

# Pruebas con LLMs

- Llama 3 70b
  - - PRUEBA ( Llama 3 70b | 66,4%)
  - - PRUEBA ( Llama 3 70b | 49% | 100 verbatims)
- Mixtral 8x7b
  - - PRUEBA ( Mixtral 8x7b | 37,67%)
  - - PRUEBA ( Mixtral 8x7b | 30,08% )
- ChatGPT-4o mini
  - - PRUEBA ( Primera prueba con GPT-4o mini | 46.91% )
  - - PRUEBA (ChatGPT-4o mini 61.54%)
  - - PRUEBA (ChatGPT-4o mini Septiembre)
  - - PRUEBA (Prueba separando según sentimientos prueba de 100 | 52.58%)
- Conclusiones

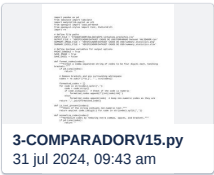
MODELO	% TP	% PARCIALES	% FP	% CLASIFICADOS
Llama 3 70b	66.40%	2.44%	25.47%	100%
Mixtral 8x7b	37.67%	12.20%	48.78%	100%
ChatGPT-4o mini	61.54%	0.55%	37.91%	100%

## Dataset general:



## Programa comparador resultados

```
1 pip install pandas tabulate matplotlib openpyxl
```



Este programa compara las columnas `CODES, PREDICTIONS` , crea un fichero CSV con el resultado de la comparación en un columna llamada `COMPARISON` , que sigue los siguientes códigos:

- 0 → Falso Positivo
- 1 → Verdadero Positivo
- 2 → Parcial Positivo
- 3 → Parcial Negativo
- 5 → No clasificado

El programa calcula la tabla con porcentajes, hay tres opciones para sacar estos resultados; Impresión por consola, imagen y fichero.xlsx. Recomendado imagen y consola (Por defecto está así). Para quitar o poner salidas modificad los booleanos.

Todas las rutas que puede usar el programa están al principio del documento para facilitar su uso.

El siguiente dataset ha servido para verificar el funcionamiento del programa de comparación (los verbatims no son representativos de los códigos):

 Dataset VALIDADOR.csv  Dataset VALIDADOR -Origen.xlsx

---

⚡ **PLANTILLA NO MODIFICAR -> COPIAR | PEGAR | RELLENAR** ⚡

### PRUEBA ( )

Nombre (multi-opción):

- ☐ Víctor
- ☐ Borja
- ☐ David
- ☐ Juan R.

Modelo LLM usado (elige uno):

- ☐ ChatGPT-4o mini
- ☐ Llama 3 8b
- ☐ Llama 3 70b
- ☐ Llama 3.1 8b
- ☐ Llama 3.1 70b
- ☐ Mixtral 8 7b
- ☐ Mixtral 8 22b

¿Se ha ejecutado el modelo en local?

- ☐ SI
- ☐ NO

Código análisis del Dataset (Primer prompt):

Resultado análisis:

Código clasifica etiquetas. (Segundo prompt):

Resultado JSON (opcional):

Resultado CSV:

Otros ficheros o comentarios (opcional):

¿Se obtienen resultados relevantes?

☐ SI

☐ NO

En caso positivo, justifica brevemente por qué:

---

## Llama 3 70b [↗](#)

### - PRUEBA ( Llama 3 70b | 66,4%) [↗](#)

Nombre (multi-opción):

☒ Borja

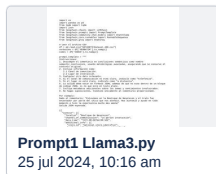
Modelo LLM usado (elige uno):

☒ Llama 3 70b

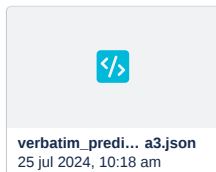
¿Se ha ejecutado el modelo en local?

☒ NO

Código análisis del Dataset (Primer prompt):



Resultado análisis:

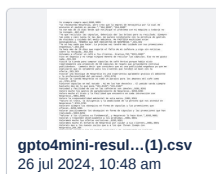


Código clasifica etiquetas. (Segundo prompt):



Resultado JSON (opcional):

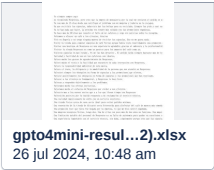
Resultado CSV:



Otros ficheros o comentarios (opcional):

Archivo Excel:

---



Métricas:

		% sobre el Total	% sobre total clasificado
Total	369		
Total Clasificad	369	100,00%	
No clasificados	0		0,00%
No clas. + FP	94		25,47%
TP+PP+PN	275		74,53%
TP (true positiv	245	66,40%	66,40%
PP (Partial posi	9	2,44%	2,44%
PN (Partial negi	21	5,69%	5,69%
FP (false positi	94	25,47%	25,47%
		100,00%	100,00%

¿Se obtienen resultados relevantes?

