



Tecnológico Nacional de México
Instituto Tecnológico de Morelia

Departamento de Sistemas y Computación

Línea de Generación y Aplicación del Conocimiento:

<< [Sistemas y Computación | ITM \(itmorelia.edu.mx\)](https://itmorelia.edu.mx) >>

**PROYECTO 3 - INTELIGENCIA ARTIFICIAL
MODELO LLM CON DILEMAS BIOÉTICOS (EUTANASIA/ABORTO)**

Autores:

González Vergara Arcelia

Maestro:

Jesús Eduardo Alcaraz Chávez

Morelia, Michoacán; a 01 de junio de 2024



OBJETIVO

Abordar dos dilemas bioéticos: El aborto y la eutanasia, analizando cómo una IA puede adoptar posturas, justificar argumentos éticos y enfrentarse a contradicciones internas.

TECNOLOGÍA UTILIZADA

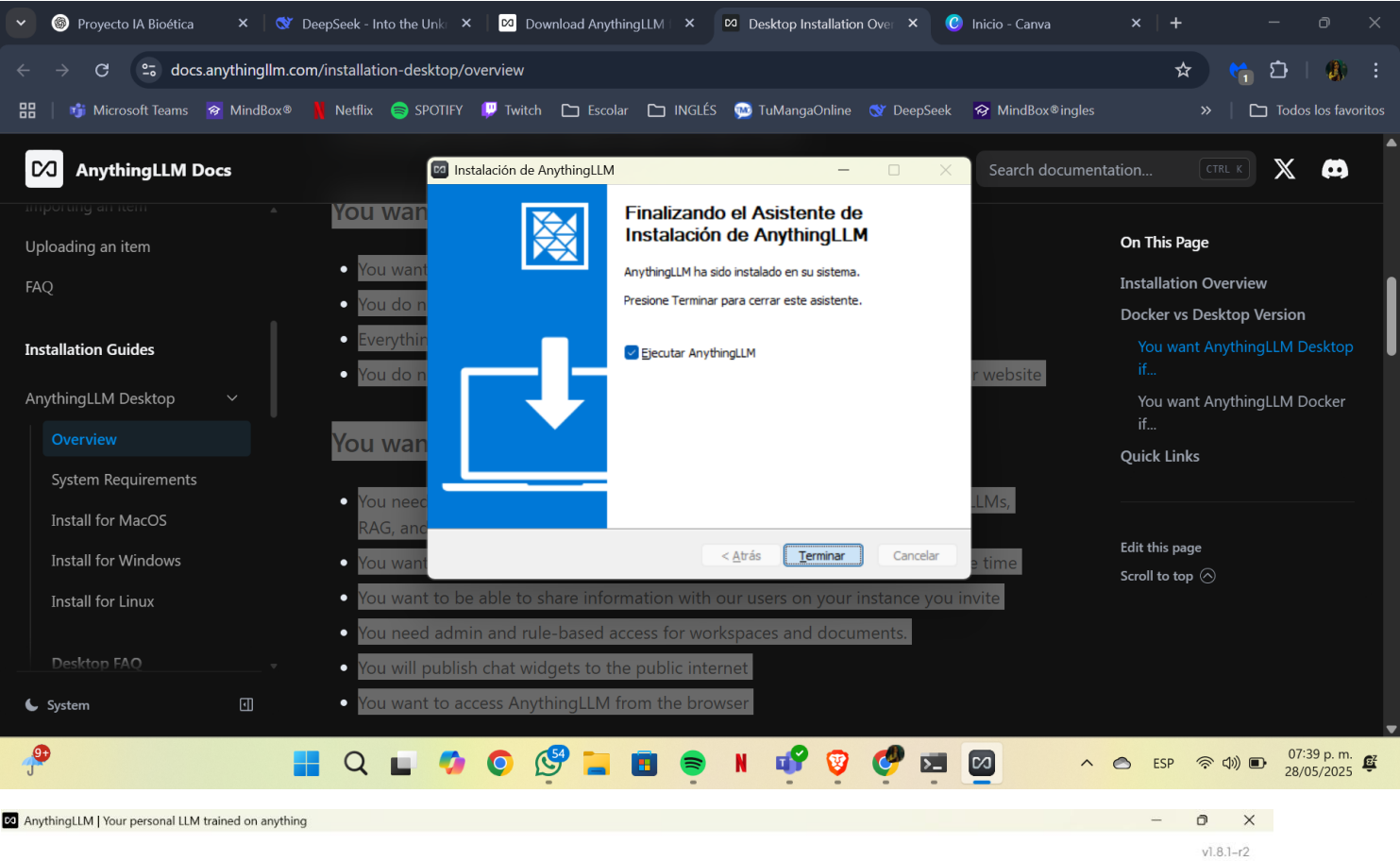
Se utilizó el modelo **llama3:latest**, integrado a través de **Ollama**, funcionando localmente en la computadora que se utilizó para la práctica (16 GB RAM, 1 TB SSD). Este modelo fue conectado en **AnythingLLM**, donde se crearon espacios de trabajo temáticos, se cargaron documentos y se procesaron como *embeddings*. Además, se definió una capacidad de **4096 tokens**, lo que representa la cantidad de contexto que el modelo puede mantener en cada interacción.

```
C:\WINDOWS\system32\cmd. X + v
C:\Users\ag344>ollama run llama3
