

Por qué frenamos ECO-1

Evidencia de dos experimentos pequeños con modelos

Manuel Nicastro · Moodinary
3 de octubre de 2026

Resumen

ECO-1 exploró dos usos: extraer entidades contextuales de comentarios sobre organizaciones y clasificar señales de moderación. Para entidades usamos LACE, un proceso de propuestas, revisión ciega de desacuerdos, una ronda acotada de razones y un piso conservador de elementos marcados. Moderación fue un clasificador ajustado por separado. Ambos experimentos se entrenaron y revisaron con datos sintéticos, pero esos resultados no respondían la pregunta del producto. En una muestra pequeña de comentarios anónimos reales, la versión nueva coincidió con 6 de 28 referencias indirectas específicas a personas o grupos; la anterior, con 15 de 28. Son observaciones de cobertura, no estimaciones de exactitud: el inventario fue manual y no hubo un conjunto de prueba con etiquetas independientes. El clasificador de moderación quedó por debajo de las referencias disponibles, y su primera medición de desarrollo estaba contaminada por entrenamiento con los mismos ejemplos. La segunda versión tampoco superó las referencias en una prueba ya expuesta. El piloto pago de adaptación se frenó antes de empezar: no pudimos confirmar acceso a las bases candidatas, aceptación de licencias, un límite de gasto efectivo ni una evaluación limpia. Cancelamos el programa de construcción de modelos, en vez de convertir puntajes sintéticos en una promesa del producto. El trabajo ahora orienta dos líneas más acotadas: pronóstico del clima emocional y reducción de costos de plataforma. Este documento registra la evidencia, los errores y los límites contables. No incluye texto de participantes.

1. Qué buscábamos con ECO-1

Moodinary recoge pulsos breves y anónimos sobre cómo está la gente en una organización. La idea inicial era hacer más útiles esas respuestas al detectar entidades contextuales y señales de moderación o seguridad. El extractor de entidades usó LACE: propuestas de candidatos, revisión ciega de desacuerdos, una ronda acotada de razones y un piso conservador de elementos marcados. El clasificador de moderación era otro sistema, separado.

En principio, LACE podría aplicarse a otras tareas que necesitan proponer una respuesta, revisarla y conservarla, descartarla o marcarla. Pero la evidencia de este trabajo solo cubre extracción de entidades. No demuestra que LACE funcione en distintos tipos de tareas lingüísticas ni que la revisión cruzada mejore la calidad en general.

El 2 de octubre presentamos públicamente los dos experimentos como previews de Nubi, con los nombres [nubi-moderation-preview](#) y [nubi-entity-preview](#). El anuncio decía que seguían en evaluación y que no estaban integrados en Moodinary. Este documento es la retrospectiva posterior y registra la decisión de frenar.

También habíamos puesto el foco en una necesidad demasiado estrecha. Un extractor propio de privacidad no era el objetivo central. La experiencia útil se parece más a mostrar referencias contextuales sobre momentos, personas y roles, lugares y temas. Un detector de privacidad no cubre todo eso, y nuestros resultados no justifican la carga de evaluación y operación de desarrollar uno para la tarea más limitada de PII.

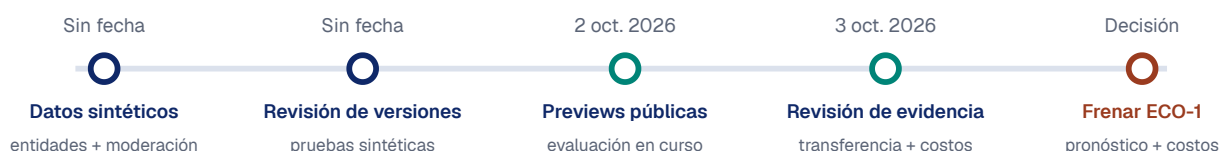


Figura 1. La secuencia es orientativa. Solo se fechan los hitos documentados; las etapas iniciales quedan sin fecha.



Moderación era un clasificador aparte; no integraba este flujo LACE.

Figura 2. LACE tal como se usó para extracción de entidades. El diagrama describe el proceso probado, no un método general ya validado.