☒ SI

☐ NO

En caso positivo, justifica brevemente por qué:

Ya que partimos de un mejor resultado a partir de este prompt.

## - PRUEBA ( Llama 3 70b | 49% | 100 verbatims) [🔗](#)

Nombre (multi-opción):

☒ Víctor

Modelo LLM usado (elige uno):

☒ Llama 3 70b

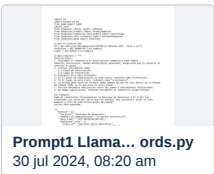
¿Se ha ejecutado el modelo en local?

☐ SI

☒ NO

Ejecución en Groq

Código análisis del Dataset (Primer prompt):



Resultado análisis:



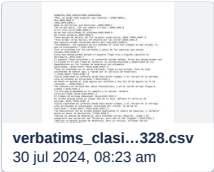
Código clasifica etiquetas. (Segundo prompt):





Resultado JSON (opcional):

Resultado CSV:



Otros ficheros o comentarios (opcional):

		% sobre el Total	% sobre total clasificado
Total	100		
Total Clasificad	100	100,00%	
No clasificados	0		0,00%
No clas. + FP	42		42,00%
TP+PP+PN	58		58,00%
TP (true positiv	49	49,00%	49,00%
PP (Partial posi	9	9,00%	9,00%
PN (Partial negi	0	0,00%	0,00%
FP (false positi	42	42,00%	42,00%
		100,00%	100,00%

¿Se obtienen resultados relevantes?

☒ SI

☐ NO

En caso positivo, justifica brevemente por qué:

El resultado no es positivo, además solo se ha probado con 100 verbatims, pero se puede apreciar que los códigos que menos acierta son los 9,95,9995 y 204 y 205

## Mixtral 8x7b [↗](#)

### - PRUEBA ( Mixtral 8x7b | 37,67%) [↗](#)

Nombre (multi-opción):

☒ Víctor

Modelo LLM usado (elige uno):

☒ Mixtral 8 7b

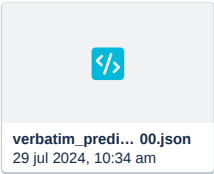
¿Se ha ejecutado el modelo en local?

☒ NO

Código análisis del Dataset (Primer prompt):



Resultado análisis:

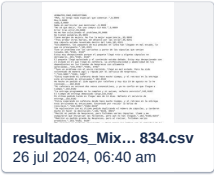


Código clasifica etiquetas. (Segundo prompt):



Resultado JSON (opcional):

Resultado CSV:



Otros ficheros o comentarios (opcional):

		% sobre el Total	% sobre total clasificado
Total	369		
Total Clasificad	369	100.00%	
No clasificados	0		0.00%
No clas. + FP	180		48.78%
TP+PP+PN	189		51.22%
TP (true positiv	139	37.67%	37.67%
PP (Partial posi	45	12.20%	12.20%
PN (Partial negi	5	1.36%	1.36%
FP (false positi	180	48.78%	48.78%
		100.00%	100.00%

¿Se obtienen resultados relevantes?

☐ SI

☒ NO

En caso positivo, justifica brevemente por qué:

- PRUEBA ( Mixtral 8x7b | 30,08% ) [↗](#)

Nombre (multi-opción):

☒ Víctor

Modelo LLM usado (elige uno):

☒ Mixtral 8 7b

¿Se ha ejecutado el modelo en local?

