Estudio de modelos de lenguaje (LLMs) para la detección de riesgo de psicosis

Julieta Pagés, Diego Fernández Slezak

Laboratorio de Inteligencia Artificial Aplicada (LIAA), Departamento de Computación, FCEyN, Universidad de Buenos Aires.

Resumen

La identificación temprana de riesgo de psicosis representa un desafío importante para la psiquiatría computacional. Investigaciones previas han demostrado que el análisis automatizado del discurso, mediante técnicas como el Análisis Semántico Latente (LSA) y el etiquetado de partes del discurso (POS-tagging), puede predecir la transición a la psicosis con notable precisión. No obstante, estos enfoques dependen de representaciones lingüísticas tradicionales y modelos de aprendizaje automático clásicos. Este trabajo tiene como objetivo reevaluar transcripciones de entrevistas clínicas, previamente etiquetadas, utilizando modelos de lenguaje de gran escala (LLMs) contemporáneos, incluyendo Gemini 2.5 flash y pro, DeepSeek y GPT-5-mini.

Se analizaron distintos enfoques de clasificación (zero-shot y few-shot) y se evaluó su desempeño no solo mediante métricas estándar como la matriz de confusión, sino priorizando el costo esperado para mitigar el impacto del desbalance de clases inherente a la muestra. Los resultados revelan una alta sensibilidad de los modelos ante variaciones en el few-shot y permiten comparar su capacidad para capturar patrones lingüísticos sutiles frente a los métodos tradicionales reportados originalmente. Estos hallazgos subrayan el potencial de los LLMs como herramientas de apoyo diagnóstico y contribuyen a la discusión sobre la implementación de tests clínicos objetivos en el campo de la salud mental.

Palabras clave: Psiquiatría computacional, LLMs, Riesgo de psicosis, Procesamiento de Lenguaje Natural, Costo esperado.

1. Introducción

La psiquiatría computacional surge como un campo que integra modelos matemáticos y aprendizaje automático para cerrar la brecha entre los mecanismos biológicos y los fenómenos clínicos. En este contexto, el discurso del paciente es una fuente de información central, sirviendo como una "ventana"que permite inferir desde el exterior los procesos internos de la mente.

Investigaciones previas han demostrado que el procesamiento de lenguaje natural (NLP) tradicional, mediante el Análisis Semántico Latente (LSA) y el etiquetado de partes del discurso (POS-tagging), puede identificar marcadores como la reducción de la coherencia semántica y la complejidad sintáctica. Estos estudios utilizaron modelos entrenados *ad-hoc* para predecir la transición a la psicosis con un accuracy de 83 % intra-protocolo, y de 79 % inter-protocolo, con un TPR de 60 % y TNR de 82 % para el segundo caso. Sin embargo, la llegada de los modelos de lenguaje de gran escala (LLMs) plantea un cambio de paradigma: "de modelos caja blanca a modelos caja negra". Cedemos control sobre el entrenamiento ad-hoc y pasamos a un modelo genérico pero con un entrenamiento increíblemente más extenso, que luego podemos ajustar mediante *prompting*. Este trabajo busca explorar si estos nuevos modelos pueden capturar de manera consistente los patrones lingüísticos identificados en los estudios anteriores y utilizarlos para clasificar el riesgo de psicósis en los pacientes.

2. Metodología

2.1. Sujetos y Tarea de Clasificación

Se utilizaron transcripciones de entrevistas clínicas de pacientes en Riesgo Clínico de Psicosis (a partir de ahora nos referiremos a este concepto como **CHR** por Clinical High Risk). La muestra incluye 5 pacientes que evolucionaron a un brote psicótico dentro de un seguimiento de dos años (CHR+) y 30 que se mantuvieron estables (CHR-). La tarea consiste en una clasificación binaria *end-to-end* donde el LLM procesa la entrevista a partir de cierto prompt, para luego asignar la etiqueta clínica (CHR+ / CHR-).