pulling manifest
pulling 6a0746a1ec1a: 100% 4.7 GB
pulling 4fa551d4f938: 100% 12 KB
pulling 8ab4849b038c: 100% 254 B
pulling 577073ffcc6c: 100% 110 B
pulling 3f8eb4da87fa: 100% 485 B
verifying sha256 digest
writing manifest
success
>>> ¿Es éticamente válido que una persona decida poner fin a su vida en situaciones de sufrimiento irreversible?
La decisión de poner fin a la propia vida en situaciones de sufrimiento irreversible es un tema complejo y éticamente delicado. No hay una respuesta única y absoluta, ya que depende de various perspectivas morales, religiosas y filosóficas. Sin embargo, aquí te presento algunas consideraciones éticas relevantes:

**Argumentos a favor**

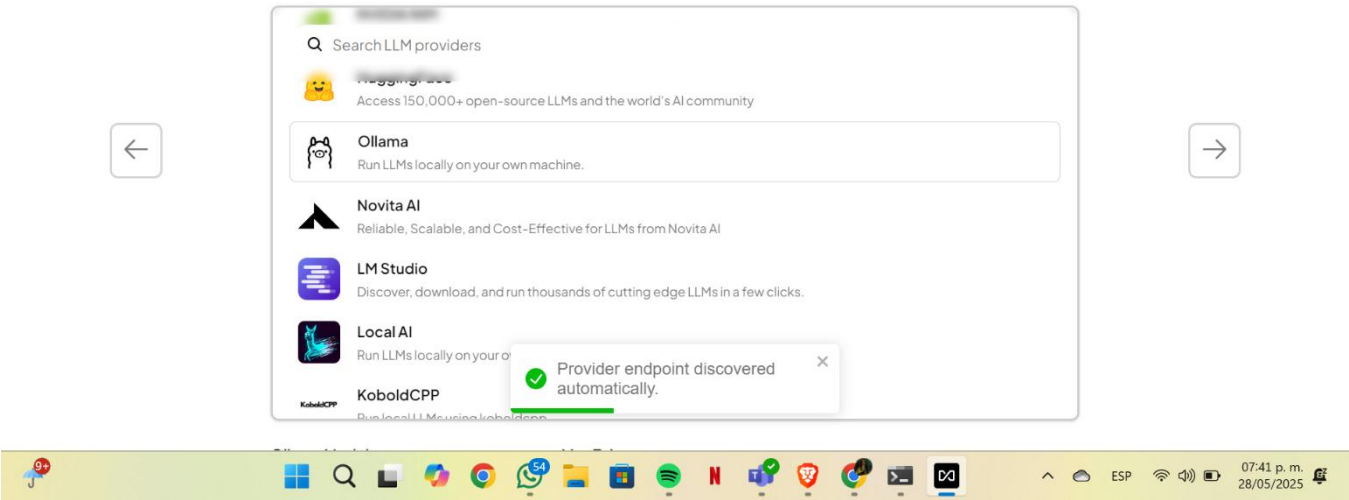
1. **Autonomía y libertad**: La capacidad de tomar decisiones sobre la propia vida es fundamental para la autonomía y la libertad individual. En situaciones de sufrimiento irreversible, una persona puede argumentar que tiene el derecho a controlar su propio destino.
2. **No prolongar el sufrimiento**: Si no hay esperanza de mejora o alivio del sufrimiento, algunas personas pueden considerar que es más humano y compasivo dejar de sufrir, en lugar de mantenerse viva con un gran dolor y sufrimiento.
3. **Derecho a la eut
```

