

Evaluación de sistemas inteligentes basada en habilidades.
Fundamentos para una taxonomía de capacidades artificiales
AFRA 2025 – Simposio Contribuciones filosóficas a la evaluación de modelos de
lenguaje

Mauro Santelli

Trabajo en colaboración con Joaquín Toranzo Calderón y Ramiro Caso

IIF-SADAF-CONICET & Universidad de Buenos Aires

19 de septiembre de 2025

El problema: evaluación sin teoría

Estado actual

- La evaluación de LLMs depende de colecciones planas de benchmarks (HELM, etc.)
- No hay marco de principios que conecte diferentes habilidades
- Probamos *qué* pueden hacer los modelos, no *cómo* se relacionan las habilidades

Problema central: la selección no es neutral

- Elegir benchmarks implica una teoría de la inteligencia
- Los enfoques actuales carecen de estructura conceptual coherente
- Faltan relaciones explicativas entre capacidades

Necesidad inmediata

Necesitamos marcos estructurados que hagan explícitos y revisables los compromisos implícitos de evaluación.

HELM: síntoma del problema

Lo que hace HELM

- Más de 50 tareas agrupadas en categorías amplias (QA, Recuperación, Razonamiento, Resumen)
- Comparación paralela y plana entre benchmarks
- Foco en rendimiento agregado sin estructura explicativa

Lo que falta

- Justificación conceptual de la selección de tareas
- Relaciones entre habilidades y correlaciones de desempeño
- Narrativa composicional que explique progresión de capacidades

Ejemplo representativo

TruthfulQA, HellaSwag y RACE requieren comprensión lingüística, integración de conocimiento y razonamiento, pero HELM las trata como puntos de datos independientes.

El desafío del loro estocástico

*“Un modelo de lenguaje es un sistema para unir secuencias de formas lingüísticas de manera azarosa... sin referencia al significado: un loro estocástico.”
(Bender et al., 2021)*

Requisitos de Bender para competencia lingüística genuina

- ① Trasfondo común (+ conciencia)
- ② Intenciones comunicativas
- ③ Teoría de la mente
- ④ Responsabilidad (accountability)

Problema abierto

¿Cómo se relacionan estas dimensiones y cómo las probamos de manera sistemática?
Necesitamos un marco que operacionalice los insights filosóficos.

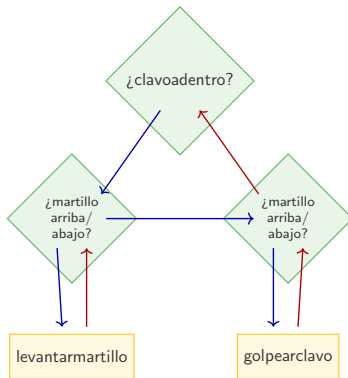
Insights clave de la filosofía

- **Debate Ryle-Stanley:** Saber-cómo vs. saber-qué
- **Brandom:** Las habilidades pueden elaborarse algorítmicamente
- **Estructura composicional:** Habilidades simples se combinan en complejas
- **Ciclos TOTE:** Test-Operate-Test-Exit como bloques constructivos

¿Qué nos aporta este enfoque?

- Caracterizaciones precisas de habilidades y sub-habilidades
- Explicación composicional para habilidades complejas
- Realizabilidad múltiple: distintos caminos a la misma capacidad
- Enfoque funcional en lo que los agentes pueden *hacer*

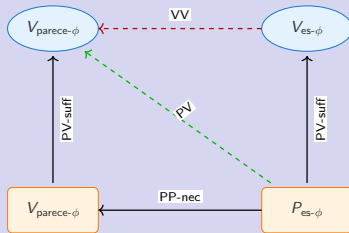
Ejemplo: Ciclos TOTE



Puntos clave

- Nodos de prueba (diamantes): verificación de condiciones
- Nodos de operación (rectángulos): acciones
- Flechas azules: ejecución hacia adelante
- Flechas rojas: bucles de retroalimentación

MUDs de Brandom: ejemplo Sellars y EPM



Poder explicativo

Entender cómo las relaciones práctica-vocabulario y vocabulario-vocabulario generan competencias complejas.