2. Entidades: buenos puntajes sintéticos, poca transferencia

Resultados de desarrollo

La primera versión del extractor se entrenó con 3.000 ejemplos sintéticos escritos por un generador. El conjunto sintético de desarrollo tenía 200 comentarios, 300 spans etiquetados y 60 comentarios negativos difíciles. No había coincidencias exactas de texto normalizado, esqueletos de plantilla, identificadores de familia ni valores de entidad entre entrenamiento y desarrollo. Sí compartían el generador, la ontología, los contextos organizacionales en español y formatos reconocibles. Ciento cincuenta de los 200 comentarios de desarrollo compartían con entrenamiento un patrón de tipos de entidad. Las etiquetas las produjo el mismo sistema de autoría; no hubo escritura ni anotación independiente por personas.

También hubo un error de formato de salida. El evaluador congelado de spans exactos dio F1 0,000 porque los offsets de caracteres generados eran inválidos y dejaron sin coincidencia los 300 spans de referencia. Un diagnóstico posterior, solo de código, ubicó el texto predicho dentro de la fuente y dio F1 0,879 en salidas interpretables; al contar los elementos malformados o imposibles de ubicar como falsos positivos, bajó a 0,861. Una ejecución de servicio autorizada por separado, sobre el mismo conjunto sintético, obtuvo 0,904 con ese protocolo de reubicación; el sistema seleccionado de revisión cruzada dio 0,672 y la referencia evaluada, 0,509. El cero original sigue en el registro: el diagnóstico posterior no lo reemplaza. La reubicación se incorporó después del fallo inicial y no es una confirmación independiente.

La segunda versión usó otro conjunto sintético de desarrollo: 280 comentarios, 259 spans y 98 negativos difíciles. El F1 exacto por texto y tipo fue 0,903; al contar como falsos positivos cinco elementos repetidos o imposibles de ubicar, el resultado complementario fue 0,894. No hubo spans aceptados en los 98 negativos difíciles. Son controles útiles de un flujo acotado, no evidencia de exactitud en el mundo real. Como las versiones usaron conjuntos y protocolos distintos, no corresponde leer esos puntajes como una comparación controlada.

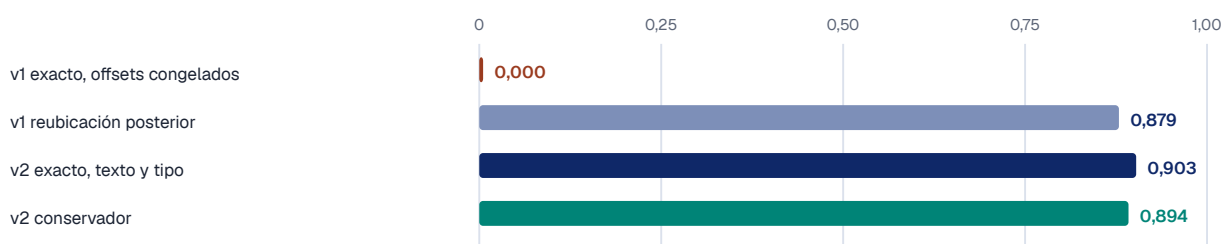


Figura 3. Solo puntajes de desarrollo sintético. Cambian las versiones, los conjuntos y los protocolos; no es una comparación controlada ni una estimación de exactitud real.

Revisión de comentarios reales

Revisamos una muestra autorizada de 90 comentarios anónimos no vacíos e inventariamos manualmente 28 referencias indirectas específicas a personas o grupos. Los spans PERSON aceptados por la primera versión coincidieron con 15 de esas 28 referencias. La versión ampliada coincidió con 6 de 28. Un diagnóstico separado

con texto normalizado encontró 7; el resultado principal sigue siendo 6. Entre los casos omitidos había referencias formuladas mediante roles de docentes y amistades. Son descripciones de categorías, no citas.



Figura 4. Coincidencias descriptivas, no recall, precisión ni F1. No había un conjunto de referencia anotado de forma independiente.

Son observaciones de cobertura únicamente. No hubo anotación independiente de todos los comentarios, conjunto de referencia adjudicado ni una estimación válida de precisión o F1 en datos reales. No publicamos texto, nombres, organizaciones, identidades ni salidas crudas. La revisión mostró además un desacuerdo de definición: el benchmark inicial interpretaba PERSON sobre todo como nombres y apodos, mientras que el producto necesitaba referencias expresadas mediante roles, vínculos y grupos acotados. Ampliar la etiqueta no mejoró la cobertura observada.

La primera exportación privada seleccionó por error una organización de demo/prueba y devolvió 59 comentarios. La descartamos como evidencia de respuestas reales de estudiantes. Ninguna fila estaba marcada como simulada, así que una marca por sí sola no bastaba para confirmar la procedencia. Después verificamos la organización correcta y exportamos IDs y texto en una transacción de solo lectura. La fuente verificada tenía 127 pulsos, 90 comentarios anónimos no vacíos y 37 campos de comentario vacíos. Volvimos a comprobar los conteos contra la misma fuente. El error anterior se registra porque cambió lo que podíamos afirmar, no porque haya cambiado los datos del producto.