2.2. Modelos y Estrategia Experimental

Se evaluaron cuatro arquitecturas: `gemini-2.5-flash`, `gemini-2.5-pro`, `deepseek` y `gpt-5-mini`. El diseño experimental incluyó variaciones en las instrucciones (*prompts*) y en la selección de ejemplos de referencia (*few-shot*) para medir la estabilidad y consistencia de las respuestas.

2.3. Métodos de Evaluación

Además de presentar la matriz de confusión, dada la naturaleza desbalanceada de los datos clínicos y la gravedad dispar de los errores, en lugar de analizar la precisión como suele hacerse en la mayoría de trabajos, se adoptó el **Costo Esperado (EC)**. Esta métrica permite asignar un costo específico a cada decisión según el escenario de aplicación, haciéndolo más genérico y sumándole una mayor interpretabilidad. En la detección de psicosis, podríamos considerar más costoso un Falso Negativo (FN) que un Falso Positivo (FP), si lo que nos interesa es ajustar el margen de pacientes clasificados como CHR, evitando perder pacientes con gran riesgo de brote psicótico. Por ello, se analiza una estructura de costos penalizando los FN con un peso de 5 frente a un peso de 1 para los FP alineada este objetivo. Se incluye también el análisis en caso de asignar costo 1 a ambos error, que es la medida inversa del accuracy.

3. Desarrollo Experimental Y Análisis de Resultados

3.1. Ajuste de Prompting y Sensibilidad Semántica

En esta etapa se evaluó la sensibiliadd de los modelos ante variaciones en la instrucción. Se definieron 10 variantes semánticas del prompt original, manteniendo la tarea de clasificación end-to-end pero con alteraciones del fraseo y nivel de detalle en la explicación. Los modelos evaluados fueron `gemini-2.5-flash`, `gemini-2.5-pro`, `deepseek` y `gpt-5-mini`.

Al analizar la distribución de la Tasa de Verdaderos Positivos (TPR) frente a la Tasa de Verdaderos Negativos (TNR) (ver Figura 1), se identificaron patrones de comportamiento distintivos entre los modelos:

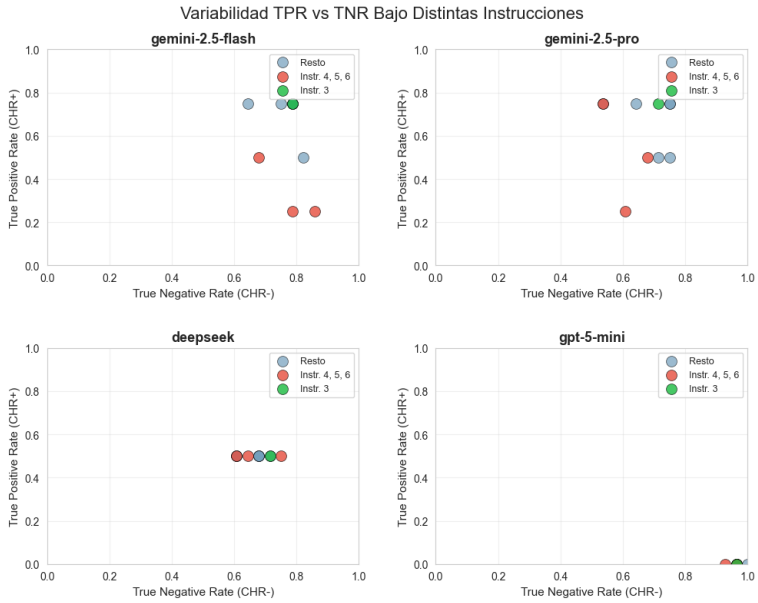
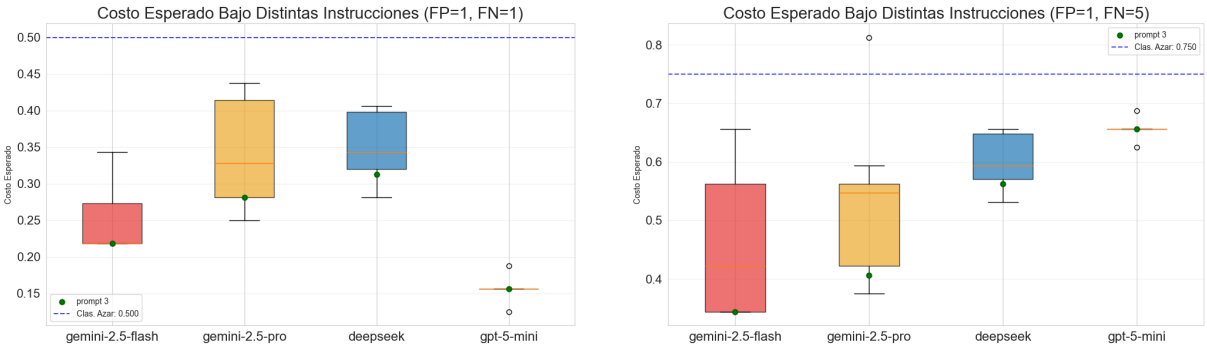