Chollet: “On the measure of intelligence”

Crítica a evaluaciones actuales

- Mayoría “basadas en destrezas” y “estrechas”
- Tareas definidas de manera precisa
- No miden habilidades amplias generalizables
- Cuatro tipos: humana, caja blanca, confrontación, benchmarks

Propuesta de Chollet

- **Habilidad:** constructo abstracto (teoría + estadística)
- **Destreza:** medida de rendimiento en tarea específica
- **Inteligencia:** eficiencia en adquirir nuevas destrezas
- Inspirado en psicometría (factor g)

Insight clave

Destreza es específica; inteligencia es capacidad meta de aprender. Falta de generalización fuera de distribución = obstáculo mayor.

Introduciendo TAAs: taxonomías de habilidades artificiales

Definición

Mapeo jerárquico de habilidades que va de lo general a lo específico.

Capacidad



Habilidad



Destreza

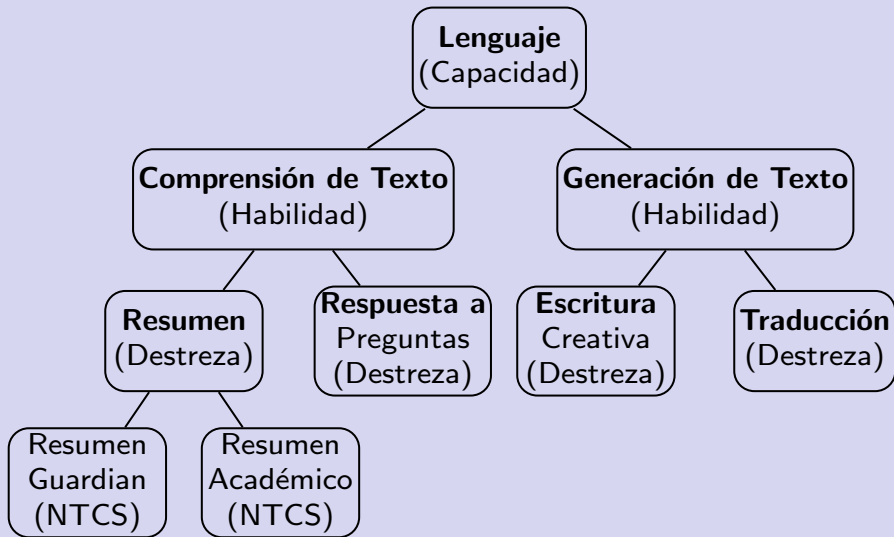


NTCS

Características clave

- Jerárquico: niveles altos son generales; niveles bajos, específicos
- Composicional: habilidades complejas se construyen desde simples
- Realizabilidad múltiple: distintos caminos para la misma capacidad
- Testeable: relaciones explícitas permiten predicción y diagnóstico

Ejemplo TAA: competencia lingüística



Estructura formal de TAAs

Relaciones entre habilidades

- **Composición estricta:** A requiere todas en $\{B, C, D\}$
- **Realizabilidad múltiple:** A se obtiene con $\{B, C\}$ o $\{B, D\}$
- **Requisitos graduales:** más de B habilita mejor desempeño en A

Ejemplo: requisitos de Bender

Sea $[LCC] = [\text{Competencia Comunicativa Lingüística}]$. Entonces $[LCC] \triangleright \{[\text{Trasfondo Común}], [\text{Intenciones Comunicativas}], [\text{Teoría de la Mente}], [\text{Responsabilidad}]\}$.

Conceptos clave

- **NTCS:** Narrow Task-Centric Skills (benchmarks específicos)
- **Destrezas:** conjuntos de NTCS estrechamente relacionados
- **Habilidades:** generalización más amplia a través de tipos de destrezas

De evaluación plana a estructurada

Enfoque actual

- Lista plana de benchmarks
- Agrupaciones arbitrarias
- Sin estructura explicativa
- Relaciones ausentes entre tareas
- Selección ad hoc

Enfoque informado por TAA

- Selección de tareas con principios
- Jerarquía explicativa explícita
- Predicción de rendimiento entre tareas
- Capacidad diagnóstica
- Pruebas dirigidas

Ejemplo práctico

- Si un modelo falla en [Resumen] y [Respuesta a Preguntas], TAA sugiere inspeccionar [Comprensión de Texto]
- Si [Comprensión de Texto] es débil → enfocar entrenamiento ahí
- Si [Comprensión de Texto] es fuerte → buscar otros cuellos de botella

