

STRUTTURA DEL PAPER

"Validating Claude with Claude: Reproducibility and Consistency in LLM Claim Validation"

Claudio Cammarano — Draft Planning Document v2 — Febbraio 2026

Aggiornamento v3

- ★ Sez. 3.1: aggiunto Peirce come fondamento del metodo abduttivo del framework
 - ★ Sez. 6.4: nota su LLM come acceleratori della ricerca linguistica (letteratura 2024)
 - ★ Sez. 7 (Conclusion): filone futuro su tassonomia NA_ILLOCUTORIO e speech act annotation
 - ★ References: +6 fonti rispetto alla versione originale → tot. 56
- v1: austin1962speech (Pre-Step 0)
- v2: chang2024language, yu2024assessing, ziems2024llms (LLM+linguistica)
- v3: peirce1931collected, peirce1903harvard, fann1970peirce (abduzione)
- ★ Appendici: aggiornate versioni tassonomia (v2.2) e Coding Manual (v2.0)

Overview

Sez.	Titolo	Stato	Parole	Note
1	Abstract	Strutturata	~200	
2	Introduction	Bozza avanzata	~800	
3	Theoretical Framework	Bozza avanzata	~1200	
4	Research Design & Methodology	Bozza iniziale	~1000	
5	Results	Da scrivere	~800	
6	Discussion	Da scrivere	~600	★
7	Conclusion	Da scrivere	~300	★
—	References	Strutturata (56 fonti)	—	★
—	Appendices	Strutturata	—	
TOTALE			~5000	

★ = aggiornato in v2 | ■ = nota operativa

1 Abstract

Stato: Strutturata

- Research question: riproducibilità/consistenza validazioni LLM
- Framework: tassonomia 7D (D1–D7), 11 categorie Content Type
- Metodologia: 3 claim × 30 repliche = 90 validazioni
- Metriche: % agreement + κ di Fleiss per dimensione
- Contributo: test empirico di validazione epistemica su LLM

Stima parole: ~200

2 Introduction

Stato: Bozza avanzata

- 2.1 Il problema: LLM come fonte di conoscenza — fiducia senza metodo
- 2.2 Gap nella letteratura: mancanza di framework epistemici sistematici per validazione LLM
- 2.3 Il bias di conferma vs sycophancy: distinzione metodologica fondamentale
- 2.4 "Ontology precedes epistemology": motivazione dell'approccio
- 2.5 Struttura del paper

Stima parole: ~800

■ Includere caso Monaco come motivating example

3 Theoretical Framework

Stato: Bozza avanzata

- 3.1 Fondamenti filosofici: Popper, Lakatos, Kuhn, Quine — e Peirce
- 3.2 Il framework a 7 dimensioni: razionale e struttura
- 3.3 D1 Content Type: la dimensione primaria e le 11 categorie
- 3.4 Dimensioni attributive D2–D7: ruolo e relazione con D1
- 3.5 Regola di Esclusione (predicati di pura esistenza)
- 3.6 Pre-Step 0: verifica atto illocutorio (Austin 1962) [v2.2]
- 3.7 Nota epistemica D4.3: asimmetria strutturale validazione causale
- 3.8 Epistemic Responsibility Check: virtue epistemology applicata

Stima parole: ~1200

■ Citare Constitutional AI (Bai et al. 2022) per ancoraggio AI safety

★ Nota aggiunta v3 — Sez. 3.1: Peirce come fondamento del metodo generativo

La struttura metodologica del progetto è abduttiva nel senso peirciano: i claim di frontiera

funzionano come "fatti sorprendenti" che generano nuove ipotesi sul framework. La domanda

implicita è sempre: quale integrazione alla tassonomia, se vera, renderebbe questo caso

meno sorprendente? È la retrodution di Peirce — l'unica forma di inferenza che introduce

idee nuove nel sistema.

Questo colma un gap nella sez. 3.1 attuale: Popper, Lakatos, Kuhn, Quine spiegano come si testano le ipotesi; Peirce spiega come si generano. Il paper ha bisogno di entrambi.

Connessione semiotica: la precedenza di D1 sulle dimensioni D2-D7 è strutturalmente analoga alla precedenza categoriale nella semiotica peirciana — il tipo del segno

(icone/indice/simbolo) determina come il segno va interpretato, non il contrario.

"L'ontologia precede l'epistemologia" è una formulazione moderna di un principio che Peirce aveva già operazionalizzato nella teoria dei segni.

Connessione lakatosiana: il ciclo abduzione → deduzione → induzione corrisponde esattamente alla struttura del progetto: i claim di frontiera generano ipotesi (abduzione); il framework le formalizza (deduzione); le 90 validazioni le testano (induzione).

Riferimenti: peirce1931collected (vol. 5, §5.180: struttura canonica abduzione)

peirce1903harvard (Harvard Lectures: abduzione come unica inferenza generativa)

fann1970peirce (monografia: abduzione e scoperta scientifica)

Tono suggerito: non introdurre Peirce come autorità esterna — riconoscere che il metodo

era già peirciano implicitamente, e che nominarlo esplicitamente è un debito da pagare.