La brecha entre datos sintéticos y reales no se explica por un único bug. El primer experimento tenía un contrato de offsets incorrecto; ubicar el texto de manera determinista arregló la medición técnica, pero no resolvió la diferencia entre menciones sintéticas pautadas y referencias indirectas reales. El puntaje sintético de alrededor de 0,90 y la observación de 6 sobre 28 no se contradicen: miden cosas distintas, con datos distintos. Solo el resultado sintético tenía un conjunto estructurado de referencias.

Qué podemos afirmar

La auditoría de entrenamiento y desarrollo no encontró solapamiento directo de texto, familias ni valores exactos. Eso no prueba independencia. Ambos conjuntos usaron el mismo renderizador y la misma ontología. Un subconjunto exploratorio con combinaciones de tipos de entidad no vistas obtuvo F1 0,817 y menor recall que el conjunto completo; solo tenía cinco familias y ningún negativo difícil. Es una alerta, no un experimento causal limpio. No se informa un puntaje de prueba independiente y el conjunto sellado no se evaluó para estos resultados.

La conclusión acotada es que el sistema aprendió patrones útiles en su tarea sintética, pero no demostramos que pudiera encontrar las referencias contextuales que necesitaban las personas en comentarios organizacionales reales. No se debería vender un flujo de privacidad apoyándose solo en desempeño sintético.

3. Moderación: el benchmark no alcanzó

La primera versión de moderación fue un clasificador compacto ajustado con 212 ejemplos sintéticos. Ese mismo conjunto de desarrollo se usó para entrenar, así que su puntaje fue dentro de muestra y no estimaba generalización. La evaluación separada de 212 ejemplos tampoco era prístina: ya se había usado en trabajo previo de referencia. La segunda versión también se evaluó sobre esa misma porción de 212 ejemplos ya expuesta.

La segunda versión mezcló 3.000 ejemplos sintéticos con 2.968 ejemplos públicos. Las etiquetas públicas se usaron solo para la tarea de moderación; las etiquetas no observadas se enmascararon y no se trataron como negativas. Las etiquetas sintéticas las asignó el generador, que tenía solo 30 familias de plantillas. El entrenamiento terminó en CPU sin costo de cómputo. La época seleccionada se eligió usando el mismo desarrollo de 212 ejemplos y después se volvió a usar una prueba pequeña ya vista. Las referencias se recibieron de trabajo previo y no se volvieron a ejecutar para esta comparación.

Etiqueta	F1 v1	F1 v2	Referencia de vecinos	Referencia de sistema completo
Moderación	0,837	0,462	0,896	0,905
Sarcasmo	0,686	0,457	0,797	0,715
Riesgo	0,942	0,389	0,959	0,964
Falso / spam	0,851	0,635	0,905	0,958

La segunda versión quedó por debajo de ambas referencias en las cuatro etiquetas. Eso no significa que no sirvan los modelos de moderación. Significa que este experimento no produjo uno mejor. La comparación estuvo limitada por una prueba pequeña y reutilizada, supervisión sintética, cobertura incompleta de etiquetas y fuentes distintas. Frenamos en lugar de seguir optimizando contra los mismos ejemplos.

La publicación del 2 de octubre atribuyó al sistema preview los valores 0,905, 0,715, 0,964 y 0,958. Después vimos que esos valores correspondían a la referencia de sistema completo que nos habían pasado, no al modelo preview. Los valores de la primera versión fueron 0,837, 0,686, 0,942 y 0,851; los de la segunda fueron 0,462, 0,457, 0,389 y 0,635 en la porción reutilizada. Corregimos la atribución en esta retrospectiva. Ambos grupos de resultados son sintéticos y limitados; ninguno estima limpiamente el desempeño con comentarios reales.

4. Por qué no lanzamos el piloto pago

Después de los experimentos en CPU consideramos adaptar una base de mayor tamaño para moderación. El piloto propuesto no llegó a entrenamiento pago ni a evaluación alojada. El acceso a las bases candidatas estaba restringido, no se había establecido la aceptación de licencias y no pudimos verificar un límite mecánico que cubriera descarga, inicio, entrenamiento, carga del adaptador, inferencia y apagado. También faltaba una validación independiente confiable. La instrucción fue mantener el piloto frenado. No hubo lanzamiento pago, carga de adaptador, solicitud de prueba ni cambio en producción.