☒ NO

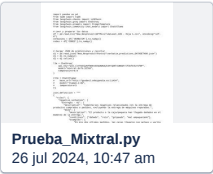
Código análisis del Dataset (Primer prompt):

Mismo que [ENLACE](#)

Resultado análisis:

Mismo que [ENLACE](#)

Código clasifica etiquetas. (Segundo prompt):



Resultado JSON (opcional):

Resultado CSV:

verbatims\_clasificados\_2607\_1135.csv

Otros ficheros o comentarios (opcional):

		% sobre el Total	% sobre total clasificado
Total	369		
Total Clasificad	369	100,00%	
No clasificados	0		0,00%
No clas. + FP	227		61,52%
TP+PP+PN	142		38,48%
TP (true positiv	111	30,08%	30,08%
PP (Partial posi	22	5,96%	5,96%
PN (Partial neg	9	2,44%	2,44%
FP (false positi	227	61,52%	61,52%
		100,00%	100,00%

¿Se obtienen resultados relevantes?

☒ SI

☐ NO

En caso positivo, justifica brevemente por qué:

Se puede ver que lanzandolo sobre mixtral ha obtenido peores resultados que en sus lanzamientos iniciales.

## ChatGPT-4o mini

- PRUEBA ( Primera prueba con GPT-4o mini | 46.91% )

Nombre (multi-opción):

☒ Víctor

☒ Juan R.

Modelo LLM usado (elige uno):

☒ ChatGPT-4o mini

¿Se ha ejecutado el modelo en local?

☒ NO

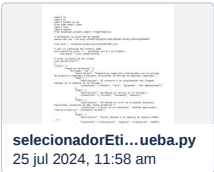
Código análisis del Dataset (Primer prompt):

Mismo que [ENLACE](#)

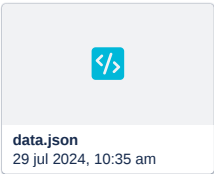
Resultado análisis:

Mismo que [ENLACE](#)

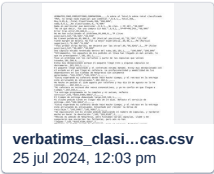
Código clasifica etiquetas. (Segundo prompt):



Resultado JSON (opcional):



Resultado CSV:



Otros ficheros o comentarios (opcional):

		% sobre el Total	% sobre total clasificado
Total	356		
Total Clasificado	356	100,00%	
No clasificados	0		0,00%
No clas. + FP	107		30,06%
TP+PP+PN	249		69,94%
TP (true positive)	167	46,91%	46,91%
PP (Partial positive)	81	22,75%	22,75%
PN (Partial negative)	1	0,28%	0,28%
FP (false positive)	107	30,06%	30,06%
		100,00%	100,00%

¿Se obtienen resultados relevantes?

☒ SI

☐ NO

En caso positivo, justifica brevemente por qué:

Partimos de esta prueba para mejorar el prompt

- PRUEBA (ChatGPT-4o mini 61.54%) ↗

Nombre (multi-opción):

☒ David



Modelo LLM usado (elige uno):

☒ ChatGPT-4o mini

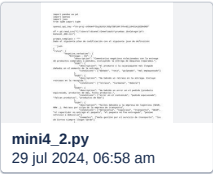
¿Se ha ejecutado el modelo en local?

☒ NO

Código análisis del Dataset (Primer prompt):

Resultado análisis:

Código clasifica etiquetas. (Segundo prompt):



Resultado JSON (opcional):

Resultado CSV:



Otros ficheros o comentarios (opcional):

		% sobre el Total	% sobre total clasificado
Total	364		
Total Clasificado	364	100,00%	
No clasificados	0		0,00%
No clas. + FP	138		37,91%
TP+PP+PN	226		62,09%
TP (true positive)	224	61,54%	61,54%
PP (Partial positive)	2	0,55%	0,55%
PN (Partial negative)	0	0,00%	0,00%
FP (false positive)	138	37,91%	37,91%
		100,00%	100,00%

¿Se obtienen resultados relevantes?

☒ SI

☐ NO

En caso positivo, justifica brevemente por qué:

Obtenemos 61.54% de acierto. Cabe mencionar que gran parte de los fallos podrían deberse a errores de codificación.

- PRUEBA (ChatGPT-4o mini Septiembre) ↗

Nombre (multi-opción):

☒ David

Modelo LLM usado (elige uno):

☒ ChatGPT-4o mini

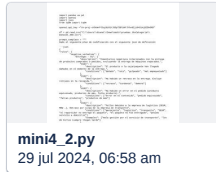
¿Se ha ejecutado el modelo en local?