Con TAAs, los requisitos de Bender se vuelven testeables

① Trasfondo común [CB]

- Sub-habilidades: conocimiento cultural, interpretación contextual
- Pruebas: razonamiento dependiente del contexto, resolución de referencias culturales

② Intenciones comunicativas [CLI]

- Sub-habilidades: planificación de objetivos, mantenimiento de coherencia
- Pruebas: diálogo dirigido por objetivos, comunicación estratégica

③ Teoría de la mente [ToM]

- Sub-habilidades: seguimiento de creencias, reconocimiento de intenciones
- Pruebas: tareas de falsa creencia, escenarios de toma de perspectiva

Impacto

TAAs vuelven empíricamente tractables debates filosóficos que antes quedaban a nivel abstracto.

Múltiples perspectivas filosóficas

TAA's acomodan diferentes compromisos teóricos

- **Funcionalistas:** foco en mapeos entrada-salida y equivalencia conductual
- **Mecanicistas:** requisitos sobre implementaciones y procesos internos
- **Pragmatistas:** énfasis en éxito contextual y comportamiento adaptativo

Beneficios del marco

- Un mismo esqueleto estructural admite distintas instanciaciones
- Hace explícito dónde divergen las visiones teóricas
- Facilita resolver disputas filosóficas vía evidencia empírica
- Permite competencia entre taxonomías sin perder comparabilidad

Ejemplo

¿Qué cuenta como “comprensión”? Distintas teorías demandan sub-habilidades diferentes y llevan a predicciones testeables distintas.

Agenda de investigación inmediata

Desarrollo técnico

- 1 Desarrollar TAAs concretas para familias de capacidades centrales (lenguaje, razonamiento, etc.)
- 2 Mapear benchmarks existentes a jerarquías de habilidades
- 3 Evaluar poder predictivo: ¿la estructura TAA explica correlaciones?
- 4 Crear herramientas diagnósticas para evaluar capacidades específicas

Trabajo colaborativo

- Compromiso constructivo con HELM y otros enfoques de evaluación
- Vocabulario compartido entre IA y filosofía
- Marcos abiertos que permitan múltiples enfoques teóricos
- Puentes entre teoría y práctica en evaluación de IA

Por qué esto importa ahora

Para investigación en IA

- Pasar del “benchmark climbing” a evaluación con principios
- Entender qué medimos realmente en LLM
- Base más sólida para discutir capacidades tipo AGI
- Estrategias de evaluación y entrenamiento más eficientes

Para colaboración interdisciplinaria

- Vocabulario común y marco conceptual compartido
- Compromisos explícitos que habilitan desacuerdo productivo
- Ruta estructurada para avanzar en problemas difíciles

Para filosofía

- Dominio de aplicación concreto para teorías de habilidades
- Nuevo banco de pruebas empírico para marcos filosóficos

Conclusiones clave

Conclusiones clave

- ❶ **Problema:** la evaluación actual de LLMs carece de estructura con principios
- ❷ **Solución:** la filosofía de habilidades aporta recursos conceptuales ricos
- ❸ **Marco:** las TAAs combinan estructura y flexibilidad
- ❹ **Beneficios:** mejores evaluaciones, vocabulario teórico más claro, colaboración productiva
- ❺ **Implementación:** múltiples perspectivas filosóficas pueden contribuir
- ❻ **Impacto:** debates abstractos se vuelven concretos y testeables

Mensaje central

No necesitamos resolver todos los problemas filosóficos para avanzar: hacen falta marcos que vuelvan explícitos, testeables y revisables nuestros compromisos teóricos.

Próximo paso

Desarrollo colaborativo de TAAs concretas para dominios prioritarios.

Preguntas clave para nuestro trabajo

- ¿Qué capacidades priorizar para las primeras TAAs?
- ¿Cómo equilibrar riqueza teórica con usabilidad práctica?
- ¿Qué estrategia seguir frente a marcos de evaluación existentes?
- ¿Dónde generan divergencia nuestros compromisos filosóficos?
- ¿Cómo validar empíricamente las predicciones TAA?

¡Gracias!

Preguntas y discusión