Fue una decisión de frenar, no un resultado fallido de entrenamiento en GPU. Ese piloto no tiene puntaje. No presentamos una comparación hipotética de calidad ni describimos como terminado algo que no se ejecutó.

5. Contabilidad de costos

Separamos uso de recursos, montos facturados, estimaciones derivadas del uso y reservas no utilizadas. No son intercambiables. Los créditos pueden dejar una factura en cero mientras el uso medido sigue siendo distinto de cero; calcular un precio a partir de tokens no prueba que haya llegado una factura; y una reserva no es gasto.

Categoría	Uso medido u observado	Facturación o estimación
Instantánea del espacio de trabajo después del entrenamiento v2 de entidades	US\$ 1,54 medidos a nivel de espacio de trabajo	US\$ 0,00 facturados tras US\$ 1,54 en créditos en la instantánea devuelta
Tres filas identificadas de recursos de entrenamiento de entidades	US\$ 1,49 en total: US\$ 0,04, US\$ 0,82 y US\$ 0,63	No hay comprobantes separados por trabajo; las filas y los resúmenes del espacio no concilian
Registro principal de inferencia de texto	2.326.196 tokens de entrada en 2.000 solicitudes completadas; 5 solicitudes iniciadas con uso desconocido	Cerca de US\$ 0,56 con el precio histórico archivado; otro estimado que incorpora incertidumbre también redondea a cerca de US\$ 0,56. Se solapan y no son facturas

Categoría	Uso medido u observado	Facturación o estimación
Evaluación DEV alojada de entidades	43.800 tokens devueltos; una solicitud incierta	No hay factura verificada
Una prueba aislada de entidades	224 tokens	No se obtuvo el monto facturado por separado
Piloto pago de adaptación de moderación no ejecutado	No hubo entrenamiento pago ni evaluación alojada	No se conoce un cargo de cómputo del piloto porque no se lanzó

Estas cifras no son un total general del proyecto. La instantánea del espacio de trabajo es de nivel de cuenta, no una factura específica del proyecto. Redondeados a centavos, el componente informado de aplicaciones efímeras y la medición agregada dan ambos US\$ 1,54; en los registros originales, el componente es levemente mayor. Las tres filas de trabajos redondean a US\$ 1,49, una diferencia aproximada de US\$ 0,05 frente al agregado, sin asignación demostrable. Conservamos la discrepancia en lugar de inventar una conciliación. Los dos valores en dólares del registro de inferencia son estimaciones de uso superpuesto; redondeados a centavos, ambos dan cerca de US\$ 0,56, y no se deben sumar. No se pudo verificar la facturación real de esas solicitudes. Los presupuestos históricos, proyecciones y reservas no son gastos.

El piloto propuesto no se lanzó y su antiguo límite de costo no es gasto. Con la evidencia disponible no se puede establecer un monto exacto total de todos los servicios.

Escala de inferencia: proxy histórico solo de tokens de entrada

Para una comparación aproximada de API, el registro histórico da un promedio de 824,985849 tokens de entrada por llamada de moderación (349.794 tokens en 424 llamadas) y 4.130,858182 tokens por pasada completa de entidades (1.135.986 tokens en 275 pasadas). Una llamada y una pasada suman 4.955,844031 tokens de entrada por comentario no vacío. Al precio histórico archivado de US\$ 0,24 por millón de tokens de entrada, son cerca de US\$ 0,0012 por comentario. Los cálculos exactos en dólares quedan en el registro interno de afirmaciones. Es un supuesto histórico fechado, no un precio actual ni una factura.

Comentarios no vacíos por mes	Moderación, solo entrada	Entidades, solo entrada	Combinado, solo entrada
1.000	US\$ 0,20	US\$ 0,99	US\$ 1,19
10.000	US\$ 1,98	US\$ 9,91	US\$ 11,89
100.000	US\$ 19,80	US\$ 99,14	US\$ 118,94

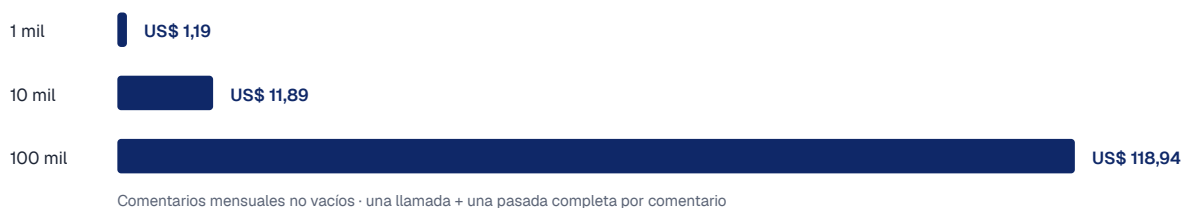


Figura 5. Ilustración lineal del mismo proxy histórico solo de entrada. No es una factura de producción ni una proyección del costo operativo total.