Figura 1: Tasa de Verdaderos Positivos (TPR) vs. Tasa de Verdaderos Negativos (TNR) ante 10 variaciones de *prompt*. En rojo tres variaciones generadas automáticamente a partir de otra, y en verde una variación que se muestra consistentemente entre las mejores opciones de cada modelo.

- `gpt-5-mini`: el modelo mostró un sesgo negativo extremo. Clasificó la mayoría de los casos como negativos (CHR-), exceptuando uno que siempre etiquetó erróneamente como CHR+, logrando una TNR cercana al 100 % pero una TPR nula o ínfima.
- `deepseek`: resultó ser el modelo más estable, con una dispersión mínima en sus resultados ante los cambios de instrucciones. El poco cambio presentado fue en el TNR y nada en el TPR, lo cuál puede estar influenciado por la disparidad de casos de cada tipo.
- `gemini-2.5`: tanto el fast como pro, exhibió la mayor variabilidad. Si bien logró los mejores desempeños individuales, su rendimiento cayó drásticamente en tres variantes de prompt específicas (generadas automáticamente por otro LLM), lo que evidencia que pequeñas variaciones en el .ªjuste de la tarea"vía prompting pueden alterar significativamente la frontera de decisión.

Además de las 3 instrucciones generadas automáticamente, 2 de ellas también presentan peor desempeño al evaluarse con `deepseek`, e incluso una de ellas logró que `gpt-5-mini` etiquete un paciente más como CHR+ erróneamente. Sumando a esto vemos que otra de las instrucciones, marcada en verde en los gráficos, se mantuvo como la mejor entre los modelos de gemini y `deepseek`. Lo que podría indicar que la variación del desempeño frente a la instrucción no es casual.



(a) Costos iguales (FP=1, FN=1)

(b) Penalización del FN (FP=1, FN=5)

Figura 2

Por otro lado, se calcularon los resultados del costo esperado en caso de tomar para ambos errores un paso de 1 (lo que hacemos indirectamente al calcular accuracy), y el costo esperado de pesar más el False

Negative (ver figura 2). Vemos que bajo el primer análisis, `gpt-5-mini` parecerían tener los resultados con mejor desempeño. Sin embargo, en cuanto aumentamos el peso del FN `gpt-5-mini`, por el contrario, pasa a ser el de mayor costo.

3.2. Exploración de Few-shot y Selección de Casos

En esta etapa se buscó evaluar el impacto del aprendizaje en contexto mediante la variación de los ejemplos de referencia proporcionados a los modelos (*few-shot-prompting*). Originalmente, los "shots"se seleccionaron manualmente, en base a la intuición de como afectaría a la clasificación según características que se notaban de los pacientes en las pruebas preliminares. En esta instancia, entonces sumamos diversas combinaciones de ejemplos, y también la opción de clasificación *zero-shot*.

