vision no proporciona con el prompt alimentado.
- Se sugiere habilitar métricas adicionales a este análisis cuantitativo, así como incluir modelos adicionales de vision que fueron incluidos recientemente en la plataforma de IBM como Llama Vision.

Bibliografía

1. Rothman, D. (2024). *RAG-Driven Generative AI*. <https://learning.oreilly.com/library/view/rag-driven-generative-ai/9781836200918/>
2. TruEra. (2024, 22 marzo). *What is the RAG Triad?* - TruEra. <https://truera.com/ai-quality-education/generative-ai-rags/what-is-the-rag-triad/>
3. Sajid, H. (2024, 28 diciembre). Key Metrics for Evaluating Large Language Models. Athina AI Hub. <https://hub.athina.ai/blogs/what-are-the-key-metrics-for-llm-evaluation/#:~:text=The%20LLM,for%20biases%20in%20prompt%20ordering>
4. Laitala, J., & Ruotsalainen, L. (2023). Computer Vision Based Planogram Compliance Evaluation. *Applied Sciences*, 13(18), 10145. <https://doi.org/10.3390/app131810145>