El cálculo es deliberadamente más estrecho que una estimación de producción. Supone un comentario no vacío por pulso y exactamente una llamada de moderación más una pasada completa de entidades. Excluye tokens de salida, reintentos, rondas LACE adicionales, agregación, solicitudes de plataforma, almacenamiento, monitoreo, revisión humana y soporte. Tampoco modela si el recorrido actual del producto haría esas llamadas con esa frecuencia. Los comentarios vacíos no generarían las llamadas asumidas, pero desconocemos qué proporción tendrán en el tráfico futuro.

No están establecidos el servicio alojado con acelerador, la capacidad en espera, el batching, la latencia, la utilización ni los precios de almacenamiento. Una fórmula general de cómputo sería precio horario del acelerador por horas activas, ajustadas por utilización, más las horas de capacidad ociosa; se desconocen el precio, la latencia de producción, la utilización y la capacidad en espera necesarios. También son desconocidos los costos

de reentrenamiento y evaluación porque todavía no se definieron el conjunto de datos, la cantidad de ejecuciones, el recorrido de inferencia ni el protocolo de revisión. Estos costos siguen sin determinarse. Este proxy histórico de API no se debe comparar como si incluyera alojamiento en GPU, reentrenamiento o un programa completo de evaluación.

6. Qué cambió

Dejamos de tratar un modelo ajustado propio como la forma predeterminada de mejorar Moodinary. La próxima línea de investigación es el pronóstico del clima emocional a partir de respuestas agregadas, sin presentar datos generados como si representaran a personas reales. La otra es bajar los costos de plataforma. Ambas se pueden contrastar con necesidades concretas del producto, sin mantener un detector débil solo porque ya se entrenó.

El trabajo sigue sirviendo como evidencia de ingeniería. Encontramos el error de offsets, vimos cómo una nota dentro de muestra puede parecer convincente y detectamos el desajuste entre la definición de las etiquetas y la necesidad del producto. También nos obligó a separar mejor el uso medido del costo facturado. Conservar estos fallos es más valioso que rebautizar el experimento como un lanzamiento.

7. Límites y manejo de datos

Todos los datos de desarrollo de entidades que se informan acá eran sintéticos y los generó el sistema de autoría. Moderación mezcló datos sintéticos y una fuente pública etiquetada, pero la evaluación siguió siendo pequeña y reutilizada. Ningún resultado demuestra desempeño con usuarios argentinos representativos. La muestra de comentarios reales se resume solo con conteos agregados. Este documento no incluye texto crudo, nombres, organizaciones, identificadores, anécdotas identificables ni salidas privadas generadas.

Esta es una retrospectiva de proyecto, no un benchmark preregistrado. Algunas decisiones de puntuación y diagnósticos aparecieron después de observar fallos. Distinguimos los resultados originales congelados de los diagnósticos posteriores, pero esa secuencia limita la interpretación confirmatoria. El proyecto no demostró un sistema listo para desplegar, una garantía de seguridad ni una ventaja general de la revisión cruzada.

Conclusión

ECO-1 empezó con una idea razonable: ajustar sistemas pequeños para tareas concretas y usar revisión cruzada para volver más confiables las salidas. La evidencia no justificó seguir. Los puntajes sintéticos cercanos a 0,90 ocultaban un problema de transferencia; el clasificador de moderación quedó por debajo de las referencias; y el piloto pago no pudo superar las barreras de acceso, presupuesto y validación. Sin datos reales adecuados y consentidos, o datos sintéticos que transfirieran mucho mejor, no había una próxima evaluación creíble.

Cancelamos el programa de construcción de modelos y conservamos sus artefactos como registro de investigación. El trabajo que sigue apunta a las preguntas del producto que todavía están abiertas.

Notas de fuentes

Los protocolos, hashes, detalles de evidencia y el mapa de fuentes de contabilidad permanecen en un registro interno de investigación. La muestra de comentarios reales no está disponible para reproducción. No se reproduce texto de participantes.