Al observar la distribución de TPR frente a TNR (ver Figura 3), los resultados revelan una gran variabilidad ante los ejemplos elegidos. A diferencia del experimento de *instrucciones*, en esta instancia incluso `deepseek` exhibió una marcada sensibilidad y una gran variación en sus resultados. No obstante, un hallazgo relevante es que tanto en la familia `gemini` como en `deepseek`, el *few-shot* original y el caso de *zero-shot* se mantuvieron cercanas y como opciones aceptables, estando entre las mejores o al menos cercanas a la media.

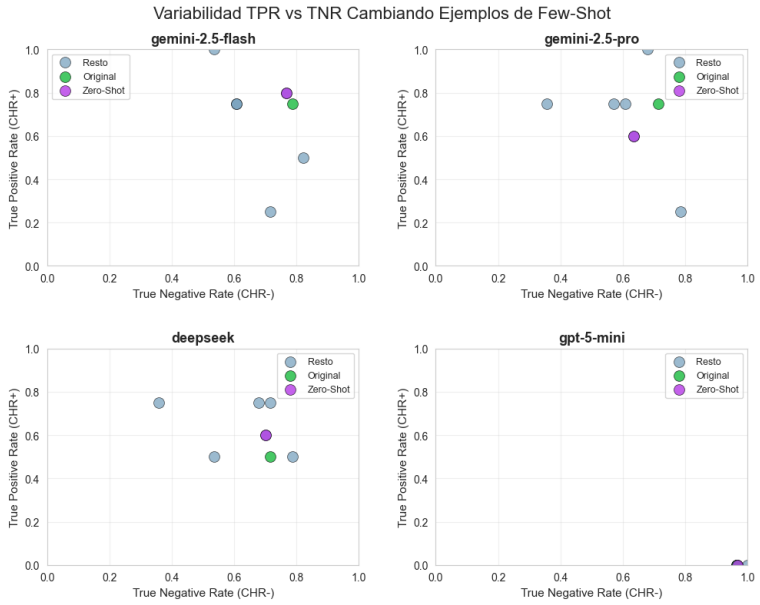


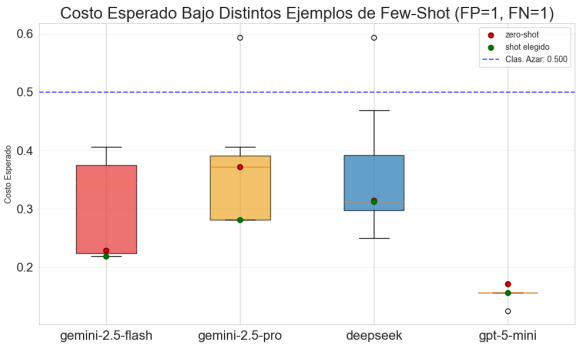
Figura 3

Al calcualr el costo esperado a estos casos, notamos que bajo una penalización balanceada ($FN = 1$), el *few-shot* original suele superar al *zero-shot*. Sin embargo, al priorizar la detección de pacientes en riesgo ($FN = 5$), se observa que el enfoque *zero-shot* pasa a ser levemente superior en desempeño para los modelos `gemini-flash` y `deepseek`.

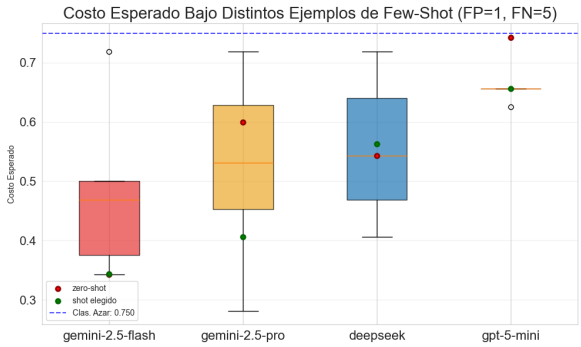
4. Discusión y Conclusiones

Antes de empezar, queremos aclarar que por el tamaño acotado de la muestra las métricas puntuales deben interpretarse con cautela, sobre todo de los casos positivos, que dependiendo de si se usó few o zero-shot son tan 4 o 5 casos, un único error en este caso representa una variación del 20-25 % en el TPR, lo que limita severamente la comparabilidad de resultados puntuales entre modelos y configuraciones.