☒ NO

Código análisis del Dataset (Primer prompt):

Resultado análisis:

Código clasifica etiquetas. (Segundo prompt):



Resultado JSON (opcional):

Resultado CSV:

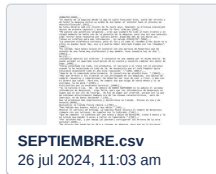
<https://docs.google.com/spreadsheets/d/1xNsUEonL36ptk4rXkbUPq0i7jSRvMilbeYEXvM7LVTw/edit?gid=1821676472#gid=182167647>

2 Conecta tu cuenta de Google

Otros ficheros o comentarios (opcional):

76% de acierto. A tener en cuenta que hay 35 comentarios, que el modelo te los predice mal, pero tras revisarlos podrían considerarse buenas predicciones.

**TENER EN CUENTA.** El dataset usado para esta prueba es el dataset de septiembre



¿Se obtienen resultados relevantes?

☒ SI

☐ NO

En caso positivo, justifica brevemente por qué:

No se ha hecho uso del .json con las units en el primer prompt. Se han puesto ejemplos de algunas categorías para dar más contexto en el segundo prompt.

---

## - PRUEBA (Prueba separando según sentimientos prueba de 100 | 52.58%) 🔗

Nombre (multi-opción):

☒ Juan R.

Modelo LLM usado (elige uno):

☒ ChatGPT-4o mini

¿Se ha ejecutado el modelo en local?

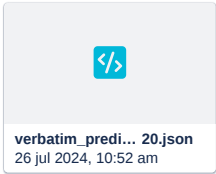
☒ NO

Código análisis del Dataset (Primer prompt)

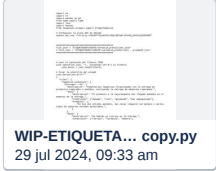




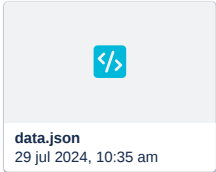
Resultado análisis:



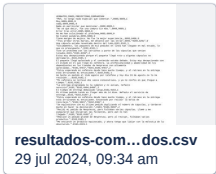
Código clasifica etiquetas. (Segundo prompt):



Resultado JSON (opcional):



Resultado CSV:



Otros ficheros o comentarios (opcional):

		% sobre el Total	% sobre total clasificado
Total	97		
Total Clasificado	97	100,00%	
No clasificados	0		0,00%
No clas. + FP	25		25,77%
TP+PP+PN	72		74,23%
TP (true positive)	51	52,58%	52,58%
PP (Partial positive)	20	20,62%	20,62%
PN (Partial negative)	1	1,03%	1,03%
FP (false positive)	25	25,77%	25,77%
		100,00%	100,00%

¿Se obtienen resultados relevantes?

☐ SI

☐ NO

En caso positivo, justifica brevemente por qué:

---

## PRUEBA ( Modificación del primer prompt y prueba con 45 verbatims )

Nombre (multi-opción):

☒ Juan R.

Modelo LLM usado (elige uno):

☒ ChatGPT-4o mini

☐

¿Se ha ejecutado el modelo en local?

☐ SI


☒ NO

Código análisis del Dataset (Primer prompt):




1-CLASIFICAD... mini.py  
29 jul 2024, 10:34 am

Resultado análisis:




verbatim\_predi... 00.json  
29 jul 2024, 10:34 am

Código clasifica etiquetas. (Segundo prompt):



2WIP-ETIQUET... copy.py  
29 jul 2024, 10:34 am

Resultado JSON (opcional):



datafase5.json  
29 jul 2024, 10:37 am

Resultado CSV:

 resultados-comparadosv2.csv

Otros ficheros o comentarios (opcional):

		% sobre el Total	% sobre total clasificado
Total	44		
Total Clasificado	43	97,73%	
No clasificados	1		2,33%
No clas. + FP	8		18,60%
TP+PP+PN	37		86,05%
TP (true positive)	31	70,45%	72,09%
PP (Partial positive)	3	6,82%	6,98%
PN (Partial negative)	3	6,82%	6,98%
FP (false positive)	7	15,91%	16,28%
		100,00%	102,33%

¿Se obtienen resultados relevantes?

- ☒ SI
- ☐ NO

En caso positivo, justifica brevemente por qué:  
La modificación del primer prompt ayuda mucho a la hora de clasificar correctamente, aun así, esto es una prueba con 45 y se procede a hacer una con dataset completo.

PRUEBA ( ChatGPT-4o mini )

Nombre (multi-opción):

- ☒ Juan R.

