# Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?

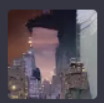
Min et al.

# Introducción

¿Cuál es el sentimiento de estas oraciones?

- ¡Qué buena comida! $\rightarrow$ Positivo
- La verdad que fue un asco. $\rightarrow$ Negativo
- El plato dejó mucho que desear. $\rightarrow$ ?

# Introducción



¿Cuál es el sentimiento de la siguiente oración?

Si por ejemplo:

- ¡Qué buena comida! -> Positivo
- La verdad que fue un asco. -> Negativo

Entonces

- El plato dejó mucho que desear. -> ?

Negativo.

# Introducción

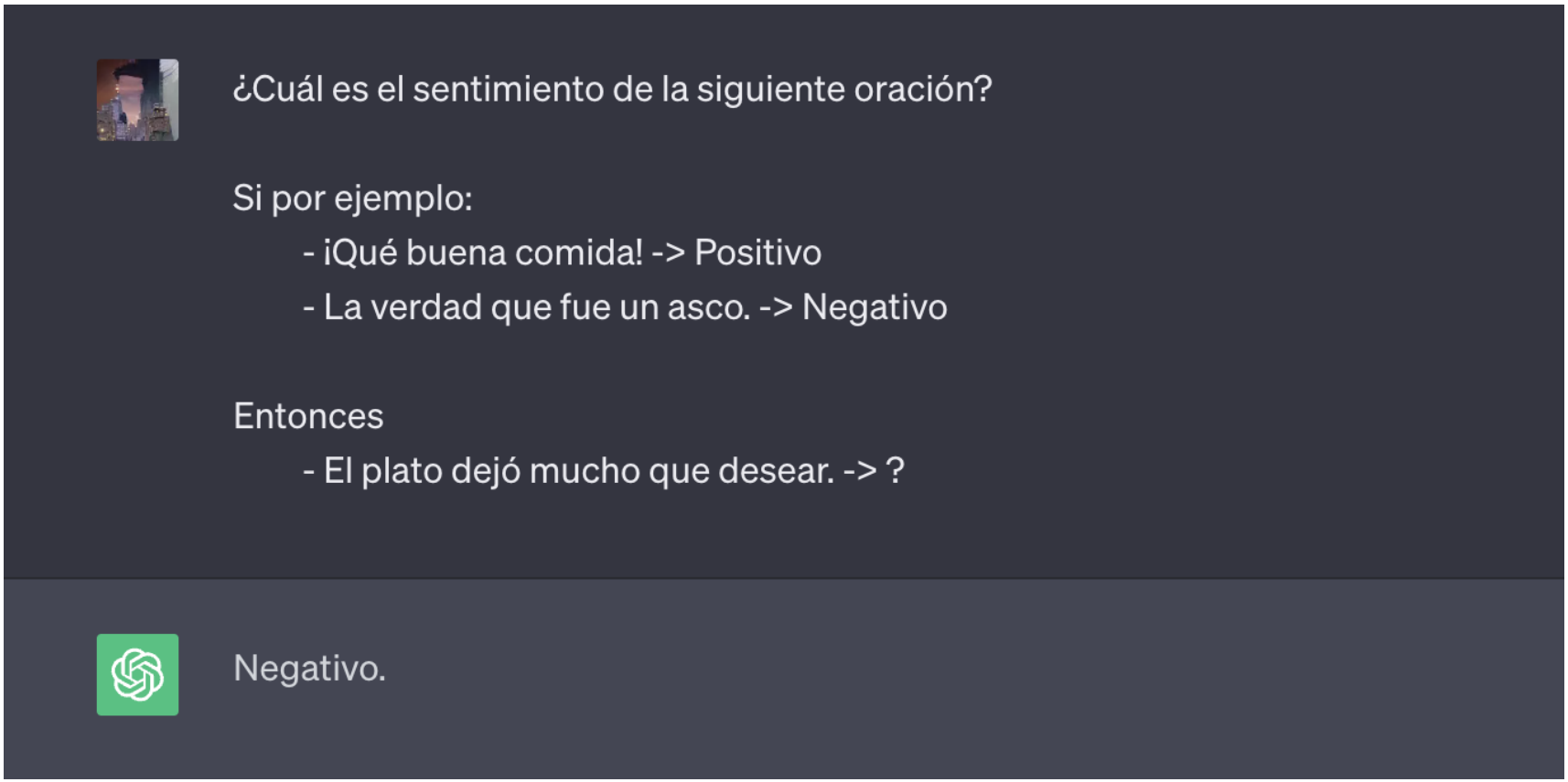
Se estudió:

- ¿Cómo elegir los ejemplos? (Liu et al.)
- ¿Cómo frasear mejor el problema? (Zhao et al.)