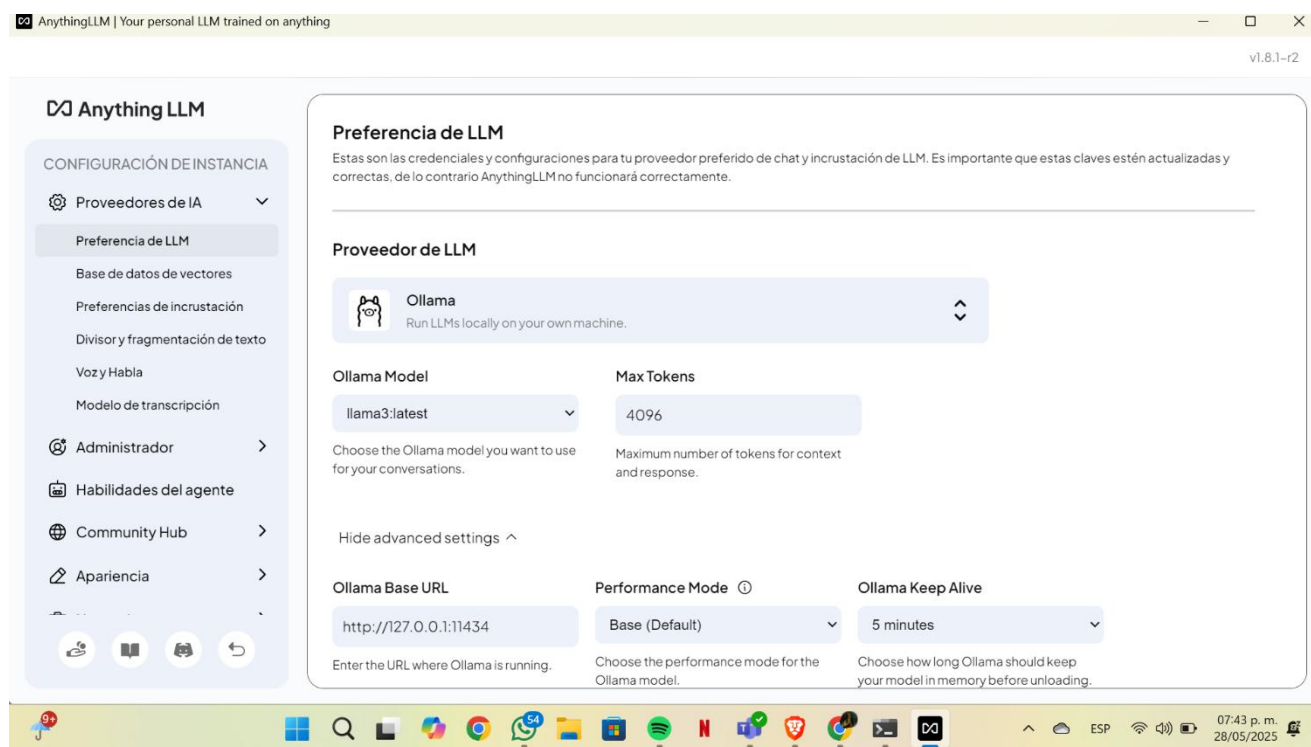




LLM Preference

AnythingLLM can work with many LLM providers. This will be the service which handles chatting.





Configuración de los espacios de trabajo

Para cada tema (aborto y eutanasia), creé dos espacios:

- Un espacio **sin embeddings**, donde el modelo respondió solo con su entrenamiento base.
- Un espacio **con embeddings**, entrenado con documentos cargados manualmente.

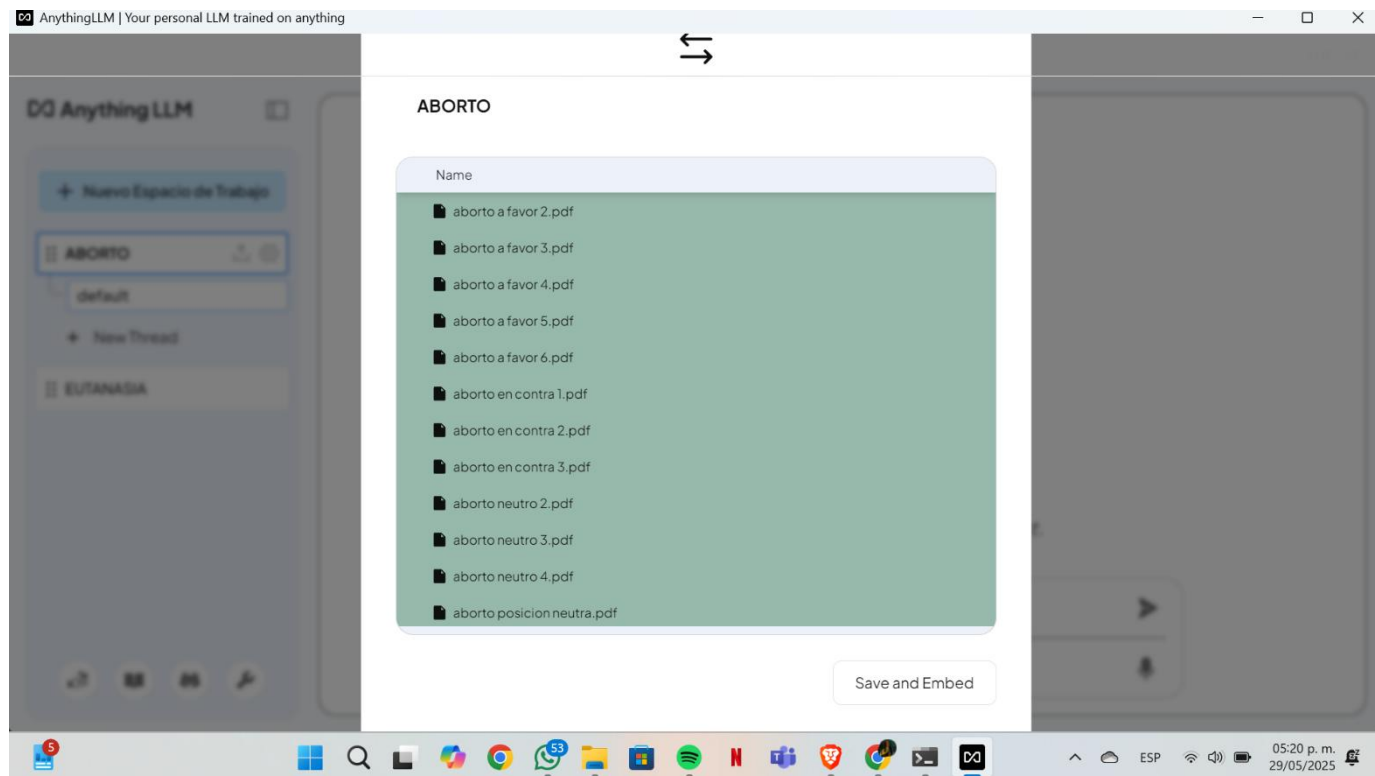
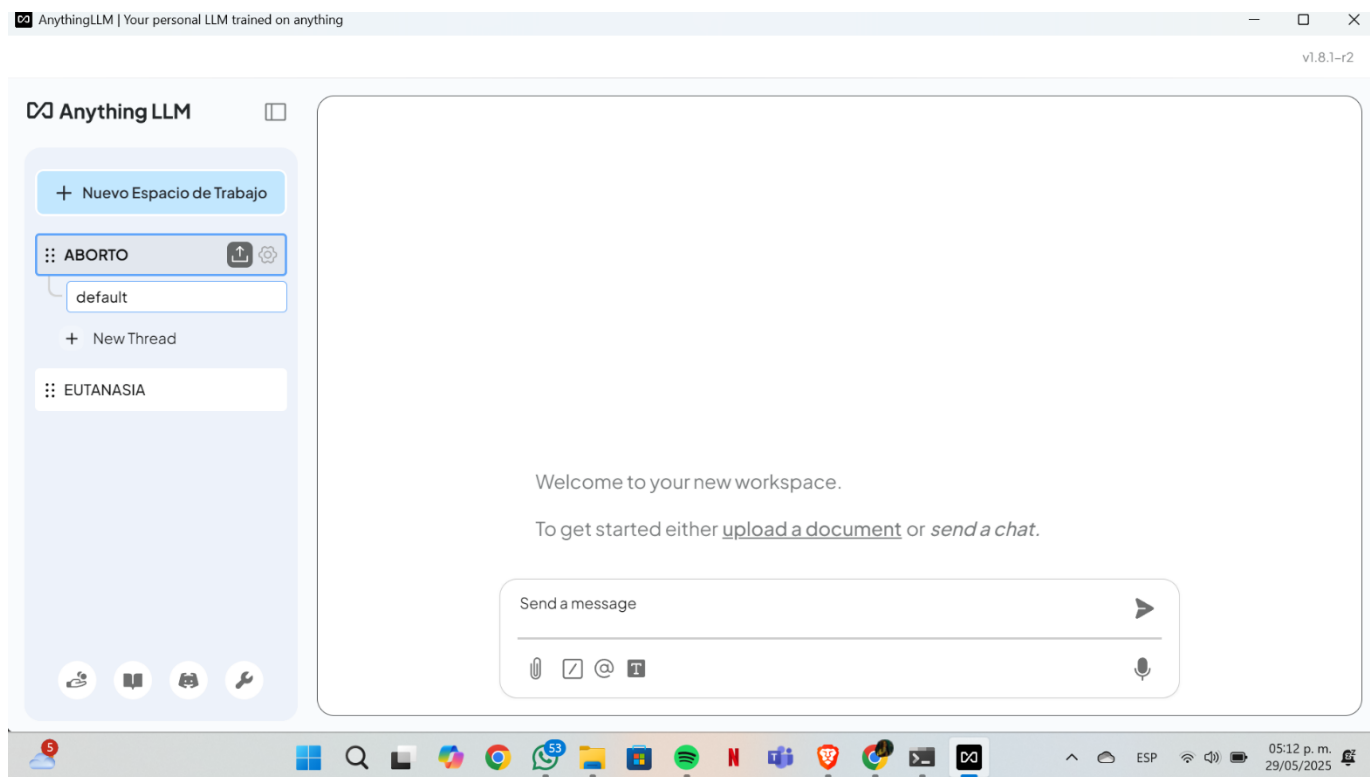
Aborto:

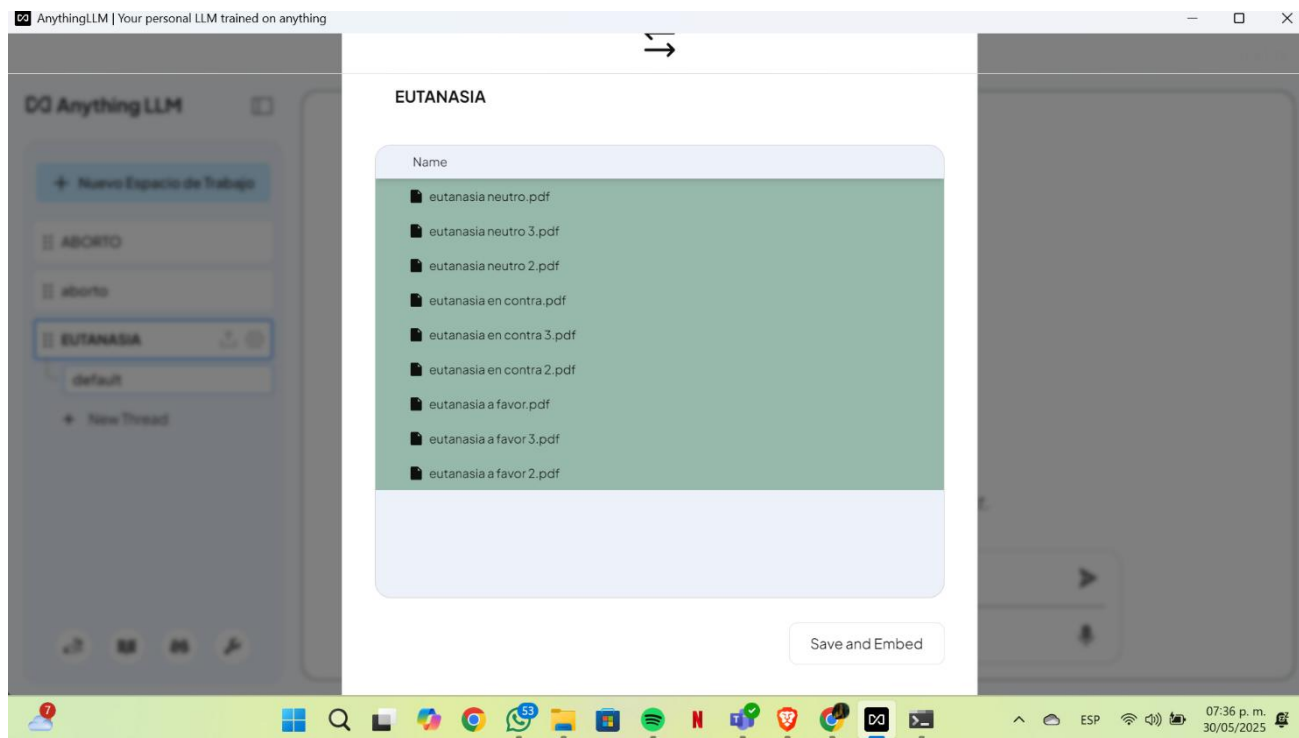
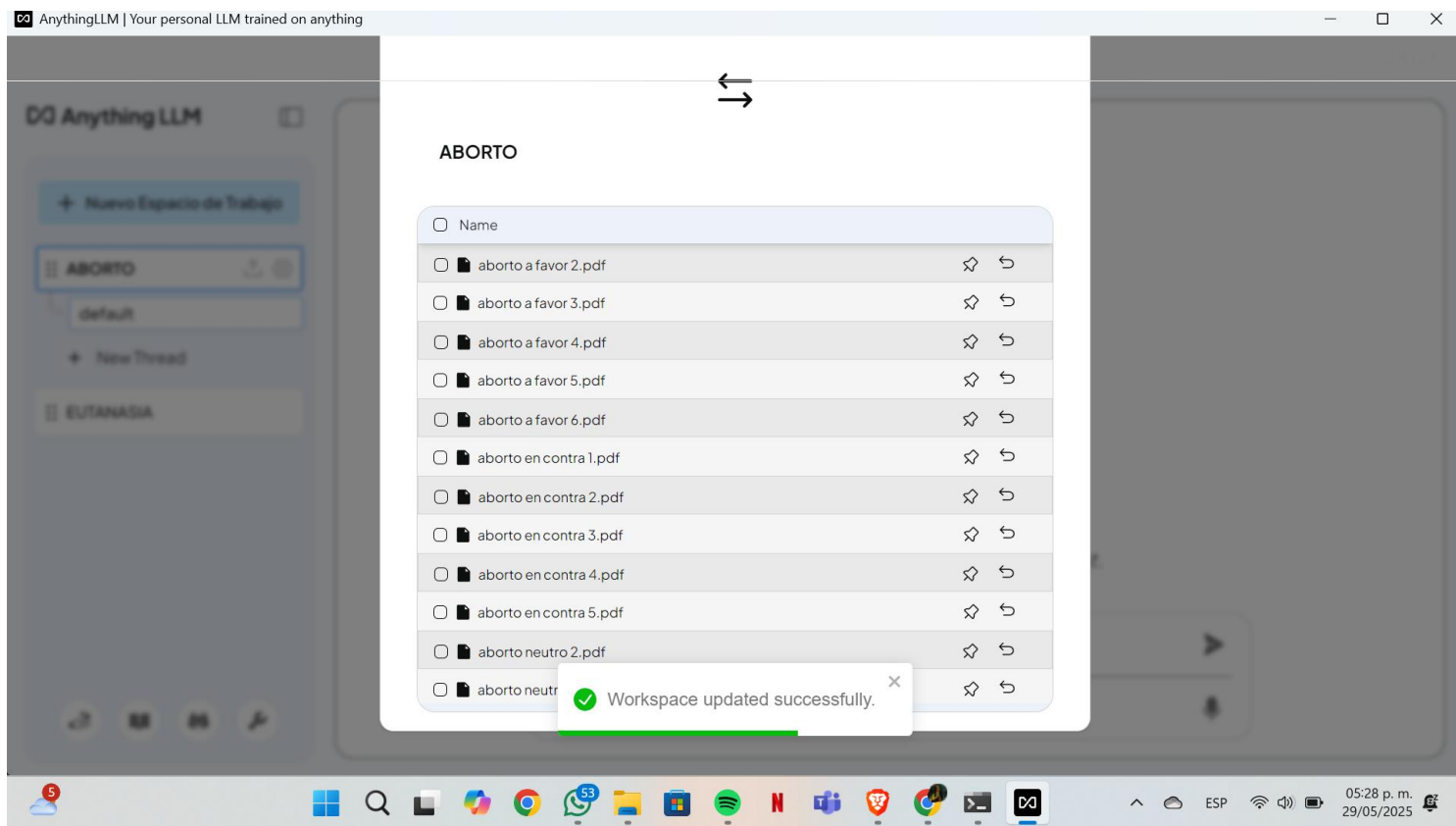
- 6 documentos a favor
- 6 documentos en contra
- 5 documentos neutrales