Luego de la exploración, podemos concluir que, si bien en esta primera exploración no se logró superar los resultados conseguidos anteriormente con los métodos NLP y ML clásico desarrollados ad-hoc, los modelos de Gemini y DeepSeek fueron capaces de capturar patrones lingüísticos sutiles que les permitieron clasificar los discursos de los pacientes con relativo éxito, superando en todos los casos a un clasificador trivial, con un TP y TN de aproximadamente 75 % para gemini en los casos mejor ajustados, y de un TP de 50 % y TN de 65 % para DeepSeek. En cuanto a gpt-5-mini, es probable que su desempeño sesgado



(a) Costos iguales (FP=1, FN=1)



(b) Penalización del FN (FP=1, FN=5)

Figura 4

se deba a que el modelo fuera, en efecto, demasiado *"mini"* para poder distinguir los patrones necesarios para esta tarea.

Otro punto que queremos mencionar es que la consistencia inter-modelo de instrucciones que mantienen un alto desempeño de forma transversal en Gemini y DeepSeek, y viceversa, instrucciones que parecen empeorarlo, indicando que, en parte, la eficacia de la clasificación que dejó de estar en el entrenamiento ad-hoc pasó efectivamente a estar en la calidad semántica de la instrucción. Pero a su vez, esto plantea la duda de si termina siendo realmente un test objetivo, como se necesitaría para utilizarlo realmente como test clínico.

Lo mismo pasa con la selección del ejemplo, el hecho de que haya alta sensibilidad a los ejemplos, sumado a que una de las combinaciones de mantuviera estable y con buenos resultados entre los modelos de Gemini y DeepSeek, podría explicarse en la naturaleza dispar de los ejemplos, y que algunos puedan terminar confundiendo más que orientar en la dirección correcta. Teniendo eso en cuenta y que los resultados del zero-shot se encuentran aceptables, sería debatible que, antes que elegir ejemplos que podrían terminar sesgando (ya sea en el sentido que preferiríamos o no), resulte conveniente no pasar ninguno, y dejar que el ajuste sea directamente sobre la instrucción brindada, lo que resulta, además, más claro para practicar.

Por último, volver a destacar que estos modelos de LLMs aún no lograron superar métodos tradicionales, la señal que presentan indica la existencia de un potencial. Sin embargo, hoy el NLP y ML tradicional aún proponen soluciones más transparentes, menos costosas y, dado el contexto medicinal, con menos posibilidad de problemas de seguridad respecto a los datos sensibles.

La investigación presentada es parte de una Tesis de Licenciatura en Ciencias de la Computación en la que aún se está trabajando. Además de lo presentado en este artículo, se está trabajando en un análisis más exhaustivo del few-shot, probando con muchos más ejemplos elegidos de manera azarosa. También se realizará un análisis de los resultados sobre cada uno de los pacientes para entender cómo afecta cada paciente en particular al clasificador y si hay casos más difíciles o confusos que otros. Sobre lo anterior, se suma la prueba de utilizar a los modelos como extractores libres de posibles síntomas prodrómicos. Por último, evaluar el segundo dataset (UCLA) para determinar si las estrategias elegidas se mantienen robustas ante nuevos datos.

Referencias

[1] Corcoran, C. M., et al. (2018). Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatry*, 17(1), 67-75.

[2] Bedi, G., et al. (2015). Automated analysis of free speech predicts psychosis onset in high-risk youths. *npj Schizophrenia*, 1(1), 1-7.

[3] Montague, P. R., et al. (2012). Computational psychiatry. *Trends in Cognitive Sciences*, 16(1), 72-80.

[4] Ferrer, L. (2025). No hay necesidad de usar sustitutos ad-hoc: el costo esperado es una métrica de clasificación basada en principios y de propósito general. *Memorias de las 54 JAIIO - ASAID*.