# Introducción

- ¿Qué aprende con estos ejemplos?

# Introducción



# Introducción

Identificaron factores importantes de los ejemplos mostrados in-context que afectan el rendimiento.

Aislaron cada factor, para ver su efecto de manera independiente.

# Introducción

Los resultados no son intuitivos.

# Introducción

Usaron 6 modelos de lenguaje basados en transformers.

Evaluaron en 26 datasets de clasificación y multiple-choice.

# Gold Labels vs. Random Labels

El primer experimento fue testear qué tanto sirve el mapeo correcto de los ejemplos a sus labels.

# Gold Labels vs. Random Labels

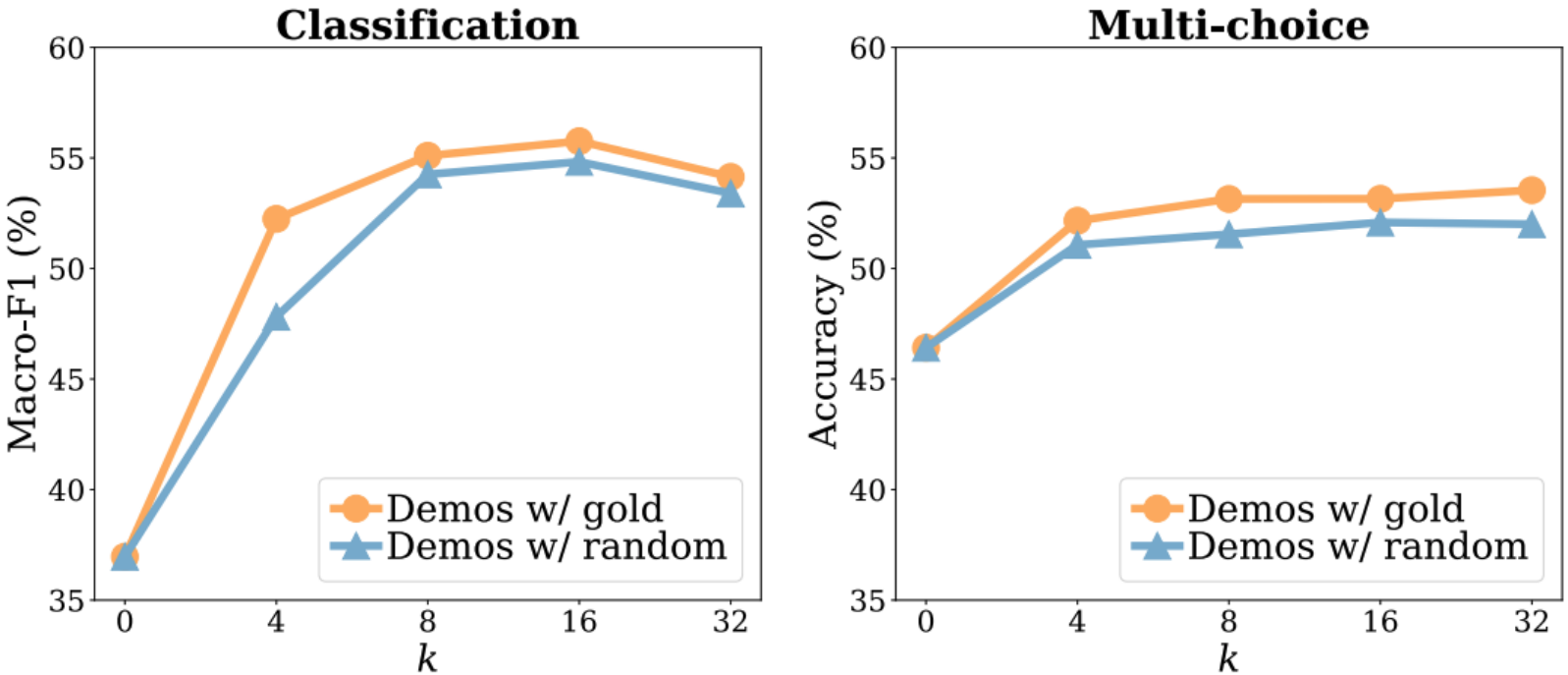
Gold

- Siamese $\rightarrow$ Cat
- Persian $\rightarrow$ Cat
- Boxer $\rightarrow$ Dog

Random

- Siamese $\rightarrow$ Dog
- Persian $\rightarrow$ Cat
- Boxer $\rightarrow$ Cat