4 Research Design & Methodology

Stato: Bozza iniziale

- 4.1 Research question formale e ipotesi operative
- 4.2 Design sperimentale: 3 claim × 30 repliche, prompt identico
- 4.3 Criteri selezione claim (copertura D1, variabilità D4/D5)
- 4.4 Protocollo di codifica: Coding Manual v2.0
- 4.5 Metriche: % agreement, κ di Fleiss, criteri di successo
- 4.6 Pre-registrazione OSF: garanzia anti-HARKing
- 4.7 Limitazioni metodologiche dichiarate

Stima parole: ~1000

■ **Includere criteri di successo espliciti: $\kappa \geq 0.61$ = accordo sostanziale**

■ **Citare [yu2024](#) assessing come precedente metodologico diretto (LLM come co-annotatore pragmatico)**

■ **Citare [ziems2024](#) llms per tensione classificazione vs generazione: rilevante per limiti sez. 4.7**

5 Results

Stato: Da scrivere (dopo raccolta dati)

- 5.1 Statistiche descrittive: distribuzione coding per dimensione
- 5.2 % Agreement per claim e per dimensione
- 5.3 κ di Fleiss: tabella comparativa D1–D7
- 5.4 Varianza inter-claim: le 30 repliche sono stabili?
- 5.5 Analisi qualitativa: casi di divergenza notevoli

Stima parole: ~800

■ **Da scrivere dopo raccolta dati — struttura già definita**

6 Discussion

Stato: Da scrivere

- 6.1 Interpretazione risultati: cosa dice il κ sulla tassonomia?
- 6.2 Dimensioni ridondanti (D6, D7B): conferma empirica
- 6.3 D4 come dimensione più preziosa: validazione attesa
- 6.4 Implicazioni per AI epistemology e AI safety

- 6.5 Sycophancy vs conferma bias: cosa emerge dai dati?
- 6.6 Applicazioni: domini ad alto rischio (medical, legal, policy)

Stima parole: ~600

■ Collegare alla letteratura Bender, Bommasani, Mitchell

★ Nota aggiunta v2 — Sez. 6.4: LLM come acceleratori della ricerca linguistica

Il presente studio si colloca in una tendenza emergente (2023–2025) in cui gli LLM vengono

utilizzati come strumenti per accelerare la ricerca su fenomeni linguistici complessi, in particolare nell'annotazione pragmatica. Questa letteratura è metodologicamente affine al

nostro progetto ma si distingue per la variabile di controllo epistemologica introdotta dalla tassonomia ontologica: non solo "l'LLM classifica correttamente?" ma "la classificabilità dipende dal tipo ontologico del claim?"

Riferimenti da citare in sez. 6.4:

- chang2024language — panorama LLM in linguistica computazionale; pragmatica come dominio fragile
- yu2024assessing — annotazione pragmatica corpus-based; GPT-4 vs umano su atti linguistici
- ziems2024llms — tensione classificazione tassonomica (fragile) vs generazione spiegazioni (forte)

Tono suggerito: il progetto contribuisce a questa letteratura aggiungendo la dimensione ontologica come variabile indipendente — ipotesi: la riproducibilità dell'LLM covaria con il tipo di claim, non solo con la difficoltà del task.

7 Conclusion

Stato: Da scrivere

- Risposta diretta alla research question
- Contributo principale al campo
- Limiti e lavoro futuro (inter-rater, cross-linguistic, altri LLM)
- Implicazioni pratiche per utenti e sviluppatori

Stima parole: ~300

★ Nota aggiunta v2 — Sez. 7: filone futuro su NA_ILLOCUTORIO e speech act annotation

La distinzione NA_ILLOCUTORIO introdotta nel Coding Manual v2.0 (Pre-Step 0) apre un filone di ricerca autonomo che il presente paper può menzionare come lavoro futuro.

Problema: le sottocategorie di NA_ILLOCUTORIO (imperativo ordinario / imperativo retorico /

imperativo apofatico come il koan) non sono deducibili deduttivamente — richiedono un corpus.

La letteratura sugli speech act (Searle 1976, Bach & Harnish 1979, ISO 24617-2) non risolve

la questione del numero di categorie empiricamente rilevanti per la frontiera asserzione/non-

asserzione nel discorso ordinario e filosofico.

Proposta per la sezione lavoro futuro:

- Costruire un corpus di claim al confine illocutorio (koan, imperativi retorici, performativi

mascherati da assertivi, pseudoasserzioni mistiche)

- Testare stabilità delle categorie NA_ILLOCUTORIO con annotatori multipli

- Verificare se il numero di categorie empiricamente stabili è deducibile o solo corpus-derivato (congettura: non più di 5-6, cfr. Searle)

Citare: Searle 1976, Bach & Harnish 1979, Levinson 1983 cap. 5, ISO 24617-2

e yu2024assessing come proof-of-concept metodologico per LLM-assisted speech act annotation.

Appendici e Riferimenti

App.	Contenuto	Stato
App. A	Tassonomia completa v2.2 (D1–D7 con esempi + Pre-Step 0)	Completa
App. B	Coding Manual v2.0 (unificato: ex-v1.0 + Addendum D4.3 + Pre-Step 0)	Completo
App. C	Walkthrough annotato con 3 claim di esempio	Completo

App. D	Pre-registration OSF (testo completo)	Completo
App. E	Output grezzi Claude per 90 validazioni + codifica	Da raccogliere
Ref.	56 fonti BibTeX (epistemologia, filosofia della scienza, Peirce, AI, linguistica computazionale)	Strutturata

Target Journal e Criteri Submission

Target	Tipo	Criteri rilevanti	Priorità
ArXiv (cs.AI / phil)	Preprint	Apertura immediata, nessun peer review, citabilità rapida	1° — immediato
Synthese	Rivista peer review (filosofia/scienze)	Epistemologia formale, filosofia della scienza applicata	2° — 6-12 mesi
Episteme	Rivista peer review (epistemologia)	Social epistemology, knowledge sources, AI	3° — alternativa