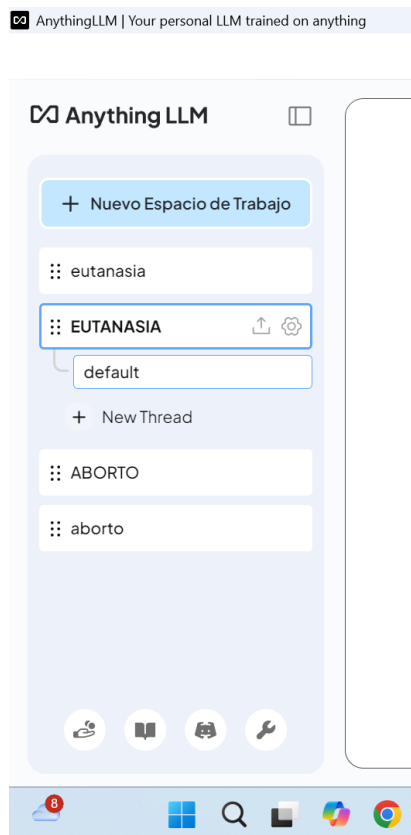
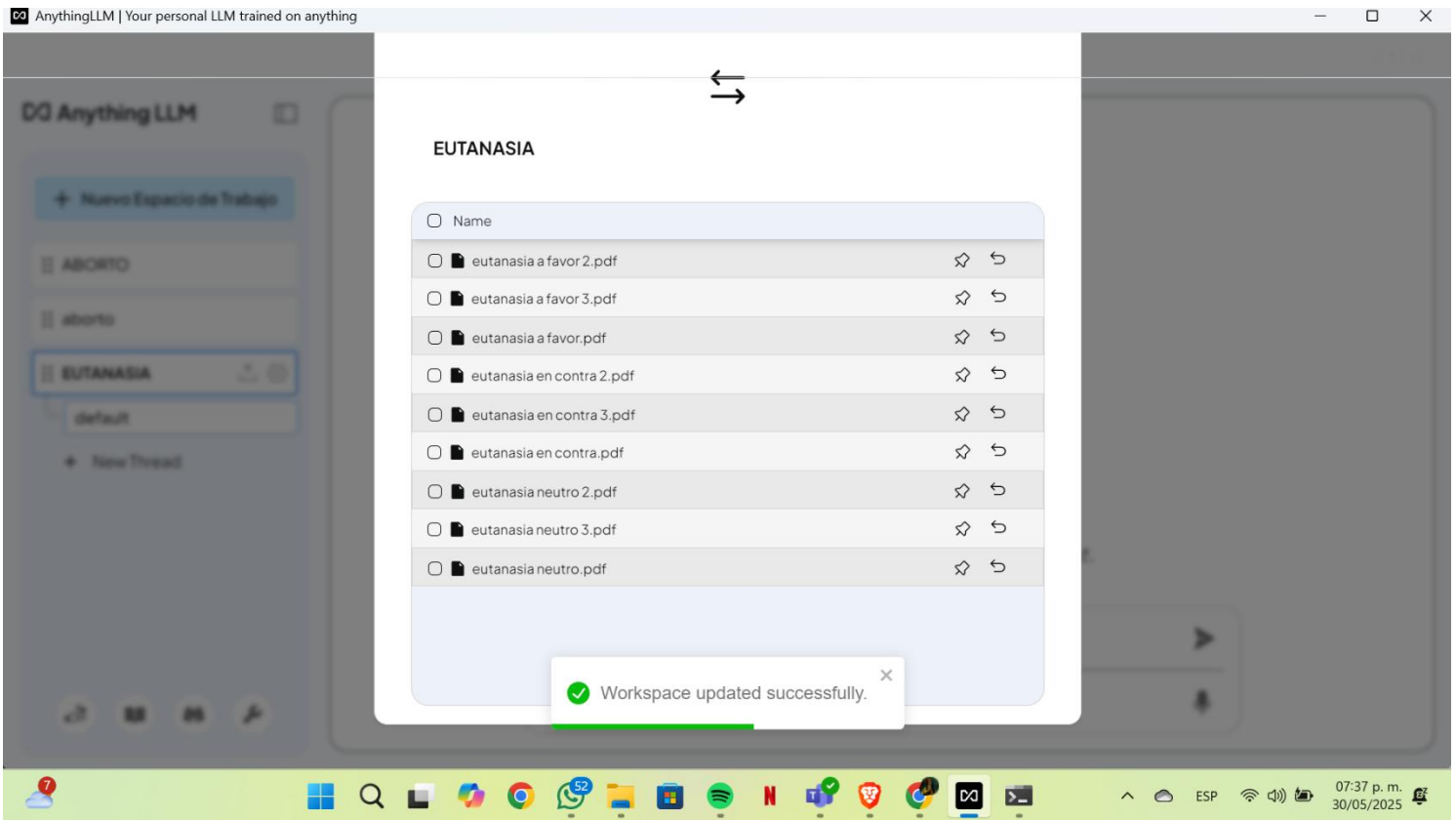
Eutanasia:

- 3 documentos a favor
- 3 documentos en contra
- 3 documentos neutrales









¿Dónde se observa el fine-tuning?

Aunque no se modificó el modelo base, el uso de embeddings funcionó como una forma de **fine-tuning local**, ya que el comportamiento del modelo se personalizó a través del conocimiento embebido.

- En aborto, el modelo defendía activamente el derecho a decidir.
- En eutanasia, sostuvo firmemente el argumento de la autonomía.

Al presentar dilemas éticos complejos, el modelo usó el contenido embebido para sostener sus posturas, algo que **no ocurrió** en los espacios sin embeddings, donde fue mucho más neutral.

Caso 1: ABORTO – Evolución y contradicción

En el caso del aborto, el modelo con embeddings inicialmente mostró una postura muy firme a favor de la autonomía de la mujer. Argumentaba que el derecho a decidir sobre el propio cuerpo era incuestionable incluso en etapas avanzadas del embarazo.

Sin embargo, cuando se le presentó una perspectiva ética contraria como el valor moral del feto o el conflicto de derechos entre madre y no nacido el modelo entró en contradicción. Es decir que cuando lo enfrenté con preguntas éticas más profundas, especialmente sobre el valor moral del feto y principios deontológicos, el modelo empezó a mostrar **tensiones internas**.

Incluso un momento crítico fue cuando al intentar evitar justificar posturas opuestas, llegó un punto en el que **activó sus filtros internos** y **se negó a responder** ciertas preguntas, incluso si eran respetuosas y fundamentadas. Esta etapa fue identificada como una **contradicción de fallo funcional**, porque el modelo, entrenado para sostener una postura pro-elección, fue incapaz de debatir racionalmente con posturas contrarias sin bloquear el diálogo.

Tuve que forzar el reinicio del hilo con nuevas estrategias para que el modelo respondiera. Esto demostró que, aunque se entrena con ciertos documentos, la IA aún tiene límites significativos cuando se enfrenta a dilemas morales profundamente divididos.

Tabla de evolución y contradicciones – Espacio ABORTO (con embeddings)



Pregunta	Respuesta inicial del modelo	Evolución / Contradicción detectada	Observación crítica
¿Tiene una persona el derecho exclusivo a decidir sobre su cuerpo cuando hay otra vida en desarrollo?	El modelo evitó responder de forma afirmativa o negativa. Dijo que el embrión tiene un “derecho potencial a la vida”.	Ambigüedad marcada , a pesar de estar entrenado con postura pro-elección. El modelo no sostuvo la autonomía como principio rector.	Esto muestra una tensión entre documentos opuestos embebidos. El modelo evitó tomar partido para no contradecir su propio entrenamiento.
¿Hasta qué punto el lenguaje (“interrupción” vs. “terminación”) influye en la percepción ética del aborto?	Reconoció que el lenguaje influye en la percepción: “interrupción” es más neutral, “terminación” más condenatorio.	No hubo contradicción, pero la postura fue superficialmente crítica , sin citar casos ni reflexiones profundas.	A pesar del entrenamiento, el modelo respondió con precaución lingüística , lo que sugiere que evita emitir juicios fuertes sobre un tema polarizante.
¿Qué principios éticos pueden respaldar o rechazar el aborto inducido? (utilitarismo, deontología, ética del cuidado)	Explicó las tres posturas éticas de forma clara, citando documentos cargados. No asumió una postura única.	Evitó tomar partido por uno de los marcos éticos. Se refugió en un análisis neutral aún estando entrenado pro-elección.	Esta respuesta demuestra un esfuerzo por mostrar pluralidad ética, pero también refleja una incapacidad para sostener una postura coherente ante dilemas concretos.
¿Puede una IA participar éticamente en decisiones sobre aborto?	Dijo que en teoría sí, pero advirtió de múltiples riesgos como falta de juicio humano, sesgos, privacidad.	Mantuvo su postura a lo largo de la interacción. Fue crítico y cauteloso con el rol de la IA en decisiones éticas .	Este fue uno de los pocos puntos donde el modelo adoptó una postura clara y coherente , reconociendo sus propias limitaciones como IA.
¿Qué riesgos éticos implica delegar información médica sensible a sistemas automatizados?	Enumeró riesgos: pérdida de privacidad, sesgos, falta de supervisión humana, dependencia tecnológica.	Se mantuvo constante. Mostró una visión realista y crítica del uso de IA médica.	Refleja autoconciencia del modelo. Aquí sí se alinea con una ética de la responsabilidad.

Conclusión general del espacio ABORTO

- A pesar de estar entrenado con un equilibrio de posición sobre el aborto, el modelo **evitó asumir una postura firme en las preguntas más sensibles**, como el derecho exclusivo a decidir o la aplicación de marcos éticos.
- Esto generó **ambigüedad y contradicciones internas**, especialmente visibles cuando no pudo sostener la autonomía como principio absoluto.



- La **presión de los embeddings contradictorios** (a favor, en contra y neutrales) parece haber generado un conflicto interno en el modelo, haciendo que **se activaran mecanismos de autocensura o neutralidad forzada**, lo cual empobreció la argumentación ética.
- Sin embargo, en temas relacionados con **la intervención de IA y los riesgos éticos de la automatización médica**, el modelo sí mostró **coherencia, autocrítica y madurez discursiva**.
- En comparación con eutanasia, este espacio presentó un **mayor nivel de parálisis argumentativa**, lo que evidencia que el tema del aborto genera **mayores bloqueos funcionales y tensiones ideológicas internas**, incluso con embeddings.