# Gold Labels vs. Random Labels

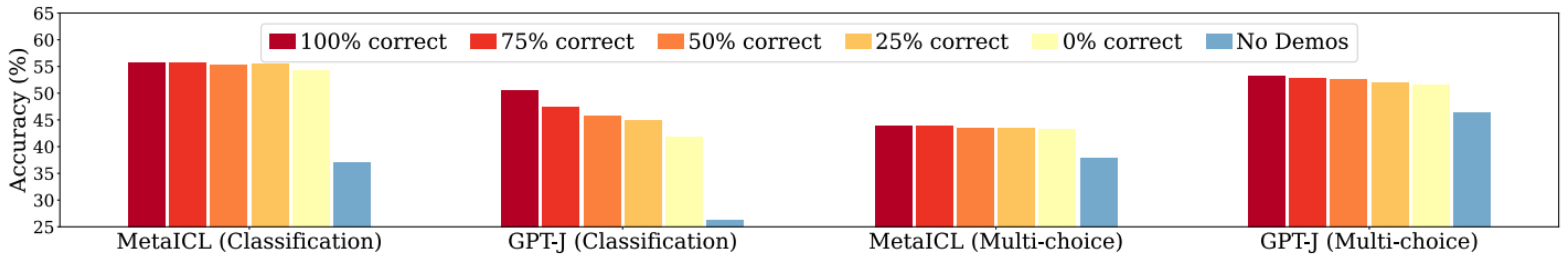


# Gold Labels vs. Random Labels

Replacing gold labels with random labels only marginally hurts performance.

Nonetheless, the models do achieve non-trivial performance on the downstream tasks.

# Gold Labels vs. Random Labels



# Gold Labels vs. Random Labels

¿Qué está pasando?

# Gold Labels vs. Random Labels

Encontraron 3 factores que afectan el rendimiento.

# Distribution of the input text

El modelo aprende qué tipos de texto pueden aparecer en el input.

# Distribución del texto del input

- Siamese $\rightarrow$ Cat
- Persian $\rightarrow$ Cat
- Boxer $\rightarrow$ Dog

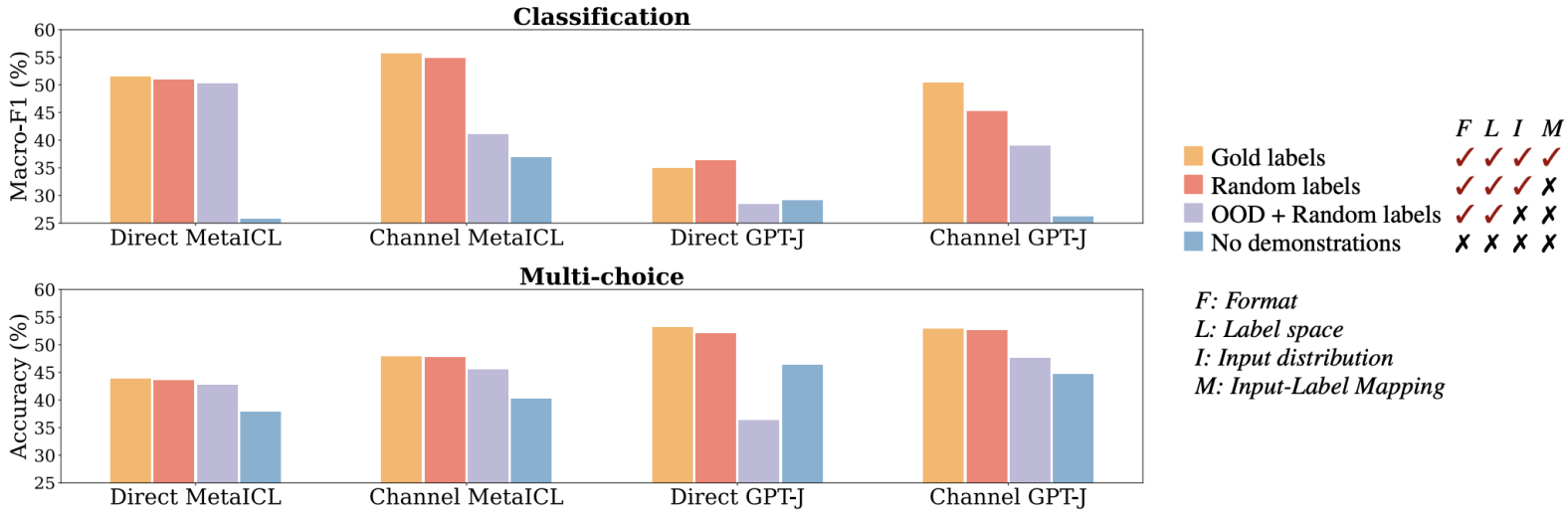
El modelo entiende que van a aparecer razas de animales como input.