Modelo LLM usado (elige uno):

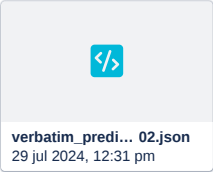
- ☒ ChatGPT-4o mini

¿Se ha ejecutado el modelo en local?

- ☐ SI
- ☒ NO

Código análisis del Dataset (Primer prompt):

Resultado análisis:



Código clasifica etiquetas. (Segundo prompt):

Resultado JSON (opcional):



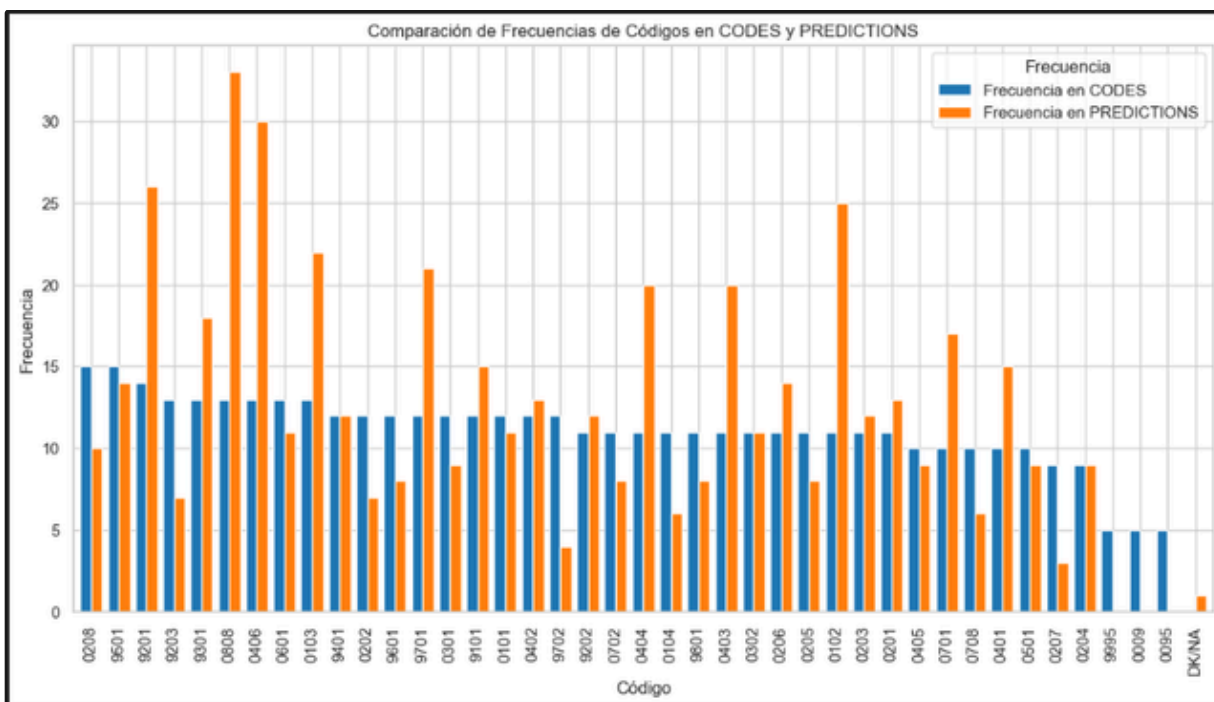
data\_0002.json  
29 jul 2024, 12:31 pm

Resultado CSV:

 resultados-comparados-todo.csv

Otros ficheros o comentarios (opcional):

		% sobre el Total	% sobre total clasificado
Total	356		
Total Clasificad	346	97,19%	
No clasificados	10		2,89%
No clas. + FP	124		35,84%
TP+PP+PN	242		69,94%
TP (true positiv	189	53,09%	54,62%
PP (Partial posi	43	12,08%	12,43%
PN (Partial neg	10	2,81%	2,89%
FP (false positiv	114	32,02%	32,95%
		100,00%	102,89%



Con el dataset corregido obtenemos !

		% sobre el Total	% sobre total clasificado
Total	356		
Total Clasificad	346	97,19%	
No clasificados	10		2,89%
No clas. + FP	120		34,68%
TP+PP+PN	246		71,10%
TP (true positiv	201	56,46%	58,09%
PP (Partial posi	35	9,83%	10,12%
PN (Partial neg;	10	2,81%	2,89%
FP (false positi	110	30,90%	31,79%
		100,00%	102,89%

¿Se obtienen resultados relevantes?