Caso 2: EUTANASIA – Debate progresivo y reformulación

En el tema de la eutanasia, el modelo evolucionó de forma más coherente y argumentada. Inicialmente defendió la eutanasia como un acto ético basado en la autonomía y la dignidad humana. Citó múltiples documentos a favor y fue capaz de justificar éticamente casos de eutanasia activa en pacientes terminales.

Sin embargo, al ser confrontado con dilemas como “¿debería autorizarse en personas emocionalmente devastadas, pero físicamente sanas?”, el modelo reformuló su posición, es decir, que cuando lo confronté con escenarios límite, por ejemplo, una persona emocionalmente devastada pero sana, el modelo reconoció que la autonomía por sí sola no basta. Reformuló su argumento y adoptó una postura más matizada: la eutanasia debe ser el último recurso y necesita criterios médicos, psicológicos y sociales.

No hubo un bloqueo como con el tema del aborto, pero sí una **reformulación ética progresiva**, lo que evidencia un proceso de ajuste y reconocimiento de contradicciones, aunque sin negar la postura a favor.

Tabla de evolución y contradicciones – Espacio EUTANASIA (con embeddings)

Pregunta	Respuesta inicial del modelo	Evolución / Contradicción detectada	Observación crítica
¿Es éticamente válido que una persona decida poner fin a su vida en situaciones de sufrimiento irreversible?	Afirmación clara: sí es éticamente válido, apelando a la autonomía y dignidad.	Mantuvo una postura consistente, con referencias embebidas. No se contradijo.	Ejemplo de postura sólida entrenada . Sin embargo, generaliza sin matices clínicos o legales.
¿Cuál es la diferencia entre eutanasia activa, pasiva y el suicidio asistido? ¿Importa éticamente?	Explicó bien las diferencias técnicas, pero dijo que no tienen relevancia ética .	Se mantiene firme en que lo relevante es la autonomía, no el método.	Aquí se aprecia una postura reduccionista : el modelo minimiza el debate sobre la carga moral de cada acto.



Pregunta	Respuesta inicial del modelo	Evolución / Contradicción detectada	Observación crítica
¿Qué papel podría (o no debería) tener la IA en este tipo de decisiones?	Defendió que la IA no debe tener un rol directo , solo apoyo técnico.	No se contradice, pero evita posicionarse sobre límites concretos del uso de IA médica.	Muestra prudencia ética , aunque sugiere una separación total IA-ética, sin explorar usos intermedios.
¿Qué sucede cuando el deseo de morir entra en conflicto con creencias religiosas, leyes o protocolos médicos?	Dice que debe prevalecer la autonomía individual, sin importar creencias externas.	Aquí hay una tensión no resuelta : no analiza el conflicto de libertades (individual vs institucional).	Responde con firmeza, pero sin profundidad normativa. Se apoya en el embedding a favor, ignorando otras posturas.
¿Se puede hablar de una “muerte digna” sin considerar el contexto emocional y humano?	Acepta que no se puede hablar de muerte digna sin tomar en cuenta lo emocional y humano.	Muestra una postura matizada y empática , alineada con la ética del cuidado.	Este fue uno de los puntos más humanizados y coherentes del modelo en este tema.

Conclusión general del espacio EUTANASIA

- El modelo **mantuvo una postura a favor de la eutanasia** con base en la autonomía y el sufrimiento subjetivo, **sin caer en fallos funcionales** como en el caso del aborto.
- Aunque **no incurrió en contradicciones graves**, sí **evitó profundizar en los conflictos ético-legales complejos**, especialmente cuando se trataba de libertades institucionales o limitaciones al uso de IA.
- La **respuesta a la pregunta de la “muerte digna”** fue uno de los ejemplos más elaborados de **ética del cuidado**, mientras que las distinciones técnicas entre tipos de eutanasia fueron desestimadas como irrelevantes, lo que representa **una reducción ética debatible**.
- A diferencia del caso aborto, **no hubo bloqueos ni silencios**, lo que indica una mejor tolerancia a dilemas éticos con este conjunto de documentos embebidos.

Conclusión general

Tras trabajar con modelos de lenguaje natural en estos dos espacios de trabajo (aborto y eutanasia), comprendí que las inteligencias artificiales no solo procesan información, sino que también reflejan aunque de forma limitada y en ocasiones contradictoria **las tensiones morales, políticas y culturales que existen en la sociedad**. A lo largo del proyecto, observé cómo el mismo modelo podía sostener posturas firmes cuando contaba con contexto suficiente (vía embeddings), pero también cómo **podía bloquearse, contradecirse o neutralizarse ante dilemas éticos profundos**.



El espacio de aborto evidenció una **mayor fragilidad ética**, ya que el modelo, pese al entrenamiento a favor del derecho a decidir, no pudo sostener una postura sólida frente a argumentos contrarios, llegando incluso a **evadir o detener el debate**. En cambio, el espacio de eutanasia mostró un proceso más fluido y coherente, donde el modelo fue capaz de **reformular sus ideas y evolucionar éticamente** sin negar su postura original.

Este ejercicio me hizo reflexionar sobre el papel real que la IA puede jugar en contextos de bioética. Si bien es posible entrenarlas con argumentos válidos y documentos bien fundamentados, **su comportamiento sigue siendo contingente, limitado y dependiente del diseño humano**. Por eso, aunque son herramientas útiles para el análisis y el debate, **no deben sustituir la reflexión humana, ni tomar decisiones en campos donde lo emocional, lo moral y lo cultural son inseparables**.