# Distribución del texto del input

- Oración aleatoria 1 $\rightarrow$ Cat
- Oración aleatoria 2 $\rightarrow$ Cat
- Oración aleatoria 3 $\rightarrow$ Dog

Y miden su rendimiento.

# Distribución del texto del input



# Distribución del texto del input

This suggests that in-distribution inputs in the demonstrations substantially contribute to performance gains

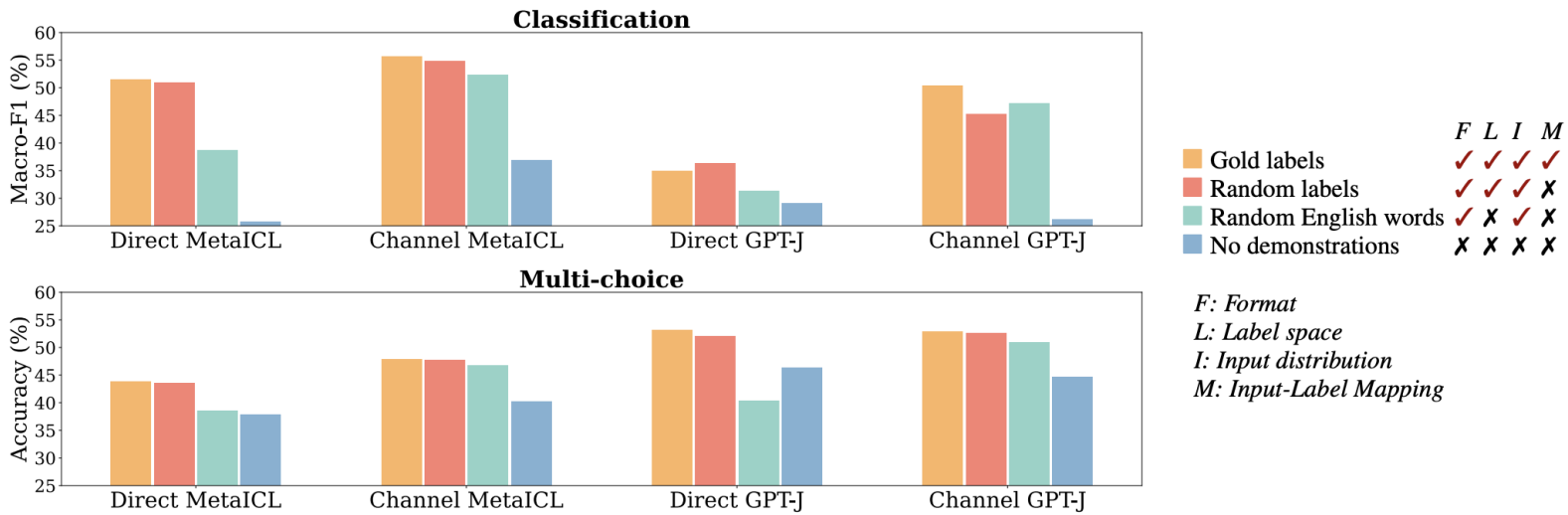
# Distribución de los labels

El modelo aprende qué tipos de labels aparecen normalmente.

# Distribución de los labels

- Siamese $\rightarrow$ Knife
- Persian $\rightarrow$ Arrow
- Boxer $\rightarrow$ Food

# Distribución del texto del input



# Distribución del texto del input

This indicates that conditioning on the label space significantly contributes to performance gains

# Formato

El modelo puede beneficiarse de saber cómo es el formato de la pregunta.

# Formato

**Only labels**
Example answers are:
Cat,Cat,Dog

**No labels**
Example inputs are:
Siamese,Persian,Boxer

# Formato

El gráfico es denso, pero...

Removing the format is close to or worse than no demonstrations.

# Meta training

Usaron MetaICL, un modelo entrenado con un objetivo para aprender bien in-context learning.

# Meta training

Patterns we see so far are significantly more evident with MetaICL.

# Discusión

¿Los modelos *aprenden* en inferencia?

# Discusión

Los autores argumentan que no. Pareciera ser que pueden ignorar las demostraciones y usar información del pre-training.

# Discusión

Creen que los ejemplos son para *localizar la tarea*, pero las habilidades fueron obtenidas en el pre-training.

# Crítica

- Gran paper.
- Los gráficos podían ser un poco densos.
- No tomaron el tamaño de un modelo como un factor que pudiese afectar sus conclusiones.

# Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?

Min et al.