- ☐ SI
- ☐ NO

En caso positivo, justifica brevemente por qué:

## Conclusiones [↗](#)

Acciones que mejoran los resultados:

- Especificar mejor el contexto en el que se da cada subtema en la definición del plan de códigos (ejemplo especificar si el subtema se trata de vía telefónica o presencial).

Pasar un prompt o otro dependiendo del sentimiento(ChatGPT-4o-mini):

- A primera vista los resultados con 100 verbatims han sido decentes, 52,58%, sin embargo sin nos fijamos en los códigos que queríamos corregir como el 0301 que lo confunde con el 9301, hemos visto que no los clasifica correctamente.

Análisis de verbatims añadiendo details/key words(Llama3 70b):

- Agregamos descripciones más extensas del plan de códigos
- En el JSON que genera el primer prompt agregamos por un lado una prueba con details y otra con key words
  - Details: "impact": "Alta satisfacción del cliente con el producto.", "responsible\_party": "Nespresso.", "expected\_action": "Mantener y mejorar la calidad del producto."
  - Key Words: "key\_words": [ "pedido", "tardo en llegar", "14 días" ]
- Falla ambos en los 204 y 205

Mejor modelo:

El modelo que mejores resultados ha dado ha sido Llama 3 70b con 66,4% de True positives.

Posible mejor Modelo:

Hemos tenido la posibilidad de hacer una prueba con tokens muy limitados sobre LLama3.1 70b y pensamos que este puede llegar a sacar incluso mejores resultados.

Hemos cogido los 10 verbatims que fallan ambos de nuestros 2 modelos y se los hemos pasado a Llama3.1. De los 10 nos ha acertado 4.

▼ 10 verbatims que fallan ambos

Admirable cómo la campaña publicitaria ha creado un nuevo segmento de clientes. Me alegra que mi cafetera ya no haga ruido después de la reparación.

[9202, 9401]

Aprecio mucho las ofertas exclusivas para miembros del club de café. La demora en la entrega de las reparaciones es frustrante.

[403, 9203]

Buen trato y calidad del producto

[9201, 9801]

Desde el 14/12/2021, he estado tratando de que solucionen una avería, pero la respuesta ha sido extremadamente lenta.

[808]

Disfruté mucho probando la variedad limitada de café, una experiencia única que siempre busco. Los anuncios son engañosos. La empresa promete cosas que no cumple.

[203, 9702]

El programa de puntos de la marca me permite obtener fantásticos obsequios. Mi pedido se ha retrasado dos veces consecutivas, y ya no confío en que llegue a tiempo.

[102, 9203]

El repartidor de MRW entregó mi paquete tarde y no ofreció ninguna explicación o disculpa.

[104]

El servicio técnico me atiende mal

[404]

El transporte pesimo, tarde e incompleto

[104]

El unico problema es la parte ecologica El café no sale muy caliente , está es la pega que yo veo

[208, 302]

▼ Predicciones Llama3.1 70b

Admirable cómo la campaña publicitaria ha creado un nuevo segmento de clientes. Me alegra que mi cafetera ya no haga ruido después de la reparación. 9202,9401 9202, 9401

Aprecio mucho las ofertas exclusivas para miembros del club de café. La demora en la entrega de las reparaciones es frustrante.

403, 9203 9203, 0403

Buen trato y calidad del producto 9201, 9801 9801

Desde el 14/12/2021, he estado tratando de que solucionen una avería, pero la respuesta ha sido extremadamente lenta. 808 404

Disfruté mucho probando la variedad limitada de café, una experiencia única que siempre busco. Los anuncios son engañosos. La empresa promete cosas que no cumple. 203, 9702 9201, 0203

El programa de puntos de la marca me permite obtener fantásticos obsequios. Mi pedido se ha retrasado dos veces consecutivas, y ya no confío en que llegue a tiempo. 102,9203 9201, 0102

El repartidor de MRW entregó mi paquete tarde y no ofreció ninguna explicación o disculpa. 104 102

El servicio técnico me atiende mal. 404 406

El transporte pesimo, tarde e incompleto 104 104

El unico problema es la parte ecologica El café no sale muy caliente , está es la pega que yo veo. 208,302 0301, 0207, 0207